



Universiteit
Leiden

The Netherlands

Using deep learning models in image processing

Xiong, Z.

Citation

Xiong, Z. (2026, May 19). *Using deep learning models in image processing*.
Retrieved from <https://hdl.handle.net/1887/4303646>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4303646>

Note: To cite this publication please use the final published version (if applicable).

Using Deep Learning Models in Image Processing

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden, op

gezag van rector magnificus prof.dr. S. de Rijcke,

volgens besluit van het college voor promoties

te verdedigen op dinsdag 19 mei 2026

klokke 11:30 uur

door

Zhan Xiong

geboren te Chongqing, China

in 1988

Promotor:

Prof. dr. ir. F.J. Verbeek

Co-promotor:

Dr. X. Tang (Southwest University)

Promotiecommissie:

Prof. dr. T.H.W. Bäck

Prof. dr. M.M. Bonsangue

Prof. dr. M.S.K. Lew

Dr. D.M. Pelt

Dr. X. Jiang (Shanghai University)

Prof. dr. E.O. Postma (Tilburg University)

ISBN: 978-94-6537-559-5

Cover designed by Shaogui Xiong.

Printed by Ridderprint.

Copyright © 2026 Zhan Xiong. All rights reserved. No part of this thesis may be reproduced in any form or by any means without the permission of the author.

The research presented in this dissertation was financially supported by the China Scholarship Council (CSC). Grant No.201306990022.

Contents

1	Introduction	1
1.1	Background	1
1.2	Preliminaries	3
1.2.1	Data Sets for Computer Vision Studies	3
1.2.2	Importance of Feature Space Learning in Computer Vision . .	4
1.2.3	Introduction to Attention Mechanisms	5
1.3	Research Questions	6
1.4	Thesis Structure	6
1.5	Terminologies and Definitions	8
2	Survey	11
2.1	Introduction	13
2.2	Developments	16
2.2.1	Convolutional Neural Networks	17
2.2.2	Visual Transformers	31
2.2.3	Hybrid Networks	46
2.2.4	Development Trend Summary	52
2.3	Techniques	53
2.3.1	Neural Architecture Searching	53
2.3.2	Neural Architecture Compression	54
2.3.3	Self-supervised/unsupervised learning	55
2.3.4	Knowledge Distillation	57
2.3.5	Diffusion Models	57
2.3.6	Deep Unfolding Techniques	58
2.3.7	Section Summary	59
2.4	Techniques	60

Contents

2.4.1	Classification	60
2.4.2	Object Detection	69
2.4.3	Image Generation	74
2.4.4	Universal Models	79
2.4.5	Medical Data Analysis	82
2.5	Conclusions	84
2.5.1	Discussions	84
2.5.2	Conclusion	87
3	Semantic Segmentation	89
3.1	Introduction	91
3.2	Related Work	95
3.2.1	Variational Bayesian methods	95
3.2.2	Clustering Algorithms	96
3.2.3	Attention Mechanism	98
3.3	Variational Gaussian Mixture Embedded Attention	99
3.3.1	Problem Formulation and Notation	100
3.3.2	Variational Bayesian Inference with Geometric Regularization	100
3.3.3	Cross-Attention as Concept Distillation	104
3.4	Implementation Details	106
3.4.1	Online Bayesian Updates for Batch Processing	106
3.4.2	Adaptions to Deep Learning	108
3.5	Experiment	111
3.5.1	Dataset and Metrics	111
3.5.2	Performance Analysis	113
3.5.3	Ablation Study	121
3.5.4	Limitations	126
3.6	Conclusions	128
4	Weakly Supervised Semantic Segmentation	131
4.1	Introduction	133
4.2	Related Work	136
4.3	Method	137
4.3.1	Probabilistic Framework	138
4.3.2	Implementation	141
4.3.3	Loss Function	143
4.4	Experiments and Results	144

4.4.1	Experimental Setup	144
4.4.2	Results and Analysis	144
4.4.3	Ablation Study	147
4.5	Conclusion	151
5	Instance Segmentation	153
5.1	Introduction	155
5.2	Related Work	160
5.3	Methods	162
5.3.1	Preliminaries	162
5.3.2	Dense Instance Segmentation Model	163
5.3.3	Lesion classifier	166
5.3.4	Implementation Details	167
5.4	Results	171
5.4.1	Datasets	171
5.4.2	Experiments	173
5.4.3	Complexity Assessment of Model Training	179
5.4.4	Ablation Study	180
5.5	Conclusion and Future Work	186
6	Conclusions and Discussion	187
6.1	Main Findings	187
6.2	Discussion	190
6.2.1	Algorithmic Perspective	190
6.2.2	Theoretical Perspective	191
6.3	Future Work	191
	Bibliography	193
	Appendix A Supplementary Materials in Chapter 5	271
	Summary	273
	Nederlandse samenvatting	275
	Curriculum Vitae	279
	Author Publication List	281

Contents

Acknowledgments	283
------------------------	------------

Chapter 1

Introduction

Motivation

This thesis aims to address the application of deep learning models in image processing from two perspectives. In the broader context, we uncover a logical clue that reveals the direction of various structural evolutions and technical developments in deep learning models. With respect to specific research topics, we focus on interpretability and applications in image classification, including semantic segmentation, object detection, and instance segmentation.

1.1 Background

Humans are very talented at jointly processing a variety of information and comprehensively understanding corresponding scenes. Through a series of sensory systems, our brain perceives the real world by integrating a spectrum of senses: vision, smell, hearing, taste, and touch. Among these senses, visual information is the most frequent and direct form of sensory stimulation. The brain will enhance the visual input signal, analyze various aspects of the information, and integrate all pieces of evidence via a complex network of neurons. By extracting information from the network of neurons, we can link observations and recognize patterns. This active brain event is referred to as thinking, which aims to reason/infer essence over evidence. Likewise, if a machine is capable of reasoning, it demonstrates Artificial Intelligence (AI), which is defined as distinct from the natural intelligence displayed by animals, including humans. Concerning natural

1.1. Background

intelligence, a process of inference reshapes neural connections by creating or deleting connections between neurons. In artificial intelligence, computational graphs, where nodes represent neurons and edges represent connections, imitate the reasoning process by adjusting edge weights. Consequently, the graphs are called **neural networks** and the process of adjusting weights is called **learning**.

In the field of computer vision, the most popular topics are mainly the following:

- **Classification** is the most fundamental task in the field of computer vision. Category cognition is used to correctly predict the class label(s) of a given image. Classification only extracts the coarsest information, yet requires the least supervision.
- **Segmentation** aims to efficiently generate a dense annotation for each pixel within a given image and classify them into different semantic regions. Semantic segmentation stops at class-level dense annotation, which only discriminates objects among different classes. Instance segmentation goes deeper toward instance-level annotation, further separating individual objects within each class.
- **Detection** is an intermediate level between classification and segmentation. The aim is to find a region-level classification that localizes each object while attaching a class label to the corresponding region.

In this thesis, we focus primarily on segmentation, object detection, and instance segmentation. Early deep learning models were based on fully convolutional neural networks, which cannot easily capture long-range dependencies between any pair of positions due to the limitation of the field of perception used in convolution operations [455]. In recent years, with the development of attention mechanisms, several deep learning models have contributed to the state-of-the-art (SOTA) in segmentation [178, 105, 368] and detection [870, 367, 808]. These SOTA neural networks, however, work as black boxes. Without clear interpretations, AI is hindered from being widely applied in realistic scenarios due to a lack of administration. Therefore, it is worth exploring mathematical explanations to understand it better.

In this chapter, we first give some insight into the datasets that we used in our experiments. Given these datasets, we explain the major aspects addressed in this thesis. We investigate feature space learning under full supervision and formulate attention mechanisms as a variational Gaussian mixture. Next, we elaborate on how to decompose a feature space into background and foreground classes and introduce additional geometric constraints to facilitate semantic segmentation with only image-level supervision. Both aspects share the importance of building good feature spaces

from which we can learn better class-specific representations. Moreover, we explored extending attention mechanisms to process very dense objects.

1.2 Preliminaries

1.2.1 Data Sets for Computer Vision Studies

The availability of image-annotation paired datasets is limited due to the significant manual effort required to collect pixel-level annotations.

The research in this thesis is based on four public datasets and one private medical dataset:

- The PASCAL Visual Object Classes dataset [167] has been widely used as a benchmark for object detection, semantic segmentation, and classification tasks. It contains 20 object categories and has been split into three subsets: 1,464 images for training, 1,449 images for validation, and a private testing set.
- The Microsoft COCO dataset [415] is a large-scale dataset that addresses four computer vision tasks: semantic segmentation, object detection, instance segmentation, and panoptic segmentation. Containing more than 200,000 annotated images, there are 80 object categories for object detection, semantic segmentation, and instance segmentation, while there are 80 thing categories plus an extra 91 stuff categories for panoptic segmentation.
- The Cityscapes dataset [123] is a large-scale database that focuses on the semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30 classes, grouped into 8 categories. The dataset consists of around 5000 fine-annotated images and 20000 coarse-annotated images, captured in 50 cities over several months, during daytime and under good weather conditions. Its images are manually selected to feature the following: a large number of dynamic objects, varied scene layouts, and varied backgrounds.
- The ADE20K semantic segmentation dataset [849] contains more than 20000 scene-centric images exhaustively annotated with pixel-level objects and object parts labels. This dataset can support tasks of semantic segmentation, object detection, instance segmentation, and panoptic segmentation. There are a total of 150 object and stuff classes, including blocky regions such as sky, road, grass, and discrete objects such as person, car, and bed.

1.2. Preliminaries

- We use images from kidney biopsies. These images are prepared in accordance with the Pathology Laboratory Protocol for kidney biopsies. The tissues were collected by core needle biopsy, Serra-fixed, rapidly processed using a routine tissue processor, then paraffin-embedded, serially sectioned at $5\text{ }\mu\text{m}$, mounted onto adhesive slides, and stained with JONES silver staining. All biopsy samples, both transplant and native kidneys, from 148 patients were collected from the multi-center archive of the Departments of Pathology at LUMC, AUMC, and UMCU in the Netherlands, or from the Department of Nephrology in Leuven, Belgium. A pathology staff member anonymized all biopsies before further analysis. The 148 Jones-stained WSIs were obtained from scanners with $\text{PPM} = 0.25$ in BIG-TIFF format¹, and 303 patches were extracted by software ASAP version 1.9 for Windows² at level 0 of the BIG-TIFF images. 249 patches are used for training and 54 patches for evaluation.

1.2.2 Importance of Feature Space Learning in Computer Vision

Feature space refers to the n -dimensional vector space on which all possible feature vectors reside. Each vector is extracted from the given data via some defined mapping functions [349]. Feature space is one of the most fundamental infrastructures on which machine learning is built to represent and learn **concepts** digitally. Given a feature space, any data can be represented as a feature vector, giving coordinates in the n -dimensional space. Then, semantic segmentation tasks in computer vision are formulated as decision boundary problems that separate different class-specific examples with minimal error. Modern deep learning models can automatically learn to extract appropriate features and infer decision boundaries in a fully supervised way. However, the training process operates as a black box, as it is too complex. To simplify the black box, we separate feature extraction from inference and focus on drawing decision boundaries in the given feature space. We can rethink the feature representation of each class over a given dataset in classification tasks according to two observations: (a) any given dataset can define a closed set of relevant classes, i.e. **foreground classes**, which implies an infinite complementary set of irrelevant classes, i.e. **background classes**. (b) Any objective function assigns different scores to different decision boundaries, thereby inevitably creating a subset of equipotential

¹<https://www.awaresystems.be/imaging/tiff/bigtiff.html>

²<https://computationalpathologygroup.github.io/ASAP/>

elements. The observation (a) indicates that it is irrational to infer background as a single class, whereas it is intractable to represent each one within precisely. The observation (b) implies that there is a set of necessary constraints that is learned to select among these equipotential elements and pushes decision boundaries toward the desired results. During training, both requests can be implicitly satisfied in fully supervised learning due to detailed target information, but this creates two difficulties for weakly supervised learning. Therefore, in this thesis, we are motivated to explore the construction of feature spaces for both fully and weakly supervised cases. For fully supervised semantic segmentation, we propose a variational Gaussian mixture model (VGMM) framework that leverages target information and constructs a feature space in which a single semantic vector represents each class. For weakly supervised semantic segmentation, we use conditional probabilistic models to complete two missing components due to using image-level target information: (1) to decompose background and foreground representations separately; (2) to integrate extra geometrical constraints to eliminate equipotential elements.

1.2.3 Introduction to Attention Mechanisms

Attention mechanisms are among the most important components in modern deep learning models because they are involved in constructing feature space. Attention mechanisms were first used for transformer models, and were initially shown to be effective in natural language processing tasks [642] and later for many computer vision works [156, 693, 674]. Attention mechanisms first take triplets (query, key, value) as input. For each query-key pair, attention probabilities are computed as feature similarities via the inner product. Finally, a new query will be reconstructed as the sum of values weighted by attention probabilities. In this thesis, we interpret attention mechanisms as soft clustering processes. Depending on the triplet (query, key, value) settings, attention mechanisms can be roughly categorized into two types. One defines a set of global queries and updates them across the entire dataset [654, 773], working in a K-means clustering manner. The other generates a set of dynamic/local queries and discards them after use [663], working in a K-Nearest Neighbors (K-NN) clustering style. In this thesis, (1) we generalize K-means attention mechanisms into probabilistic models and explore the learned feature spaces from deep learning models; (2) we integrate K-NN attention mechanisms with regional neural networks (R-CNNs) and diffusion models for dense instance segmentation.

1.3 Research Questions

To detail aspects of segmentation and detection algorithms, we propose the following research questions:

- RQ1 (Chapter 2) What are the current developments of deep learning models and their applications in image processing?
- RQ2 (Chapters 3) How can we explain an attention mechanism as a probabilistic model that captures global cross-class correlations in semantic segmentation tasks?
- RQ3 (Chapters 3) How can we integrate geometric constraints into a probabilistic model and efficiently solve it?
- RQ4 (Chapters 4) How can we interpret weakly supervised semantic segmentation as a feature space decomposition problem using conditional probabilistic models?
- RQ5 (Chapters 4) How can we eliminate the equipotential of the loss surface in the weakly semantic segmentation using geometric regularization?
- RQ6 (Chapters 5) How can we equip an attention mechanism with an R-CNN in a diffusion manner in support of efficiency and generalization of dense instance segmentation tasks?

1.4 Thesis Structure

The remainder of this thesis is structured according to the research questions presented in the section 1.3. These are subsequently presented in four research chapters and a concluding chapter.

- **Chapter 2** provides a novel perspective for organizing the development of deep learning models in the image processing domain: image classification, image segmentation, and object detection (related to the RQ1). The survey provides a blueprint for broadening our understanding of related techniques to address challenges in the computer vision community. It provides a comprehensive introduction to contemporary deep learning systems and their core components, presented chronologically. In addition, we introduce other stand-alone techniques

to improve the performance of deep learning models. Moreover, we collect representative work across a variety of image processing applications: classification, detection, image generation, and an important sub-field: medical data processing.

- **Chapter 3** narrows down our research topic on semantic segmentation and revisits the important attention mechanism with a K-means style. We introduce a variational Gaussian mixture model (VGMM) that efficiently captures global cross-class correlations (related to RQ2). Similar to MoCo [243], our VGMM can be used as a light-weight plug-in. All parameters will be updated using momentum. Our model can analyze the feature spaces of deep learning models and probe residual queries (as in RQ3). The experimental results show that our approach achieves performance comparable to that of the compared methods.
- **Chapter 4** further explores the function of the attention mechanism in a weakly supervised paradigm where we can exploit the long-range contextual features to compensate for the missing supervision information. Specifically, we formulate weakly supervised semantic segmentation as a conditional labeling assignment problem. Based on contrastive learning [87], we first decompose the feature space into background and foreground sub-spaces via a binary segmentation (refer to RQ4). Then, subsequent class-specific segmentation is just inferred within the foreground subspace through attention mechanisms. Finally, we use level-set methods for total-length and minimal-region constraints, ensuring that the semantic segments are smooth (refer to RQ5). The experimental results show that our approach achieves performance comparable to the other methods.
- **Chapter 5** extends our research range onto instance segmentation/object detection, and proposes a generative method to tackle complicated medical data: kidney pathology images. Specifically, the chapter focuses on an anchor-free instance segmentation approach that can process a large number of objects (varying from 300 to 1000). Our framework combines diffusion models, R-CNNs, and Transformers (related to RQ.6). The diffusion models will provide random Gaussian noise as initial bounding boxes to replace anchors, and iteratively refine these noisy boxes into predicted bounding boxes. The Transformers obtain pooled regional features from previously refined boxes, extract regional features from them, and interpret all features as dynamic queries (related to RQ6). Among different object proposals, attention mechanisms can capture long-term dependencies. Based on these proposals, RCNNs will predict final instance segments, thereby alleviating resource costs. The experimental results

1.5. Terminologies and Definitions

show that our approach outperforms the compared methods and can perform domain translation.

- **Chapter 6** elaborates the connection among the previous chapters and summarizes our contributions for exploring feature space, attention mechanisms, and probabilistic frameworks in computer vision tasks. In addition, we briefly describe the remaining challenges and potential improvements for the proposed approaches. Lastly, current challenges motivate future work on interpretability, cross-modality integration, and reliable data generation.

1.5 Terminologies and Definitions

In this section, we provide descriptions of common terms used in neural networks.

- **Feature**, used in deep learning, is an N -dimensional matrix (or formally called a tensor), containing representative information to describe key aspects of given data for a certain task.
- **Convolution** is a linear summation of element-wise products between two functions. One function, $f(x)$, is called the input function. The other, $g(x)$, is called the kernel (function). The domain of kernel D normally is a subset of the domain of $f(x)$ and $g(x)$ acts as weights to accumulate $f(x)$ into a real number $H(x)$:

$$H(x) = \sum_{i \in D} f(x - i)g(i) \quad (1.1)$$

- **Activation**, or activation function, is a family of non-linear functions to enhance the representation ability of neural networks. The most commonly used activation functions are the sigmoid, rectified linear unit (ReLU), softmax, and softplus.
- **Pooling** is an approach to down-sampling feature maps by summarizing the presence of features in patches of the feature map. Two common pooling methods are average pooling and max pooling, which summarize the average and the maximum activations of a feature, respectively.
- **Layer** logically is an atomic module that must be executed as a basic unit during runtime. The module can consist of several operations. Normally, a standard layer has convolution(s) and activation(s) w/o a pooling.

- **Pyramid** is a kind of data structure stacking with N layers. A layer l_k at k -th level has a larger size than that of l_{k+1} where $k \in \{0, 1, 2, \dots, N\}$.
- **Fusion** is a kind of mathematical operation $f(\cdot)$ used to integrate N features into M features (normally $M = 1$). For instance, element-wise sum, element-wise product, or concatenation.
- **Self-Attention**, or called attention in short, is a kind of feature that acts as a modulation. The modulation only responds to relevant regions over the feature and removes noise or negligible parts.
- **Ground truth** denotes a correct result to a certain task justified by a human. For instance, manually labeled binary segmentation serves as ground truth for training a semantic segmentation network.
- **Loss function** is an objective function that measures the difference/error between the output of a network and the given ground truth. The cumulative loss can guide the training network in adjusting its internal parameters to reduce error.
- **Annotation** normally relates to ground truth used in segmentation tasks. A dense annotation denotes per-pixel labeling, where each pixel is uniquely assigned to a class—for instance, a binary segmentation for all cars in an image. Additionally, instance-level dense annotation assigns each pixel to an object with a unique ID. For example, a segmentation for each car in the image. Notably, an instance annotation is not a binary image since each instance occupies a unique ID.
- **end-to-end** is a term to indicate if a network can output its prediction in the same format as the ground truth. For instance, if a ground truth is an instance annotation, then an end-to-end network also outputs an instance annotation without any other add-on post-processing procedures.
- **Fully convolutional network** means the whole network is built without a fully-connected layer. For more details, please refer to [444].
- **Multi-scale**, or called multi-resolution and cross-scale, is a term to indicate that features of different layers have different sizes due to the existence of pooling functions. A multi-scale network fuses features from different layers into enhanced features that combine context information and accurate spatial resolution.
- **Regression**, or regression analysis, estimates the relationships among variables. It includes many techniques for modeling and analyzing several variables when

1.5. Terminologies and Definitions

the focus is on the relationship between a dependent variable and one or more independent variables.

- **instance-awareness** is a term to indicate that if a network can sense individual identities sharing the same class label. For instance, if a network can correctly label every individual car in an image with a unique instance ID, it has instance-awareness.

Chapter 2

A Survey of Deep Learning in Image Processing: Developments, Applications and Challenges

Journal Artificial Intelligence Review (under review).

Chapter Summary

The rapid advancement of deep learning models (DLMs) in computer vision is driven by their capacity to unify feature representation and task inference into end-to-end automated frameworks, thereby surpassing manual feature engineering. This survey categorizes DLMs into Convolutional Neural Networks (CNNs), Visual Transformers (ViTs), and hybrid models based on their core components (convolutional vs. attention-based), systematically reviews their technical evolution, and critically analyzes their strengths and limitations. CNNs, the pioneering paradigm, leverage the locality of convolution kernels for efficient hierarchical pattern extraction. However, their reliance on multi-layer stacking to aggregate global context and rigidity to geometric variations has motivated innovations such as deformable convolutions to expand receptive fields. ViTs, emerging later, prioritize the globality of attention mechanisms to model long-range dependencies directly, but face computational bottlenecks and noise sensitivity, which are addressed by sparse attention and hierarchical tokenization. Hybrid models now dominate as optimal compromises, combining the locality of CNNs and globality of ViTs to balance efficiency and expressiveness. Moreover, we emphasize universal techniques—neural architecture search (NAS), model compression, and self-supervised learning—as indispensable tools to enhance robustness and adaptability across architectures. Despite progress, challenges persist in interpretability, data efficiency, and real-world robustness. We advocate prioritizing lightweight hybrid designs, explainable attention-convolution combinations, and cross-modal systems integrating more data beyond vision. This survey not only systematically reviews the technical evolution of DLMs but also proposes a forward-looking roadmap for adaptive, scalable, and trustworthy vision systems, grounded in current technological trajectories.

Keyword: Machine Learning Deep Learning Convolutional Neural Network
Vision Transformers CNN-Transformer Architectures

2.1 Introduction

The rapid advancement of machine learning (ML) has profoundly reshaped modern society, embedding itself into nearly every aspect of daily life. From personalized recommendations in e-commerce and content moderation on social platforms to real-time language translation and autonomous decision-making systems, ML technologies have become indispensable tools for fully exploiting the value of vast amounts of data (big data) and unleashing their immense potential. At its core, the capacity of ML systems to automatically discover hidden patterns in multi-modal inputs (i.e., text, images, audio, and sensor data) has exceeded the limitations of traditional approaches and advanced the computational boundaries to complex, high-dimensional information. Among the diverse sub-fields, computer vision stands out as a cornerstone due to its direct relevance to automating the interpretation of visual data.

In computer vision, traditional systems relied heavily on handcrafted feature engineering, in which experts manually designed filters (e.g., edge detectors, texture descriptors) to extract structure-dependent representations for task inference. While friendly to interpretability, these methods lack adaptability to inherent variations and complexities present in real-world data, such as lighting variations, occlusions, and deformable objects, hindering generalization across data structures or distributions. For instance, conventional techniques for lesion detection or cellular segmentation in biomedical imaging demand laborious parameter tuning to accommodate diverse tissue structures and imaging modalities. By contrast, ML-based approaches, particularly deep learning (DL) methods, circumvent these limitations by automating the construction of feature representations directly from data. By training on large-scale datasets, DL networks have achieved superior performance across diverse tasks like object detection, image segmentation, data simulation, drug discovery, and pathological diagnosis.

The explosive growth of the literature on deep learning for computer vision poses a unique challenge: organizing decades of research into a clear, logical structure that captures the core contributions, advancements, and limitations in the field. Existing surveys often categorize works by application domain or chronology, but this approach risks obscuring the underlying evolution of model design principles. To address this, we propose a novel organizational perspective rooted in the mathematical foundations of core computational components, namely *convolution kernels* and *attention mechanisms*. By dissecting their formulations, we systematically reveal the inherent strengths and limitations of Convolutional Neural Networks (CNNs) and Visual Transformers (ViTs), as well as their interconnections, paving the way for the foreseeable emergence of hybrid

2.1. Introduction

models that combine their advantages.

Convolutional operations, characterized by fixed receptive fields and weight sharing, excel at capturing localized spatial patterns through inductive biases like translation equivariance. These properties make CNNs highly efficient for processing complicated grid-structured data via a hierarchy of convolutional layers, while maintaining computational efficiency and good generalization performance, particularly on relatively small datasets. However, their rigid receptive fields and static kernel parameters hinder modeling of long-range dependencies and adaptation to complex geometric variations. Attention mechanisms, conversely, employ global receptive fields and compute long-range pairwise interactions across all input elements, thereby capturing the so-called global context. This flexibility allows ViTs to dynamically adjust weights based on the input and to benefit from global reasoning, such as in panoptic segmentation or cross-modal retrieval, but at the cost of quadratic computational complexity, sensitivity to noisy tokens, and data-hungry.

Compared to previous surveys, our survey proposes a novel organizational structure grounded in the mathematical and computational foundations of the models. The deep mathematical kinship naturally motivates the design principle of hybrid architectures that use flexible dynamic aggregation schemes across different feature scales, balancing convolutional locality with attentional globality for efficiency and expressiveness. For instance, deformable convolutions that learn adaptive yet sparse sampling grids can capture non-local representations from high-resolution feature maps with small overhead, whilst sparse attention mechanisms can model dynamic long-range correlations from low-resolution feature maps to support customized global inference for all input data. Such connections underscore the inevitability of hybrid designs, where complementary inductive bias of convolutional kernels enables efficient task-agnostic extraction of commonality (i.e., building a shared feature space), whilst the dynamic global contexts in attention mechanisms support flexible task-specific extraction of individuality (i.e., performing input-dependent task inferences), to tackle complex vision tasks.

In addition, this survey extends beyond mere taxonomy to encompass universal advancements, including neural architecture search, model compression, self-supervised learning, diffusion methods, and deep unfolding techniques. These can improve the overall performance of DL models, transcending architectural boundaries. Moreover, we illustrate the remarkable progress achieved in visual applications, including generic computer vision tasks (e.g., classification, detection, data generation) and a dedicated section 2.4.5 for medical data analysis. Given limited annotations, multi-modal data, and high-reliability requirements, it is worth paying attention to specialized archi-

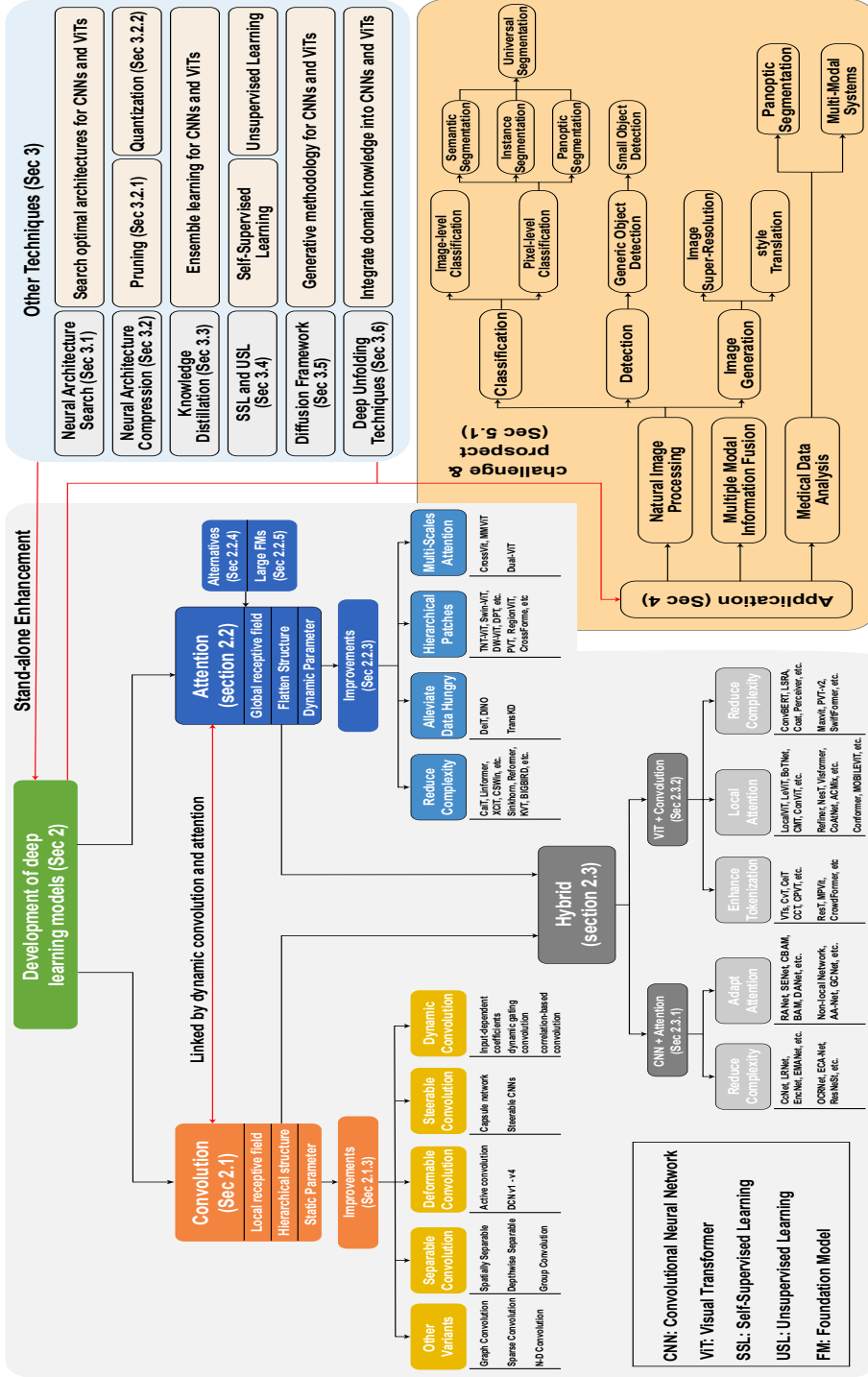


Figure 2.1: A simple visualization of the organization of the survey.

2.2. Developments

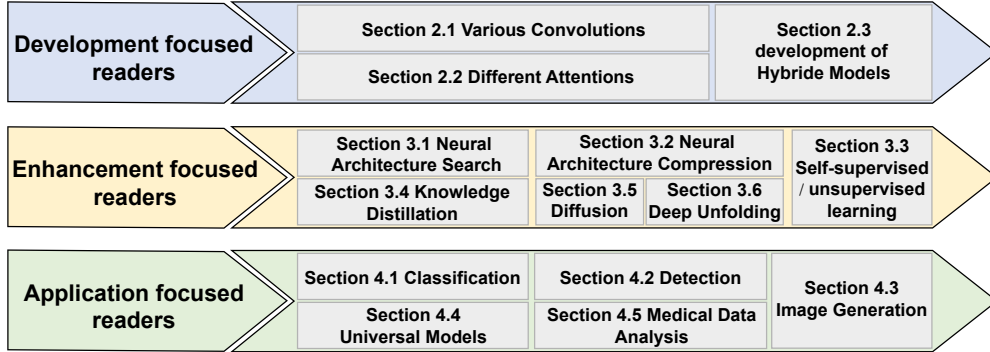


Figure 2.2: Text roadmap for each of the specific sections of the survey.

tectures and learning paradigms for medical vision tasks. Furthermore, we explore current challenges and future research directions. From the perspectives of lightweight model compression, interpretability enhancement, and multi-modal fusion, we delineate possible technical pathways for designing scalable, interpretable, and robust computer vision systems.

In conclusion, the survey presents a comprehensive architectural coverage. It provides a systematic review of the technical evolution, strengths, and limitations of the three major paradigms: CNNs, ViTs, and Hybrid models for organizing decades of research and revealing the underlying evolution of design principles, as depicted in Figure 2.1. To facilitate survey perusal, we have provided a text roadmap (Figure 2.2) to quickly locate the key sections.

2.2 Developments

Deep learning [257, 349] is a specific kind of machine learning. From a systematic perspective, a deep learning system typically has two phases: training and testing. At the training stage, a system consists of four components - a dataset that feeds valid raw data into the system, a representation model extracting machine-oriented discriminative features, a task-specific estimation metric (also known as loss/cost/object function) to measure the error from machine predictions to human expectations, and an optimization algorithm for automatic parameter tuning - while at the testing stage, it freezes all hidden trainable parameters and is ready to use. For improved representational capabilities, extracting features across multiple scopes and fusing information at different scales are essential. In fact, the evolution of deep architectures

in computer vision is fundamentally driven by the pursuit of an optimal balance between local and global feature aggregation, and between static and dynamic parameterization. Currently, there are three main architectures for image processing in computer vision: Convolutional Neural Networks (CNNs) established the initial paradigm, leveraging local, static kernels for efficient hierarchical feature extraction; Visual Transformers (ViTs) shifted the paradigm by introducing global, dynamic attention to capture long-range dependencies directly; Subsequently, Hybrid architectures emerged as a synthesis, strategically combining convolutional locality with attentional globality to achieve a more efficient and expressive balance. In this section, we elaborate on the progression and explicitly connect each architectural innovation to its underlying mathematical principles.

2.2.1 Convolutional Neural Networks

CNN is a feed-forward neural network that can extract features from data using convolution operators, as shown in Figure 2.3. Unlike traditional machine learning methods that rely on manual feature construction, CNN automates feature extraction by convolving data with a stack of differentiable composite functions. If each function denotes an artificial neuron, the extraction process is analogous to visual perception [349], where each neuron can respond to various features and transmit the information to the next neuron.

Foundations of CNNs

The basic module in a CNN comprises mainly four operations, with optional functions: a convolution operation, a normalization operation, an activation operation, and an optional pooling operation, along with additional residual connections or random dropouts.

Standard convolution

Standard convolution operations are the most fundamental elements in neurons, designed to process grid-structured data (e.g., images and videos). Originally developed from signal processing, convolutions were initially designed to detect patterns with predefined kernel weights. In deep learning, however, they have evolved into learnable feature extractors, in which kernel weights are optimized via backpropagation based on the input data. The essence of a convolution involves sliding a small filter over the input. The kernel performs pairwise multiplication between kernel weights and

2.2. Developments

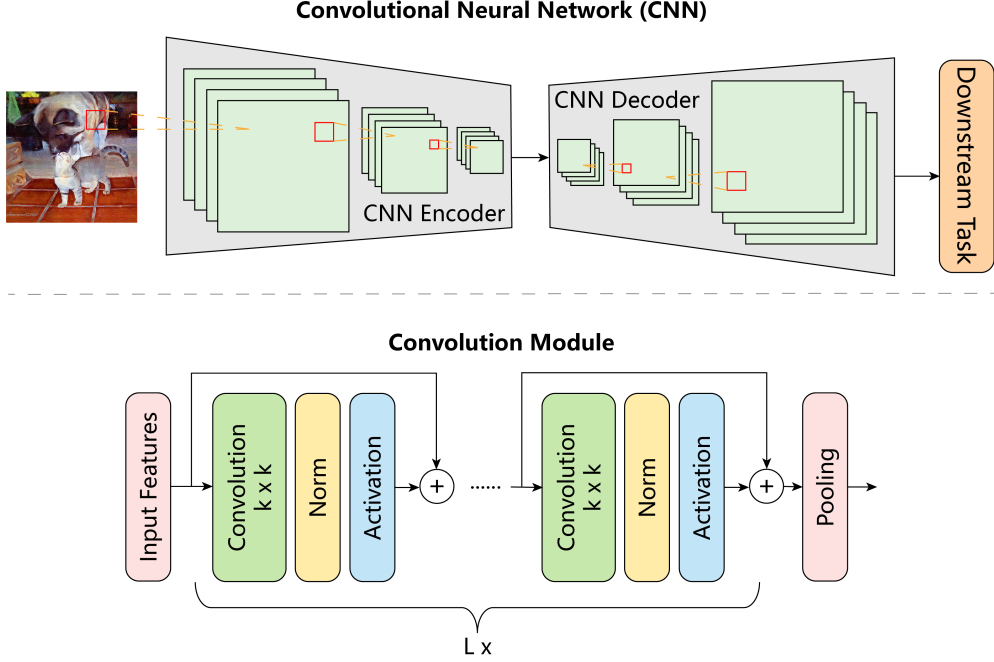


Figure 2.3: Schematics graphical representation of a generalized framework for a Convolutional Neural Network (CNN) with L blocks.

the corresponding input elements at each sliding position. This process captures local patterns and relationships within a local field, enabling the model to acquire essential features.

Mathematically, a generalized standard convolution ¹ in machine learning works as a sliding-window operation over the input feature maps. For clarity, we begin with a special case of a generalized 2D convolution and then extend the definition to higher dimensions. Consider an input feature map $X \in \mathbb{R}^{H \times W \times C_{in}}$, where H and W are the spatial height and width, and C_{in} is the number of input channels. A generalized 2D convolution contains C_{out} learnable kernels and **independently** applies each kernel to produce one channel of the output feature map $Y \in \mathbb{R}^{H' \times W' \times C_{out}}$. Therefore, focusing on the dynamic process of the arbitrary kernel f_c benefits understanding its hyperparameters. Depending on different roles, all arguments can be divided into three groups: 1) stride s decides the sliding displacement from the previous window center to the next; 2) padding size p ² controls the valid range of sliding window centers

¹To show the geometrical characteristics of convolutions, dilated convolutions are also included as one type of standard convolutions in this paper.

²Although different padding methods depend on the various boundary conditions, we argue that

along the boundary to avoid discarding edge information; 3) kernel size k (usually odd), dilation rate $d \geq 1$, and input channels C_{in} jointly decide a 3D window for local operation.

a. Window Sliding with boundary condition

To preserve edge information, the input X is often extended by adding p rows/-columns of values (normally 0) around its borders. Then the window size k , step length s , and the padding size p jointly defines a function to map the sliding window centers (x, y) from the coordinates (x', y') on the original input:

$$x' = p + x \times s + (k - 1)/2, \quad y' = p + y \times s + (k - 1)/2 \quad (2.1)$$

Likewise, considering the stride s and padding size p , the output size are given by:

$$H' = \lfloor \frac{H + 2p - d \times (k - 1) - 1}{s} + 1 \rfloor, \quad W' = \lfloor \frac{W + 2p - d \times (k - 1) - 1}{s} + 1 \rfloor \quad (2.2)$$

Common padding schemes are valid when $p = 0$ and same (chosen p to get $H' = H$ and $W' = W$ when $s = 1$).

b. Local window operation

It is common to overlook that each 3D kernel defines only a 2D grid, since the grid is shared isotropically across all channels. Denote an arbitrary kernel as a 3D weight tensor $f_c \in \mathbb{R}^{k \times k \times C_{in}}$, then the window is defined by a regular sampling grid $\mathcal{R}$ that contains k^2 offsets stretching by dilation rate d :

$$\mathcal{R}_d^k = \left\{ (d \cdot i, d \cdot j) \mid i, j \in \left\{ -\left\lfloor \frac{k}{2} \right\rfloor, \dots, \left\lfloor \frac{k}{2} \right\rfloor \right\} \right\} \quad (2.3)$$

where $\lfloor \cdot \rfloor$ denotes a floor function. Let each element $(\Delta x, \Delta y) = (d \cdot i, d \cdot j) \in \mathcal{R}$ corresponds to one offset point, centered at $(0, 0)$. When $d = 1$, the offsets are the integer coordinates of a $k \times k$ square centered at $(0, 0)$, inducing a standard dense $k \times k$ grid. When $d > 1$, the kernel samples are spread out, producing a dilated (sparse) sampling pattern that effectively enlarges the receptive field without increasing the

they do not influence the mathematical representation.

2.2. Developments

number of parameters (e.g., in dilated convolution). Inside the window, the input values are weighted by the kernel weights and summed across all input channels:

$$Y(x, y, c_{out}) = \sum_{(\Delta x, \Delta y) \in \mathcal{R}_d^k} \sum_{c_{in}}^{C_{in}} X(x' + \Delta x, y' + \Delta y, c_{in}) \cdot f_c(\Delta x, \Delta y, c_{in}) \quad (2.4)$$

Notably, the kernel applies the same grid isotropically across the spatial dimensions: the same set of offsets is used at every window center, but the summation extends over all input channels. Thus, the kernel is inherently **a channel-wise aggregation followed by spatial sliding**, which can serve as the foundation for depthwise separable convolution.

Other Components

Pooling operation is necessary for CNNs because it can expand the receptive field of the convolution without a significant computational overhead. A pooling (i.e., down-sampling) operation partitions the input feature map into multiple regions, and each region is pooled into a statistical summary. With fixed receptive fields, successive convolutional operations can effectively search over larger spatial scopes using reduced feature maps. The most commonly used pooling functions are max pooling [858] (selecting the mode of one region) and average pooling [41] (selecting the mean of one region). Therefore, pooling operations enable the hierarchical structure of CNNs and introduce a strong inductive bias that the learned feature maps must approximately satisfy the slow-variation assumption. Otherwise, a statistical summary (e.g., mode or mean) cannot precisely describe the sampling region. Applying pooling techniques can significantly reduce computational complexity and accelerate model training. There is some theoretical research [40, 306] that gave a detailed analysis of pooling operations.

Normalization operation is a crucial technique in CNNs that standardizes feature vector distributions to have zero mean and unit variance. Distribution standardization can reduce internal covariate shifts between layers induced by parameter updates. Consequently, each layer will operate under consistent conditions, stabilizing the training process. In addition, normalizing feature vector magnitudes to a more compact range can significantly reduce the influence of extreme-value outliers. Without extreme features, normalization can mitigate vanishing (too small) / exploding (too large) gradients and alleviate overfitting by preventing feature learning from relying too heavily on outlier features with extreme values. Therefore, these are crucial for fast convergence

and generalization, especially in deep architectures. Common normalization techniques in CNNs are designed to standardize along different dimensions for various application scenarios. Batch Normalization [288] computes the mean and variance within every mini-batch and its performance is dependent on the batch size; Layer Normalization [19] normalizes across all feature channels of one feature map independently of the batch; Instance Normalization [634] normalizes at one feature channel within one feature map independently; Group Normalization [701] normalizes groups of feature channels within one feature map, providing a balance between batch and instance normalization.

Activation operation plays a crucial role in introducing nonlinearity into CNNs. Without non-linear transformations, a stack of linear transformations would essentially act as a single linear transformation, limiting the ability to learn complex patterns and relationships in data [552, 553]. For this reason, activation functions are applied to each neuron’s output, and non-linear transformations are introduced before the input of the next neuron. Similar to the function of a neuron in the human brain, an activation function acts as a stimulus between two neurons, controlling the transmission of information from one to the other. By stacking many convolutional layers and activation functions, deep neural networks can approximate any function, thereby greatly enhancing their ability to fit data. Within CNNs, there are mainly two types of typical activation function types: 1) map functions that non-linearly transform the input to different representation spaces, including sigmoid, tanh, and softmax; 2) gate functions that filter out irrelevant information include Rectified Linear Units (ReLU) [487] and its variants like Parametric ReLU [246], Randomized ReLU [730], Gaussian-error linear unit (GELU) [251], SiLU [478], Exponential Linear Unit (ELU)[118], Squareplus [28]. In summary, activation functions enable CNNs to model intricate patterns and decision boundaries through nonlinearity, thereby enhancing their performance across various tasks.

Residual connection, also known as a skip connection, is a shortcut that allows a layer to bypass multiple intermediate layers and connect directly to the computation of the destined layer. This simple modification has profound implications for addressing the degradation problem, where increasing the network depth leads to worse performance due to training difficulties. In very deep networks, gradients can become vanishingly small during backpropagation, making it difficult to train the early layers. Residual connections provide a direct path for gradients to flow, stabilizing the training process. Originating from ResNet [248], the residual connection enables the deep structures

2.2. Developments

in CNNs. Recently, residual connections have also been successfully applied to other types of neural networks, such as recurrent neural networks (RNNs) and transformers.

Dropout facilitates regularization in the network by randomly omitting some units or connections with a predetermined probability. This random dropping of connections or units yields different sub-network architectures, from which one combination can be interpreted as an approximation to the mixture-of-experts [594]. Each expert can generate weak predictions, and the ensemble of all experts can eventually enhance generalization and mitigate overfitting.

Characteristic of CNNs

Within a CNN, each standard convolutional layer extracts information from a neighborhood $\mathcal{R}$ of sampled locations and is responsible for detecting a distinct pattern in the input. During training, the model undergoes backpropagation and gradient descent to learn optimal convolutional filters. The essence enables the model to discern meaningful patterns in the data for feature representation automatically. From the perspective of extracting features at different levels, the mathematical definition of standard convolution forces the design of CNNs that must be characterized in the following three respects:

1. **Local receptive field:** Each neuron is connected only to a few neighboring neurons based on the sampling locations $\mathcal{R}$. This design promotes the extraction of short-term correlations because each neuron learns a specific pattern within its local scope. Local receptive fields are computationally efficient, focusing on small regions to capture local details.
2. **Hierarchical structure:** Extracting long-term dependencies by convolutions with increasing receptive field is computationally expensive. A rational alternative is to use a hierarchical structure that can gradually decrease the sizes of feature maps stage-wise via stride convolutions or pooling operations. With significant computational overhead, convolutions can sample from larger effective scopes, but at the cost of losing spatial information. Consequently, this architecture can exhibit different features from local to global, but it introduces a strong inductive bias: the learned feature maps must approximately satisfy the slow-variation assumption. Otherwise, a statistical summary (e.g., mode or mean) cannot precisely describe the sampling region.
3. **Static Parameter:** the weights in convolutional kernels are static because the

same kernel is applied across all sliding locations. This weight sharing reduces the number of parameters and accelerates computation, but introduces a strong inductive bias toward translation equivariance. The bias means that if the input feature is translated, the output changes accordingly. In other words, pattern detection is not related to spatial locations. Using static parameters enables training with less data.

Improvements of CNNs

However, standard convolutions face limitations in computational efficiency, geometric transformations, and adaptability to complex geometric variations. To address these limitations, recent works focus on three directions: (1) reducing computational complexity via separable convolutions, (2) enhancing geometric adaptability through deformable kernels, and (3) improving dynamic feature aggregation with steerable or dynamic convolutions. We organize these improvements by technical focus and highlight their potential application scenarios.

Separable Convolution

Separable convolution is a convolutional operation in deep learning that aims to reduce computational complexity by decomposing the convolution into a product of simpler matrices. They are particularly well-suited to scenarios with limited computational resources, such as mobile devices or real-time applications. By factorizing larger kernels into multiple smaller kernels, a separable convolution offers a compelling trade-off: significantly reduced computational cost and model size while maintaining acceptable performance. According to different splitting schemes, there are two main types of separable convolutions: spatially separable and depthwise separable convolutions.

Spatially separable technique decomposes a single standard convolution into a series of lower-dimensional convolutions on different spatial dimensions. This tensor decomposition can be achieved using tensor decomposition techniques, like Singular Value Decomposition (SVD) [142, 291, 143, 824, 392], Tucker decomposition [325, 412, 337], CP decomposition [347, 14, 852], and other decomposition methods [497, 662, 765, 569, 353, 651]. Given a set of decomposed components, a standard kernel can be expressed as the tensor product of multiple smaller kernels. For instance, a spatially separable 2D convolution kernel with spatial dimensions $F \times F$ can be split into two 1D kernels: $F \times 1$ and $1 \times F$. By applying these two filters sequentially, the spatially separable convolution achieves the effect of a full $F \times F$ convolution but

2.2. Developments

with significantly lower complexity: reducing from $\mathcal{O}(F^2)$ multiplications down to $\mathcal{O}(F)$. The main advantage is that there is no performance degradation after kernel separation, unless the matrix is decomposable. However, spatial separable convolutions have limited application in CNNs because not all convolution kernels are separable. Therefore, this decomposition may narrow the search for all possible kernels during training, leading to a performance trade-off. Hence, the applications of spatial separable convolution are significantly limited.

Depthwise separable technique is an efficient alternative that splits tensor computations of multiple convolutions into two steps: depthwise convolutions and pointwise convolutions. Depthwise convolutions apply a separate kernel to each input channel, capturing spatial patterns independently for each channel. Pointwise convolutions then use 1×1 convolutions to combine the output channels from the depthwise step, effectively aggregating the information. Depthwise separable convolutions significantly reduce the number of parameters and computation while maintaining model performance, making them popular in mobile and embedded applications. Therefore, the decomposition does not rely on the special properties of kernel functions. By decoupling the spatial filtering from cross-channel filtering, depthwise convolution achieves higher computational efficiency, reducing complexity from $\mathcal{O}(F^2 \cdot C_{\text{out}})$ to $\mathcal{O}(F^2 + C_{\text{out}})$. F is the kernel size and C_{out} denotes the channel number of the output feature maps. Depthwise separable convolution [211, 612, 417, 132], therefore, is well-suited for resource-constrained environments. The variants of MobileNet [265, 563, 264], ShuffleNets [823, 457], and Xception [112] are popular CNN architectures that use depthwise convolution to reduce model size and improve inference speed without significantly compromising performance.

Group convolution is another solution to the depthwise separable technique that divides depth channels into a group of smaller filters based on the assumption that each group of filters is mutually independent. Therefore, it simplifies the original full-matrix multiplication to a diagonal block-matrix multiplication along the depth dimension. Therefore, sparsity in the off-diagonal blocks improves efficiency and reduces the complexity by a factor of g , where g is the number of groups. Group convolution was first applied to AlexNet [340] to reduce the number of parameters. Then, ResNeXt[716] overlaps smaller convolution blocks to approximate ResNet’s convolution blocks. Moreover, CondenseNet [276] introduced sparsity-inducing regularization and divided group convolution training into several steps. However, previous methods require manual group selection. GroupNet [834] combines network architecture search (NAS) with group

convolution, enabling automatic learning of the number of groups via an end-to-end approach.

Deformable Convolution

Deformable convolution is an advanced variant of standard convolution that extends the receptive field at low computational cost. In segmentation or detection tasks, it is crucial to pay closer attention to the contents within semantic / object regions. In other words, not all locations in a receptive field contribute equally to an output response. Locations within regions are much more important for sampling than those outside the region. Therefore, uneven sampling requires convolution operations to account for geometric shapes. Constrained to the regular structure of the receptive field, standard convolution evenly samples at fixed locations and, therefore, cannot efficiently tackle geometrical variations. That means a stack of standard convolutions can linearly expand their theoretical receptive field scope to cover arbitrary geometrical shapes with increasing computational complexity. In contrast, deformable convolution enhances geometric modeling by learning input-dependent sampling offsets. The core idea is to transform the shape of receptive fields from a regular grid to arbitrary shapes. Specifically, deformable convolutions define sub-networks to learn additive disturbances $(\Delta p_x, \Delta p_y)$ alongside the static offset $(\Delta x, \Delta y)$ in a regular grid (refer to Eq.2.3). Then adaptive offsets can be generated as: $\Delta x' = \Delta x + \Delta p_x$ and $\Delta y' = \Delta y + \Delta p_y$. Therefore, a static convolution kernel has been replaced with spatially variant operators that use input-dependent geometric information to adaptively sample at more important locations. Under the assumption of local continuity in feature representation, the deformable convolutions have dynamic, receptive fields which are equivalent to larger receptive fields with sparse non-zero elements in standard convolutions. Compared to standard convolution with kernel size k , deformable convolution incurs an additional cost for offset prediction (equivalent to one extra $k \times k$ convolution). It requires at least bilinear interpolation for sampling at non-integer positions. The total complexity is approximately 2 times that of a standard convolution. However, by enabling adaptive receptive fields, it often achieves higher accuracy with fewer layers, offering favorable trade-offs between efficiency and accuracy in tasks that require geometric flexibility, such as object detection and segmentation. Active convolution [300] was an early work to augment sampling locations with learnable offsets. However, the offsets were static model parameters shared across different spatial locations, making them less general and less adaptive. DCNv1 [127] extends the sampling offsets as a dynamic output from

2.2. Developments

other sub-networks, allowing offsets to be adaptive to different geometrical shapes at different locations. Subsequently, DCNv2 [869] introduced an extra learnable modulation scalar for each offset, allowing it to adjust the importance of the sampled location, thereby providing additional flexibility. To reduce computational cost, DCNv3 [672] introduced depthwise separable convolution to compute modulation scalars. Besides, it splits deformable convolution into groups and normalizes the modulation scalars at each sampling point using the softmax function. Each group has its own sampling offsets and modulation scalars, and can have different spatial aggregation patterns. Therefore, it can enhance feature extraction by ensembling different patterns with normalized modulation scalars. Recently, DCNv4 [723] aims to speed up the computation by removing softmax normalization and optimizing memory access to minimize redundant operations.

Steerable Convolution

Steerable convolution is a convolutional operation used in computer vision to handle geometric transformations, such as rotations, efficiently. Unlike standard convolutions, which must use large amounts of static filters to enumerate all possible orientations, steerable convolutions aim to construct a vector space to represent arbitrary orientations. Inspired by the mixture-of-experts' architecture [255] in capsule networks [256, 2], steerable convolutions also encode the rotation of a pattern in the input as a linear combination of basis kernels. Normally, the combinatorial coefficients that are a set of learnable parameters can be interpreted as dynamic routines [559, 220]. At the same time, the disentangled basis filters can be understood as expert sub-networks dedicated to different geometrical transformations. Therefore, the steer convolutions define a representation space in which each orientation or transformation can be decomposed into (normalized) coordinates of basic pattern units. Cohen et al. [119] proposed steerable convolutional operations based on group representation theory. They designed equivariant filter banks and represented geometrical variations as linear transformations in a group. The constructed space increases the flexibility of equivariance at low computational complexity. Based on the 2-dimensional steerable convolution, Weiler et al. [690] proposed the 3-dimensional steerable convolution that maps rigid body transformations in 3-dimensional Euclidean space onto the representation space. Besides, the Euclidean group $E(2)$ [689], and its subgroups are given as general solutions to the kernel space to adapt to the rotation and reflection of planar images. In summary, steerable convolutions are suitable for orientation / pose-sensitive scenarios, including

biomedical image analysis, 3D object recognition, and object detection.

Dynamic Convolution

Dynamic convolution is a more general concept in CNNs that enhances feature representation without significantly increasing computational cost. Instead of using a large set of static kernels to enumerate all possible feature representations, dynamic convolution defines generation routines of feature expression. It dynamically selects appropriate kernels for the input mini-batch. The dynamic characteristic implies constructing implicit representation spaces to handle more complicated contexts, thereby compressing redundancy in the set of static kernels. According to different generation methods, there are mainly three types of dynamic convolutions: 1) One of the most used types is to generate input-dependent aggregation coefficients over parallel fixed convolution kernels. This type of dynamic convolution still satisfies the **parameter sharing** constraint, i.e., it applies the same dynamic kernels to all locations. Inspired by the architectures of mixture-of-experts, the linear combination coefficients are obtained by different dynamic route functions running on the input [743, 97, 829]. Noted, capsule networks and steerable convolutions can also be regarded as special cases of dynamic convolution with dynamic combinatorial coefficients. 2) The second popular type is to generate gating information that can guide convolution to execute only on selected locations, extending the property of **parameter sharing**. Normally, gating information is generated by sub-networks and represented as binary masks, where non-zero values indicate the importance of each sampling location. Therefore, the gating masks induce sparsity in the data, thereby greatly reducing computational complexity. Spatially dynamic gating convolutions [180, 643] exploit the sparsity in the spatial domain and only execute conditionally on important selected regions. Likewise, depthwise dynamic gating convolutions [268, 598] account for the importance of different feature channels and selectively train only the relevant channels across various inputs. Consequently, combining spatial and depthwise gating information is possible for more generalized dynamic gating convolutions [381, 364]. Noted, deformable convolutions with modulation scalars can be categorized into the spatially dynamic gating convolution. 3) The last type is to directly use the output feature maps from other (sub-)networks as the kernel contents [305, 130]. Therefore, running the dynamic content convolutions over all locations of the input feature maps can be interpreted as collecting correlation information between two feature maps [460, 459]. Overall, dynamic convolution offers the most flexible design for convolution operations,

2.2. Developments

leveraging input-dependent adaptability to enhance model performance.

Other variants

Other variants include convolutions that emphasize application scope rather than improvements to the definition. We briefly list the following application directions of convolution operations:

Graph convolution is a generalized operation that extends the concept of convolution from Euclidean space to non-Euclidean space. Specifically, irregular data structures can be represented as graphs, where nodes represent entities and edges imply relationships. One of the most important advantages of graph structure is that it can efficiently represent complicated mutual correlations via an adjacency matrix. Therefore, graph convolution simplifies aggregating information from connected edges around each node and the learning of intuitive and scalable node representations at the graph level. Deep learning models that stack multiple graph convolution operators are called Graph Convolutional Networks (GCNs) [543]. GCNs can effectively discover inherent data structures and capture complex relationships among nodes, whereas traditional convolutional operations struggle to do so. GCNs have found applications in various fields, including social network analysis, drug discovery, and recommendation systems. The main challenges of GCNs are multiple folds: (a) Scalability, where large graphs result in high computational complexities. (b) Static nature that cannot adapt to evolving graph structures. (c) Limited expressiveness that hinders the capture of complex graph structures, especially degrading and over-smoothing node identifiability as graph convolution layers deepen. (d) Limited generalization that GCNs cannot transfer their information from one graph structure to another, especially between heterogeneous graphs.

Sparse Convolution is a specialized convolution operation designed to handle sparse data where most input values are zero efficiently. It allows computations to be performed only on non-zero data points, reducing computational and memory costs compared to traditional dense convolutional operations. Sparse convolution is typically implemented in three steps: (1) build an index system to store locations of non-zero values. (2) Build a RuleBook to indicate how to accumulate the values of pairwise multiplication between non-zero elements and corresponding kernel weights. (3) Execute all computations in parallel based on the instructions in the RuleBook. Therefore, sparse convolutions are often adopted for high-resolution images (i.e., medical & microscope images) or

3D data (i.e., point clouds). The main advantage of sparse convolutions is that they extract information from sparse data with high speed and low memory usage, since the overhead of the construction of the RuleBook can be ignored. SCNN [420] applied a two-stage decomposition over the channels and convolutional kernels and converted them into sparse matrix multiplication. However, the main disadvantage is their limited generalization, which imposes an ambiguous restriction on data sparsity. Therefore, the choice between sparse and traditional convolution depends on the specific application and the data characteristics.

Generalization to Higher Dimensions

The principles of 2D convolution naturally extend to higher dimensions, enabling deep learning models to process volumetric data such as 3D medical scans (CT, MRI), 3D/4D microscopy, video sequences, and point clouds. Let $X \in \mathbb{R}^{H \times W \times D \times C_{in}}$ denote a 3D input volume, where D denotes the depth (or temporal) dimension. A 3D convolution kernel is a 4D tensor $f_c \in \mathbb{R}^{k \times k \times k \times C_{in}}$ for each output channel c with kernel size k , and the sampling grid $\mathcal{G}$ expands to three spatial dimensions:

$$\mathcal{G}_d^k = \left\{ (d \cdot i, d \cdot j, d \cdot l) \mid i, j, l \in \left\{ -\left\lfloor \frac{k}{2} \right\rfloor, \dots, \left\lfloor \frac{k}{2} \right\rfloor \right\} \right\} \quad (2.5)$$

where $d \geq 1$ is the dilation rate (typically uniform across axes for isotropic data). The stride s and padding p are defined along all three dimensions. The coordinate- and size-mapping formulas are analogous. The output at location (x, y, z) is computed as:

$$Y(x, y, z, c_{out}) = \sum_{(\Delta x, \Delta y, \Delta z) \in \mathcal{G}_d^k} \sum_{c_{in}}^{C_{in}} X(x' + \Delta x, y' + \Delta y, z' + \Delta z, c_{in}) \cdot f_c(\Delta x, \Delta y, \Delta z, c_{in}) \quad (2.6)$$

3D CNNs have become instrumental in medical image analysis and video understanding. For example, 3D U-Net [117] extended the successful U-Net architecture to volumetric segmentation by replacing 2D convolutions with 3D ones, enabling the model to leverage inter-slice context, which is crucial for accurate organ or lesion delineation. In video action recognition, C3D [628] demonstrated that straightforward 3D convolutional networks can learn powerful spatiotemporal features directly from raw video clips. However, processing 3D data with standard 3D CNNs is often computationally expensive. Several efficient architectures [717] have been developed for efficiency. Sparse 3D Convolutions [420, 204] mined spatial sparsity in the high-dimensional data and

2.2. Developments

computed operations only on non-zero voxels, dramatically reducing computation. Beyond dense volumetric grids, irregular 3D data, such as point clouds and meshes, require specialized architectures because of their non-Euclidean structure. Instead of voxelizing point clouds, Point Cloud Networks directly processed unordered point sets. PointNet [525] and PointNet++ [526] used shared multilayer perceptrons (MLPs) and symmetric pooling functions to achieve permutation invariance. 3D Graph Neural Networks [327] modeled point clouds as graphs, where edges represent local neighborhoods, and message passing aggregates information. Moreover, separable 3D Convolutions [717], inspired by 2D depthwise separable convolutions, have factorized the 4D kernel into a 3D spatial convolution (per channel) followed by a $1 \times 1 \times 1$ pointwise convolution, reducing parameters and computation. Likewise, to handle irregular geometries in volumetric data, deformable convolutions [127] have been extended to 3D, allowing the sampling grid to adapt to the data shape.

3D computer vision tasks often require tailored loss functions. The Dice loss [476] is widely used in medical image segmentation to handle class imbalance. For point cloud reconstruction, Chamfer distance [169] and Earth Mover’s distance [557] measure the similarity between two point sets. In summary, extending deep learning to 3D and volumetric data introduces unique challenges in computational efficiency and representation. While 3D CNNs serve as a direct generalization of 2D convolutions to dense, grid-structured volumes, alternative architectures like GNNs and point-based networks are better suited to irregular 3D structures, enabling impactful applications across multiple domains. Future directions may include more efficient 3D attention mechanisms and hybrid models that combine the strengths of CNNs, GNNs, and Transformers for 3D data.

CNN Summary

CNNs are characterized by local receptive fields, a hierarchical structure, and static kernel parameters. These inductive biases grant them remarkable data efficiency and hardware-friendly computation. However, their fundamental limitations in modeling long-range interactions and adapting to complex geometric variations motivate the search for architectures with global receptive fields and dynamic aggregation—principles that form the foundation of the next architectural paradigm.

2.2.2 Visual Transformers

Transformer models originate from natural language processing (NLP) tasks and outperform recurrent neural networks (RNNs) [345] in modeling element-wise correlations within input sequences. Unlike the Long Short-Term Memory (LSTM) mechanism in RNNs, transformers can ignore spatial distances and directly model long-range pairwise interactions via self-attention. Owing to their strong representational capabilities, transformers have intrigued the vision community to adapt their architectures to computer vision tasks. The variants applied in computer vision have performed as well as, or better than, CNNs. To overcome the inherent locality and staticity of CNNs, Visual Transformers (ViTs) adopt a different design philosophy. By replacing convolution with a global self-attention mechanism, ViTs directly model global pairwise dependencies among all input tokens, without a hierarchical structure, enabling dynamic, content-dependent feature aggregation. As depicted in Figure 2.4, it consists of an encoder and a decoder with several transformer blocks of the same architecture. The encoder generates embeddings from the inputs, while the decoder uses all the embeddings and their incorporated contextual information to generate the output sequence. Each transformer block comprises a multi-head self-attention layer, a feed-forward neural network, a residual connection, and layer normalization. In the following, we describe how the shift from local/static to global/dynamic operations redefines architectural design while introducing new challenges in computational complexity and data efficiency.

Foundations of ViTs

In contrast to CNNs, a ViT uses a process called tokenization to partition an input image into a sequence of tokens. In addition, a basic ViT module comprises three components, with optional additional functions: multiple positional embeddings, an attention mechanism, and a normalization operation, along with residual connections and a Feed-Forward Network/Multilayer Perceptron (MLP).

Tokenization

Unlike CNNs, which directly process grid-structured data (e.g., images or videos), ViTs require tokenization to adapt self-attention mechanisms from NLP. Specifically, because treating individual pixels as tokens is computationally infeasible, ViTs partition the input image into patches. Each patch is then linearly projected into a vector embedding, and the sequence of these embeddings serves as the input to the transformer. Each

2.2. Developments

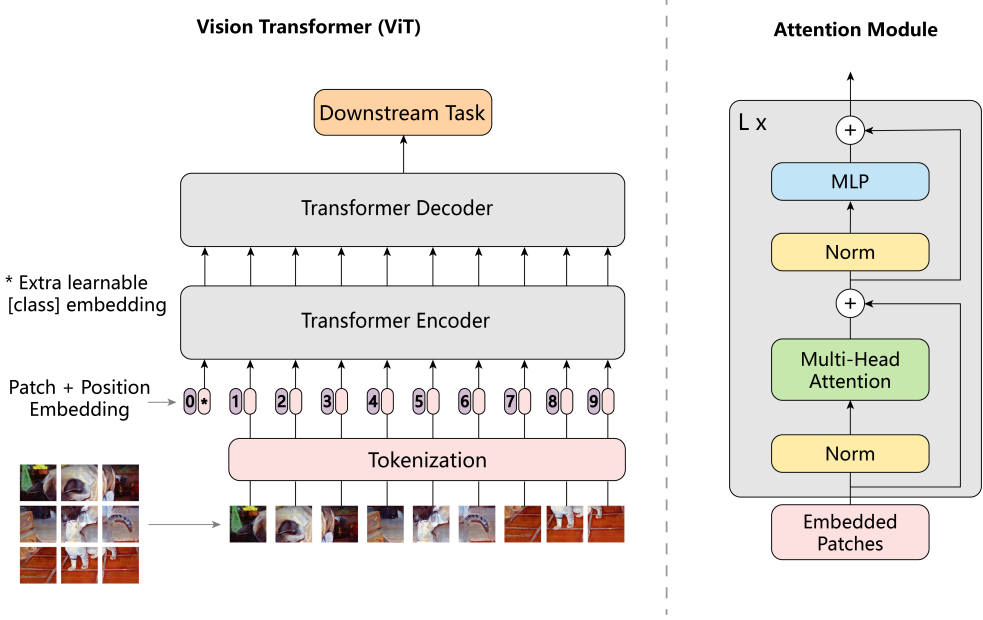


Figure 2.4: Schematic representation of a generalized framework for Visual Transformer (ViT) with L blocks.

patch embedding serves as a "super-node" of the input and contains information about all pixels in a group. Typically, each patch embedding is referred to as a visual token, and the conversion of grid-structured data into sequential tokens is called tokenization. The vision transformer [156], which is the origin of pure ViTs, partitions an input into fixed-sized non-overlapping patches, flattens each patch into a 1-dimensional vector, and projects all vectors onto a higher-dimensional embedding feature space using linear layers:

$$X_{patch}^{N \times D} = \text{tokenize}(I^{L \times C}) \quad (2.7)$$

where I is a flattened input with length L and feature channels C ; X_{patch} is a sequence of tokens with length N and feature channels D . Note that $N = \frac{L}{P}$ in which the L is divisible by the area of a patch P ; otherwise, the input image will be padded with zero elements.

Positional embedding

Because the self-attention mechanism is designed for sequential data, grid-structured data (e.g., images or videos) is flattened into 1D vectors, losing spatial structure. In

addition, the global pairwise correlation in self-attention is invariant to the position of each element. That means spatial information cannot be integrated into computation. Therefore, extra positional embeddings $X_{position}$ are added to the sequential input to retain spatial information:

$$X = X_{patch} + X_{position} \quad (2.8)$$

Different position embeddings fall into three categories:

Absolute position embedding (APE) is a vector that represents the unique absolute embeddings to the position of one element in a sentence [53]. It can use element embeddings to encode positional information.

Relative position embedding (RPE) technique is primarily used to incorporate relative position information into the attention mechanism [696, 115]. The RPE technique is based on the idea that the spatial relationships between patches carry more weight than their absolute parameters. The relative distance between patches determines the lookup process. Although the RPE technique can be applied to sequences of varying lengths, it may increase training and testing time.

Convolutional position embedding (CPE) methods take into account the grid structure of the input and employ convolutional operations to gather position information with zero-padding [693, 674, 289].

Self-attention mechanism

The self-attention mechanism is the primary operation in ViTs and plays a crucial role in explicitly modeling relationships among entities in sequential data. It calculates pairwise correlations between each location and all others, aggregates all long-term dependencies into a single entity, and represents it as global contextual information [642]. The self-attention mechanism transforms the input sequence(s) into three different embedding vectors, namely query, key, and value, as named in NLP configurations. We introduce a more generalized definition of the self-attention mechanism, which can benefit in explaining the developments of other variants and drawing connections with convolutions:

$$Y_i = \frac{1}{C(X)} \sum_{\forall j} f\left(g_q(X_i), g_k(X_j)\right) g_v(X_j) \quad (2.9)$$

Where i is the index of the visual token in the output sequence, j is the index that enumerates all possible positions. $X \in \mathbb{R}^{L \times C}$ is the input sequence and $Y \in \mathbb{R}^{L \times C}$

2.2. Developments

is the output sequence. Note that L is the length of the sequences, and C is the number of channels of each vector. A pairwise function f computes a correlation scalar between i and all j . The unary functions g_q, g_k, g_v transform the input sequence X into embedding vectors of query, key, and value, respectively. The pairwise function $f(\cdot)$ is normalized by a factor $\mathcal{C}(X)$.

The standard single-head self-attention mechanism can be instantiated as follows:

$$g_q(X) = \text{Linear}(X), \quad g_k(X) = \text{Linear}(X), \quad g_v(X) = \text{Linear}(X) \quad (2.10)$$

$$\frac{1}{\mathcal{C}(X)} f\left(g_q(X), g_k(X)\right) = \text{Softmax}\left(\frac{g_q(X) \cdot g_k(X)^T}{\sqrt{d_k}}\right)$$

where $\text{Linear}(\cdot)$ denotes linear projection function, $\text{Softmax}(\cdot)$ denotes the softmax normalization function, and $\frac{1}{\sqrt{d_k}}$ is the scaling factor with d_k as feature channels of the key vector. Then, the standard single-head self-attention mechanism can be reformulated into a popular formula:

$$q_i = g_q(X_i), \quad k_j = g_k(X_j), \quad v_j = g_v(X_j), \quad (2.11)$$

$$\text{Attention}(q_i, K^T, V) = \sum_{\forall j} \text{Softmax}\left(\frac{q_i \cdot k_j^T}{\sqrt{d_k}}\right) \cdot v_j$$

where q_i, k_j , and v_j are the embedding vectors of query, key, and value at location i and j , respectively. In addition, K^T and V are the transposed key and value embedding matrices, respectively.

It is limited by the capacity of single-head self-attention to attend to all important positions. Multi-head self-attention (MHSA) [642] is utilized as the real component in ViTs. Like group convolution, MHSA assumes several mutually independent feature groups and decomposes the embedding spaces of the query, key, and value into multiple sub-spaces. Applying single-head self-attention across sub-spaces in parallel can not only reduce computational complexity within groups but also capture a diverse range of complex interactions across groups. MHSA comprises multiple self-attention blocks, each of which projects the concatenation of its outputs into the output space. The separation enables MHSA to focus on multiple portions and to capture the relationships

among them effectively:

$$\text{Head}_i^k = \text{Attention}(q_i^k, K^k, V^k) \quad (2.12)$$

$$\text{MHSA}(q_i, K, V) = \text{Concat}(\text{Head}_i^1, \text{Head}_i^2, \dots, \text{Head}_i^k) \cdot W^{Cat}$$

where Head_i^k is the k -th output of group-wise self-attention at position i , and W^{Cat} represents the learnable weights to transform the concatenation of all heads into the final output.

Other components

Feed-Forward Network (FFN) is applied after the self-attention modules to obtain more complex attributes from the input sequences. It consists of two linear project functions and a nonlinear activation function

$$\text{FFN}(X) = W_2 \cdot \sigma(W_1 X) \quad (2.13)$$

where W_1 and W_2 are the weight matrices of the two linear project layers, and $\sigma(\cdot)$ represents the non-linear activation functions, such as GELU [251]. The dimensionality of the hidden layer is $d_h = 2048$ by default.

Normalization with residual connection consists of a residual connection and a normalization operation after each self-attention module. This design can stabilize information flow and improve performance. The output of these operations can be described as

$$\text{Norm}(X + \text{Attention}(X)) \quad (2.14)$$

where X is the input of the self-attention module, and Norm is a normalization layer, like layer-normalization [19]. A variant pre-normalization [23, 664, 156] is also widely used. It reorders the computation and inserts normalization into the residual connection before self-attention.

Characteristic of Transformers

Besides CNNs, ViTs differ in their architectural design due to their self-attention mechanism. Because self-attention is designed for sequential data, we assume the input data is a sequence of length L .

2.2. Developments

1. **Global receptive field:** Each visual token in the input sequence is connected to every other token, regardless of distance. Globality can explicitly capture long-range dependencies and accumulate pairwise relationships to build global contextual features. However, the adopted global receptive fields, rather than sampling, result in a computational complexity of $\mathcal{O}(L^2)$ that quadratically increases with the length of the input sequence.
2. **Flatten structure:** Since the self-attention mechanism directly adopts a global receptive field, there is no need to use a hierarchy of layers to expand receptive scopes gradually. Instead, the ViT stacks multiple self-attention modules with the same structure to introduce nonlinearity and improve the model’s capacity to capture complex relationships in the data. Notably, the global receptive fields reduce the inductive bias of locality in ViTs that require feature representations with small spatial variances.
3. **Dynamic Parameter:** The aggregation weights in the self-attention mechanism are dynamic because they are computed as pairwise correlations and vary depending on the input. The dynamic parameters can be interpreted as gating information that adaptively highlights significant content and captures complex relationships from the input. Unlike CNNs, which enforce translation equivariance with static kernels, dynamic weights in self-attention allow the model to prioritize input-dependent contexts rather than rigid spatial invariance.

In summary, the self-attention mechanism is closely related to dynamic gating convolution 2.2.1. Within a global receptive field, gating information is computed as pairwise correlations based on the input content and can highlight all important locations for each query element. Compared to CNNs, ViTs impose fewer inductive biases and, in theory, can process a broader range of data.

Improvements of the Attention Mechanisms

Given that images involve more dimensions (e.g., spatial dimension), noise, and redundant modality, they are believed to be more difficult to capture robust long-range dependencies than sequential data. Following NLP conventions, ViTs use tokenization to group pixels into a small number of visual tokens, each representing a semantic concept in an image. By capturing long-range dependencies among tokens, self-attention mechanisms offer significant flexibility in feature representation learning, as their dynamic weights adapt to each input image. However, the adaptability comes with four limitations: (1) the quadratic computational complexity increases with the

sizes of images; (2) the data-hungry requires a large scale dataset (e.g., JFT-300M [599] dataset with 300 million images) for pre-train otherwise impedes fine-tuning of downstream tasks with smaller datasets; (3) fixed patch size cannot provide a plausible granularity to describe the true semantic parts in different images at various scales. (4) long-range dependencies at different scales cannot freely flow across feature maps at different scales. To address one or more limitations, a series of variants of the Vision Transformer [156] have been proposed to improve attention mechanisms and modify the architectures.

Reduce computational complexity

The quadratic complexity of self-attention hinders its applicability to long sequences, i.e., high-resolution images. The core bottleneck stems from the heavy matrix multiplication used to compute pairwise dependencies. Two main categories simplify the self-attention computation process.

Similar to the core idea of separable convolution, the *matrix decomposition* factorizes the large matrix of self-attention into a product of several sub-matrices with lower ranks. CaiT [627] proposed class attention (CA) that computes attention between class tokens and patch tokens. CA reduces the complexity and is used with self-attention to build Transformers with very deep architectures. Geng et al. [193] proposed the "Hamburger" module, which reformulates matrix decomposition as a low-rank reconstruction process. Based on the low-rank matrix, it can efficiently compute dependencies among different tokens. Linformer [670] proposed a different decomposition scheme. Based on the distributional Johnson–Lindenstrauss lemma [418], it first projects the key and value matrices to low-rank matrices and then computes attention between the query and the low-rank matrices. Performer [113] proposed a novel Fast Attention Via positive Orthogonal Random features (FAVOR+) approach that can efficiently model kernelizable attention mechanisms beyond softmax. Katharopoulos et al. [318] proposed linearized attention, which rearranges the original self-attention as a linear dot product between a query vector and two intermediate matrices. The advantage is that both intermediate matrices are computed once and reused for all query vectors. Nyströmformer [722] adapted the Nyström method to approximate standard self-attention. XCiT [6] proposed cross-covariance attention that computes attention across feature channels instead of tokens. Cross-covariance attention has linear complexity with respect to the feature dimension rather than the number of tokens. Singularformer [700] used neural networks to learn the singular

2.2. Developments

value decomposition process and decompose the attention matrix into the product of three matrices with lower ranks. SeT [604] factorized self-attention into pixel-wise local and patch-wise global attention, reducing the computational cost while retaining the information at different granularities. CSWin [154] decomposed self-attention into horizontal and vertical strips in parallel that form a cross-shaped window, reducing the computational cost.

Another direction seeks to find sparsity patterns in the matrix and approximate self-attention with *sparse attention*. Given an attention matrix $A \in \mathbb{R}^{L \times L}$ of sequence length L , sparse attention only computes pairwise interactions on a subset $\mathcal{S} = \{(i, j) | A(i, j) \neq 0\}$, reducing the quadratic complexity $O(L^2)$ to near-linear. Sparsity enables processing of high-resolution images with large L , making sparse attention preferable for tasks such as segmentation or generation on large inputs. Therefore, there has been great enthusiasm to explore different sparsity patterns $\mathcal{S}$. Child et al. [110] introduced sparse factorizations of the attention matrix and proposed fast attention kernels for training. Yun [788] argued that a sparse attention model can universally approximate any sequence-to-sequence function. The results showed that sparse Transformers with $\mathcal{O}(L)$ connections per attention layer can approximate the same function class as the dense model with L^2 connections. The Sinkhorn Transformer [618] is based on differentiable sorting of internal representations and computes quasi-global attention using only local windows, thereby improving the memory efficiency of the self-attention module. Reformer [332] observed that the softmax function in self-attention will be dominated by a small subset of keys closest to the query in computation and defined these sparse sets by finding nearest neighbors in high-dimensional spaces. Based on locality-sensitive hashing (LSH), Reformer proposed a simplified sparse attention mechanism, called LSH attention, to reduce computational complexity. Likewise, KVT [663] adopted the K-NN clustering method to find top- k sparse similar tokens. Inspired by graph theory, many works [116, 590, 791] have formulated attention mechanisms as graphs and have adopted various graph sparsification methods to reduce computational complexity. Chung et al. [116] adopted random graph theory and approximated the dense graph with sparse graphs. Spielman et al. [590] based on spectral similarity of graph Laplacians and proved that every graph has a spectral sparsifier of nearly linear size. A construction algorithm was proposed to build a dense graph from sparse graphs. BIGBIRD [791] used a directed graph and proposed sparse attention, where each query attends over a small set of randomly chosen keys. Inspired by deformable convolution, DAT [707] emphasized the importance of key-value pairs in a data-dependent manner and focused on relevant patches, thereby capturing informative features from sparse

sets.

Alleviate data-hungry

ViTs yield modest results when trained on mid-sized datasets such as ImageNet [138], with 1.2 million images, achieving lower accuracy than CNNs with comparable parameter counts. Because transformers lack some inductive biases inherent to CNNs—such as translation equivariance and locality—they do not generalize well when trained on insufficient data. However, Han et al. [225] found that training ViTs on large datasets (14 million to 300 million images) overcame the limitations of inductive bias, resulting in excellent fine-tuning performance on downstream tasks even with fewer data samples. Many variants are being explored to improve data efficiency in ViTs trained on smaller datasets. Many ViT architectures use knowledge distillation to transfer knowledge from a larger network (i.e., the teacher model) to a smaller network (i.e., the student model) [219]. The DeiT [626] is the first work to demonstrate that transformers can be trained on mid-sized datasets (e.g., 1.2 million ImageNet examples) in a relatively short training time. Using a pre-trained Vision Transformer as the teacher model, DeiT learned to capture the most important pattern in tokens and produced similar results with fewer parameters. DINO [55] introduced self-supervised learning and proposed a form of self-distillation with no labels. The key difference is that the teacher model was built from past iterations of the student network and was updated using an exponential moving average of the student weights, i.e., serving as a momentum encoder. To avoid model collapse, the student model was trained to predict the teacher model’s output after a centering and sharpening layer. TaT [413] argued that the one-to-one spatial matching strategy in a patch-based manner cannot always be maintained in distillation learning because different architectures may cause semantic inconsistency between patches from the teacher and the student model, respectively. Instead, TaT proposed a one-to-all matching knowledge distillation pipeline that allows each teacher patch to teach the entire set of student features dynamically. To avoid intractable computational complexity arising from correlations in large feature maps, TaT proposed a two-step hierarchical distillation pipeline, i.e., patch-group and anchor-point distillation, to reduce computational burden and achieve decent results. TinyViT [697] proposed a fast distillation method in the pre-training stage and built a family of tiny yet efficient ViTs suitable for devices with limited resources. In addition, TransKD [432] is the first framework for distilling knowledge from large-scale semantic segmentation transformers. By focusing on comprehensive distillation across different knowledge sources within

2.2. Developments

the transformer-specific building, TransKD distills information from both feature maps and patch embeddings.

Hierarchy of dynamic patches

In ViTs, an image is first divided into a grid of patches, projected to a linear embedding, and treated as a sequence of tokens. Each token corresponds to a semantic region in an image. Since images are highly complex, with abundant detail and color information, the granularity of patch division is insufficient to extract semantic regions at different scales and locations. In addition, a better image partition can reduce computational redundancy from irrelevant patches in self-attention. Consequently, many researchers have explored different architectures for flexible patch generation. Two main strategies to dynamically adjust patch sizes are (1) Patch-fused Self-Attention and (2) Pooling-based Self-Attention.

Patch-fused Self-Attention fuses fixed-sized patches into windows with dynamic sizes to better cover semantic parts. Moreover, the merging strategy progressively reduces the patch sequence length and constructs multi-scale representations. T2T-ViT [780] proposed a Tokens-to-Token (T2T) transformation that applied a sliding window algorithm to combine neighboring tokens into a single token with overlapping recursively. T2T-ViT adopted an efficient backbone and reduced the parameter count by half due to shorter token sequences, enabling training from scratch on the mid-sized ImageNet dataset. Similar to Network In Network (NIN) [507], TNT-ViT [226] proposed a Transformer in Transformer (TNT) architecture that adopts a hierarchical patching scheme. In each TNT module, smaller patches (visual words) were fused into larger patches (visual sentences) without overlapping. Within the TNT architecture, local correlations among visual words are captured within each visual sentence, while global correlations are extracted simultaneously across visual sentences. DPT [103] proposed a Deformable Patch (DePatch) module, which learns to adaptively split the images into patches in a deformable way with learnable positions and scales. The method can generate dynamic patches that capture image semantics based on their content. Swin-ViT [441] proposed a shifted-window architecture. This framework generated windows from patches at the same layer without overlapping and computed local self-attention within each window. To gradually capture long-range dependencies across layers, Swin-ViT shifted window partitioning within successive blocks and built connections between windows across layers. Several works [286, 174, 544, 817, 587, 440] improved the shifted window strategy. Shuffle Transformer [286, 174] utilizes the spatial shuffle operation instead of

shifted window partitioning to allow cross-window connections. DW-ViT [544] explored the upper limit of the effect of window settings on model performance with dynamic multi-scale windows; VSA [817] employed window regression modules to predict the size and location of the target windows from data; CSWinTT [587] utilized a cyclic shifting window strategy at multi-scales.

Pooling-based Self-Attention, inspired by hierarchical structures used in CNNs, gradually reduces the sizes of feature maps and expands the effective coverage of patches. PVT [673] introduced a progressive shrinking strategy and generated a pyramid of patches without overlapping. The pyramid structure can represent semantic parts in images at various scales. PVT proposed spatial-reduction self-attention (SRA) to capture long-term dependencies at low-resolution key-value pairs. RegionViT [65] combined a pyramid structure with an efficient regional-to-local (R2L) self-attention mechanism. Like TNT-ViT, RegionViT adopted regional self-attention (RSA) to learn global information, while local self-attention (LSA) was used to learn local features within regions. Like PVT, PiT [252] also used a pooling layer to reduce token length. Based on pyramid structures, Twins [114] and CAT [408] alternately performed local and global attention across the pyramid layers. MS-ViL [816] proposed a mechanism that combines local attention with global memory. The local attention only captured short-term dependencies among neighboring tokens, whereas global memory propagated information over long distances, as in Longformer [34] in NLP tasks. Focal Transformer [747] sampled all surrounding tokens for each query on every pyramid layer and computed local self-attention among concatenated neighboring tokens across multi-scales. CrossFormer [676] progressively enlarged group size along the pyramid for patch embedding and decomposed self-attention into long/short distance attentions, capturing local and global visual cues.

In summary, hierarchical tokenization exponentially reduces sequence length, thereby lowering the cost of subsequent attention layers. The reduction is because complexity quadratically scales with token count. Consequently, Hierarchical tokenization is especially beneficial for high-resolution inputs and tasks requiring multi-scale reasoning, such as object detection and semantic segmentation, providing a better accuracy-efficiency trade-off than fixed-resolution transformers.

Attention across scales

Although ViTs can propagate global context information sequentially through transformer decoder layers, this information cannot freely flow across different scales. Many

2.2. Developments

approaches applied multiple ViTs in their architectures to extract long-term dependencies from multi-scale features and to directly interact with global information across any pair of scales. CrossVit [66] proposed a ViT architecture having dual branches and defined a novel type of attention mechanism (called cross-attention) that enables information transfer at different scales. The cross-attention assigned the class tokens to query and used the concatenation of class and patch tokens as the key and value. Therefore, the simplified attention mechanism computed only long-term dependencies between class tokens and image patch tokens, and removed the most expensive matrix computation involving the image patch tokens, thereby achieving linear computational complexity. Similarly, MMViT [439] adapted dual branches and introduced Cross-Pooling-Attention to fuse information at different views. Moreover, Dual-ViT [763] extended cross-attention with two pathways: a semantic pathway that compressed token embeddings into global semantic queries and a pixel pathway that mined finer local details. The global semantic queries served as useful prior information, guiding the refinement of the feature maps at the pixel level.

Efficient Alternatives Beyond Attention Mechanisms

Although many improvements aim to mitigate the quadratic scaling of attention mechanisms through approximations, Visual Transformers still struggle to achieve a trade-off between efficiency and performance as sequence length increases. This fundamental bottleneck has spurred intense research into deeper-level architectural alternatives. Among these, structured state space models (SSMs), rooted in classical control theory, have recently emerged as a theoretically grounded option that entirely avoids attention, favoring recurrent state transitions with linear-time complexity. Mamba [207], an outstanding representative, proposes a selective, data-dependent SSM that challenges the necessity of global attention for high-performance sequence modeling and offers a compelling, linearly scalable alternative. Although challenges remain, such as optimizing training stability and further validating scaling at the very largest scales, Mamba has firmly established itself as a cornerstone in the landscape of post-attention efficient foundation models. Its emergence underscores that the future of sequence modeling may not be a choice between attention or alternatives, but an intelligent synthesis of complementary computational principles.

Selective SSMs as Linear-Time Alternatives

At its heart, Mamba departs from the Transformer’s fixed, input-agnostic context

mixing. It is built upon a structured, parameterized SSM that maps a 1-D input sequence $x(t)$ to a latent state $h(t)$ and output $y(t)$ through continuous-time differential equations, which are then discretized for digital computation. In contrast to attention’s dynamic weighting, the conceptual counterpoint of Mamba is selectivity, which converts the matrices governing dynamics and input projection into functions of the current input x_t . Specifically, attention computes pairwise relevance scores across all tokens. At the same time, the selective SSM modulates its internal state transition dynamics and input sensitivity at each step based on the incoming token. Selectivity allows the model to propagate or discard information contextually throughout the sequence, emitting a “hidden state” that carries a compressed, relevant history. Crucially, thanks to a hardware-aware parallel algorithm inspired by parallel prefix scans, this inherently sequential process can be trained efficiently. During inference, Mamba operates with linear-time complexity $O(L)$ and a constant memory footprint relative to sequence length L , as it only needs to maintain a fixed-size hidden state, eliminating the massive KV cache of autoregressive Transformers.

Theoretical Unification and Hybrid Models

Based on a profound theoretical connection, there is an established link between SSMs and Transformers through the structured state space duality (SSD) framework [133]. It shows that both attention and SSM-based layers can be derived as different instantiations of computations on semi-separable matrices. This mathematical duality reveals that Mamba and attention are not opposed but lie on a unified continuum of sequence-mixing primitives, offering a rigorous lens for understanding and potentially unifying these seemingly disparate approaches. Therefore, a prominent research direction is the exploration of hybrid models that integrate Mamba blocks with attention blocks. Architectures such as Jamba [404] demonstrate that such hybrids can outperform pure Transformers or pure Mamba models with equivalent total compute, leveraging Mamba for efficient long-context processing and attention for precise, localized reasoning. Its success has catalyzed a broader reassessment of the fundamentals of sequence modeling, leading to practical hybrid systems and deeper theoretical unification.

Scaling Towards Universal Large Foundation Models

Apart from the previous research trajectory of reducing ViT parameter scales for computational efficiency, a significant yet opposite paradigm has emerged: embracing scale and complexity to pursue universal functions and extreme generalization. The

2.2. Developments

extreme generalization has led to the rise of large-scale foundation models, which are pre-trained on web-scale data and exhibit remarkable zero-shot or few-shot adaptability across a wide range of downstream tasks, effectively functioning as general-purpose cognitive systems. Their development represents a pivotal shift from task-specific models to universal intelligence across multiple modalities.

Large-Scale Visual Foundation Models

As a self-supervised learning framework built on ViTs, the DINO family has been a significant contributor to the quest for universal visual representations. Analogous to DINO [55], which uses large-scale, curated data from diverse sources, DINOv2 [502] produces generalized visual features that work across image distributions and tasks without fine-tuning, making them exceptionally effective drop-in replacements for traditional backbones in diverse visual applications. Furthermore, DINOv3 [576] introduced the pivotal Gram anchoring and effectively addressed the known yet unsolved issue of dense feature maps degrading during long training schedules. Therefore, DINOv3 could produce high-quality, dense features for various vision tasks, surpassing previous self- and weakly supervised foundation models. In summary, the DINO family demonstrates that with carefully designed self-supervised objectives and large-scale, curated data, ViTs can learn a comprehensive model of visual appearance and structure without labor-intensive manual annotations.

In parallel with advances in visual representation, the Segment Anything Model (SAM) [330], and its successor, SAM2 [537], stand as landmark achievements in universal segmentation frameworks. SAM introduced a promptable segmentation model trained on an unprecedented dataset of over 1 billion masks. Its core innovation lies in decoupling the heavy image encoder (a ViT-H) from a fast prompt encoder/mask decoder, demonstrating zero-shot transfer to new image distributions and tasks. SAM2 further advances promptable visual segmentation for images and videos. The model is equipped with a streaming memory that stores information about the object and its previous interactions, allowing it to generate masked predictions throughout the video and effectively correct them based on the object’s memory context from previously observed frames. The SAM family exemplifies how scaling data, model capacity, and task diversity (through prompting) can yield a model with unparalleled generalization in spatial understanding. For more information about large-scale visual foundation models, please refer to the survey [17].

Multiple-Modal Models

While visual foundation models have dramatically advanced computer vision tasks, their capabilities are inherently bounded by the visual modality. They often require extensive task-specific labeling and struggle with open-world semantic understanding and generalization beyond predefined categories. This limitation has catalyzed a pivotal shift towards multi-modal foundation models that integrate and align information from diverse data sources. By learning joint representations across modalities, these models harness the complementary strengths and rich supervisory signals present in uncured, web-scale multi-modal data.

Due to the limitations of visual foundation models that require a closed set of predefined categories, Visual-Language Foundation Models (VLMs) have transcended this boundary by anchoring visual representation learning in the open and semantically rich modality of natural language. The foundation of this paradigm lies in learning a joint embedding space in which VLMs can pull together semantically matched images and texts while pushing apart mismatched pairs. Therefore, the mainstream research on VLM focuses on learning more robust and versatile image-text correlations. The pioneering work CLIP [530] established the efficacy of large-scale contrastive pre-training for zero-shot transfer. To mitigate data noise and scale limitations, ALIGN [303] demonstrated training on billions of noisy pairs, while DeCLIP [386] and UniCL [748] improved data efficiency through self-supervision and label integration. For a finer-grained understanding, FILIP [761] and GLIP [376] introduced region-word alignment, while FLAVA [578] integrated masks for flexible guidance. Architectural innovation has evolved from dual-encoder towers (CLIP) to unified single-tower designs [630, 297] to enable efficient cross-modal fusion. Furthermore, models like DetCLIP [760] and FIBER [157] extended VLMs from image-level to region-level tasks, paving the way for open-vocabulary object detection and segmentation. This trajectory illustrates a clear path from learning global image-text associations to mastering fine-grained, task-agnostic visual-semantic correspondence. For more information about large-scale visual-language foundation models, please refer to the survey [810].

Building on the foundational architecture and training paradigms established by VLMs, the multi-modal learning paradigm has been successfully extended to a wide range of domains, demonstrating remarkable versatility. In audio-visual learning, models align spectrograms with textual descriptions to perform tasks such as sound event localization and captioning. The video-language domain leverages temporal alignment between video frames and transcripts or instructions to enable video question

2.2. Developments

answering and action recognition. In affective computing, multi-modal sentiment analysis models fuse visual (facial expressions), acoustic (prosody), and linguistic (text) cues to infer emotional states. The medical imaging field integrates radiology images with clinical reports to build foundation models for automated diagnosis and report generation. Furthermore, robotics utilizes models that combine visual perception, language instructions, and sensorimotor data for task planning and human-robot interaction. These expansions consistently rely on core VLM principles—cross-modal alignment, contrastive or generative pre-training on paired data, and zero-shot generalization—thereby demonstrating the transferability of the paradigm and establishing a unified framework for multimodal intelligence.

ViT Summary

While ViTs excel at capturing global context through dynamic attention, their lack of inherent spatial inductive biases, such as locality and translation equivariance, and their high computational complexity present significant barriers for efficient and robust visual modeling. The bottleneck provides compelling motivation for hybrid architectures that integrate the complementary strengths of convolutional locality and attentional globality into a unified, synergistic framework.

2.2.3 Hybrid Networks

Although ViTs can efficiently capture long-term dependencies among visual tokens, they cannot flexibly represent local semantic parts at multiple scales. Except for the aforementioned improvements 2.2.2, using a simple linear layer remains less straightforward than convolution for extracting local features with spatial geometric information within each visual token. Besides, because ViTs lack some inductive biases inherent to CNNs, such as translation equivariance and locality, they require large datasets (e.g., JFT-300M) to overcome these biases and generalize. Empirically, a self-attention mechanism in early layers works similarly to a standard convolution, showing locality and sparsity in pattern extraction [122, 312]. These limitations have motivated applying convolution operations to ViTs, especially in the early layers, to reduce computational redundancy. On the contrary, CNNs can be trained on mid-sized datasets (e.g., ImageNet). Still, they cannot efficiently capture long-range interactions because of their local receptive fields. To combine the strengths of CNNs and ViTs, considerable effort has been devoted to hybridizing them. Hybrid architectures are a deliberate synthesis that combines the computational efficiency and spatial priors of CNNs with

the expressive power and global context modeling of ViTs. The advantage is achieved not by mere concatenation, but through deep architectural integrations that embed convolutional operations within attention mechanisms (Con-Trans) or augment convolutional feature maps with attention-based feature recalibration (Attn-CNNs). The subsection explores how hybrid models navigate trade-offs between locality/globality and between static/dynamic aggregation to achieve state-of-the-art performance across diverse vision tasks.

Attention enhanced CNNs

Inspired by non-local means operations [45], global or local self-attention mechanisms are reformulated as special instances of dynamic convolution. Given a feature map, non-local operations compute the response at a position as a weighted sum of the features at non-local positions in the feature map, regardless of the distance between them. Therefore, equipped with the capability to model long-range interactions in a feed-forward fashion, CNN architectures have demonstrated superiority for feature extraction by expanding the receptive field and focusing on different spatial locations based on their importance, as depicted in Figure 2.5.

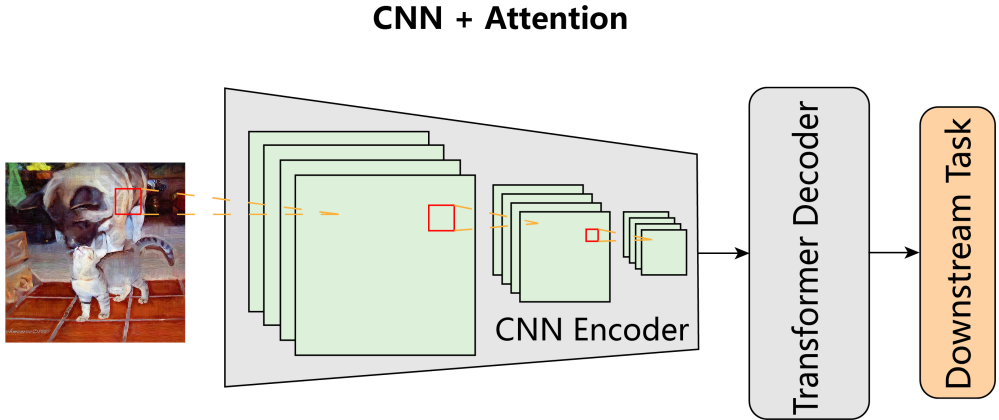


Figure 2.5: Schematic of a generalized framework for the hybrid framework of CNN with an attention mechanism.

Adapting Attention for CNNs

Attention mechanisms, as instances of dynamic convolution, can adjust their aggregation weights in response to changes in the visual input based on pixel similarities. Many Attn-CNNs applied attention modules to higher-dimensional visual data than to sequential

2.2. Developments

data for adaptation. RANet [652] stacked attention modules with residual units to generate attention-aware features adaptively varying as layers go deeper. SENet [268] focused on the channel relationship and proposed a "Squeeze-and-Excitation" (SE) block that adaptively captured interactions between channels. Non-local Network [678] introduced a non-local operation as a generalization of spatial attention, computing a weighted sum of features across all positions. CBAM [691] combined spatial and channel attention modules sequentially, while BAM [508] and DANet [183] combined them in parallel. A²-Nets [98] proposed a double attention block that aggregated and propagated informative global context from the spatial-temporal representative space of input images and videos. AA-Net [33] introduced two-dimensional relative position encoding [570] into CNNs to maintain translation equivariance for handling images and applied MHSA in parallel to a standard convolution operation as an augmented operation. GCNet [52] proposed a three-step general framework that unified the Non-Local network and SENet to model global context. Ramachandran et al. [533] proposed a local self-attention layer that aggregates only within a specific window centered on each pixel and can be used in place of convolutional operations in CNNs. Zhao et al. [836] proposed two types of vector attention, pairwise and patchwise self-attention, to replace conventional attention with scalar weights. Both attention mechanisms learned vector weights for the spatial and channel dimensions simultaneously. Pairwise attention was a set operator that computed interactions within a local scope of aggregation. In contrast, the patchwise self-attention was an instance of the dynamic convolution operator that adaptively generated the local aggregation weights. FEDER [234] proposed a learnable wavelet transformation and applied spatial/channel attention and convolution to mine different frequency bands. Likewise, FADNet [461] captured long-term dependencies within the frequency domain and converted them into query vectors that contained global input-dependent contexts.

Reducing Computational Complexity

Although self-attention enables modeling of global contextual information, it is both memory- and computation-intensive. The dense attention map computation has a high complexity of $O(N^2)$, where N denotes the number of pixels in feature maps. To reduce this computational burden, many Attn-CNNs have explored matrix factorization or simplification to improve efficiency. One stream of research employed sparsity or locality properties to simplify self-attention mechanisms. CcNet [287] proposed the criss-cross attention module for each pixel position that approximated the dense matrix

multiplication with two simple vector multiplications along one criss-cross path. By recurrent criss-cross attention, each pixel position can capture long-range interactions from other pixels with a complexity of $O(2\sqrt{N})$. LRNet [267] proposed local relation networks to compute attention in a local window with lower complexity adaptively, and introduced geometric priors into CNNs. Other work has explored approximating the self-attention mechanism using simpler matrix multiplication. Similar to the core idea behind cross-attention (see 2.2.2), most efforts have focused on adapting cross-attention to achieve linear complexity for CNNs, extracting high-level semantic information between class tokens and pixels in feature maps. EncNet [806] introduced channel attention to highlight significant feature channels via a context-encoding embedding. The context vector was the output of cross-attention, which used the Euclidean distance rather than the cosine distance to compute similarity. EMANet [379] interpreted the cross-attention as a Mixture of Gaussian models solved by the Expectation-maximization algorithm. It updated static class tokens and re-estimated feature maps to remove unnecessary noise. On the contrary, OCRNet [784] introduced object-region representations as dynamic class tokens and obtained pixel-region relations via cross-attention to reconstruct object-level contextual representations. ECA-Net [665] is dedicated to reducing the complexity of the channel attention mechanism and proposes an efficient channel attention module with fewer parameters. Moreover, ResNeSt [807] introduced the Split-Attention block to implement channel attention across different groups of feature channels within multi-path feature representations, resulting in a simple, unified computational block with fewer parameters.

Convolution enhanced ViTs

Convolutional operations help mitigate several inherent drawbacks of visual transformers, including the lack of inductive biases (e.g., reliance on large-scale pretraining data), high computational complexity, and difficulty in capturing local low-level context and positional embeddings. Incorporating convolutions into visual transformers enables these hybrid models to leverage the strengths of both architectures, leading to improved robustness, efficiency, and performance across various computer vision tasks, as depicted in Figure 2.6.

Enhancing Tokenization

VTs [692] integrated convolutional operations into ViTs to enhance tokenization and proposed an efficient VT-FPN module to extract multi-scale feature maps with fewer

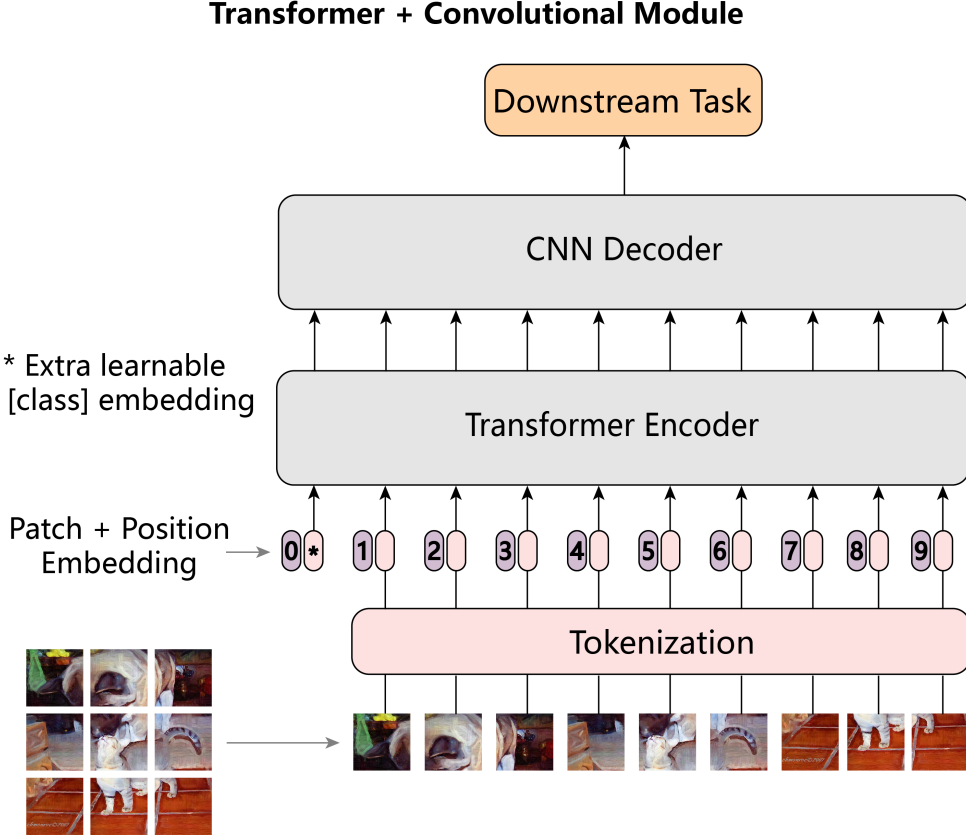


Figure 2.6: A generalized framework for the hybrid framework of CNN with an attention mechanism.

FLOPs than conventional FPN. Xiao et al. [711] substituted the original tokenization process with several stacked 3×3 kernels with a stride 2, facilitating the stability and generalization of ViT for downstream tasks. CvT [693] replaced Linear projections with convolutions to capture the spatial structure and low-level details for tokenization and adopted a hierarchical structure to progressively reduce the token length, imitating the impact of spatial down-sampling as in CNNs. CeiT [779] integrated the advantages of CNNs in extracting low-level features and enforcing locality into an image-to-token module that extracts patches from the generated local context. CCT [232] introduced a convolutional tokenization process and proposed a sequence pooling scheme to utilize multi-scale information. In addition, CCT can be trained from scratch on smaller datasets, e.g., CIFAR10, a remarkable property not possible with traditional ViTs. CPVT [115] utilized depth-wise convolutions to define Positional Encoding Generators,

incorporating positional information into patch embeddings. ResT [818] replaced the patch embedding as a stack of overlapping convolution operations with a stride on the token map and incorporated depthwise convolutions and adaptive position encoding to tackle varying image sizes. MPVit [359] integrated convolution operations with overlap to generate patch embeddings at different scales and proposed a multi-path architecture to interact global and local features across scales. CrowdFormer [750] proposed an overlap patching transformer block to capture the global contextual information in a top-down manner to extract overlapping patches at different scales.

Integrating locality into Attention

LocalViT [393] introduced depthwise convolution into Feed-Forward networks and brought locality into self-attention mechanisms. LeViT [203] adopted a pyramid architecture with pooling, similar to the LeNet [350], and explored the convolutional behavior in the transformer patch projection layer. BoTNet [592] adapted ResNet Bottleneck Modules that replaced the spatial convolutions with global self-attention. CMT [213] integrated convolution operations into the FFN of each transformer block to promote the correlation among neighboring tokens. ConViT [161] designed a dual path that attached a parallel convolution branch to a transformer branch for soft integration of convolutional inductive biases via a Gated Positional Self-Attention (GPSA). Refiner [851] observed the over-smoothing issue of MHSA in deep architectures and directly integrated convolution kernels into the self-attention mechanism, thereby augmenting local patterns in the attention matrices and establishing a coaction between global context aggregation (via self-attention) and local context modeling (via convolution). NesT [835] adopted a hierarchy of blocks to represent images at different scales and aggregated blocks across the hierarchy using convolution and pooling operations, thereby communicating information among blocks. Visformer [102] performed a comparative study between convolution operations and self-attention modules. It revealed that replacing self-attention modules with residual convolutional blocks [716] in the early stages can effectively capture both spatial and local features while requiring fewer FLOPs. CoAtNet [131] introduced relative attention that combined depthwise convolution with self-attention and vertically stacked convolution and attention in the shallow layers. Likewise, ACMix [506] addressed an underlying relation between convolution operations and self-attention mechanisms and combined them in parallel as a lightweight aggregation operation. MOBILEViT [469] combined the strengths of CNNs and ViTs to build a lightweight and low-latency network for mobile vision

2.3. Developments

tasks. Conformer [515] adopted a concurrent structure consisting of one CNN and a transformer branch. It proposed a Feature Coupling Unit (FCU) to integrate local features into global representations.

Reducing Computation Complexity

ConvBERT [311] simplified self-attention calculation by proposing a new module - called span-based dynamic convolution - that combines the fully connected and convolutional layers. Long-Short Range Attention (LSRA) [703] reduced the computational cost of the self-attention mechanism by using convolutional operations to capture local information. Efficient-Attention [571] decomposed the pairwise matrix multiplication with two channel-attention functions and converted the matrix production into vector multiplication. Coat [737] devised a conv-attentional mechanism that approximated the softmax attention matrix as a product of two linear functions, inspired by LambdaNets [32]. Perceiver [292] combined cross-attention with latent transformers that mapped arbitrary high-dimensional input onto a fixed-sized latent space, reducing computational complexity. Maxvit [632] introduced a Multi-Axis attention mechanism that integrates local blocked attention with dilated global attention. This design enables global-local spatial interactions across arbitrary input resolutions while maintaining only linear computational complexity. PVT-v2 [674] proposed a linear spatial reduction attention module that reduced the spatial dimension of feature maps to a fixed size via average pooling, enjoying linear computational and memory costs like a convolutional layer. EfficientViT [47] proposed lightweight multi-scale attention to handle high-resolution input via ReLU-based global attention [318]. SwiftFormer [567] introduced a novel, efficient additive attention mechanism to replace the quadratic matrix multiplication operations with linear element-wise multiplication.

2.2.4 Development Trend Summary

The progression from CNNs to ViTs to Hybrids illustrates a clear conceptual trajectory: starting with strong, locally biased priors, moving towards flexible, globally aware computation, and culminating in architectures that strategically balance both. This evolution is governed by the continuous interactions between the core mathematical operations of convolution (local, static) and attention (global, dynamic). The resulting hybrid models are not merely a combination of two components, but a new class of architectures where locality and globality, static and dynamic, coexist and cooperate, defining the current frontier in visual representation learning.

2.3 Other Techniques

Some stand-alone techniques are used in deep learning to improve automation and expand the range of applications. Neural Architecture Search (NAS) is a technique that reduces the difficulty of designing efficient and general architectures for given complex datasets. Specifically, NAS formalized model architectures as a set of learnable hyperparameters and automated hyperparameter tuning via dedicated training processes. In addition, Neural Architecture Compression (NAC) explored patterns of feature vectors and mapped them to similar vectors using fewer bits. The compression aims to deploy powerful deep learning models in resource-limited scenarios, such as mobile and embedded devices, without significantly degrading performance. Moreover, Knowledge Distillation (KD) is used to transfer knowledge from source models to target models. The knowledge transfer enables the easy generation of new models as an ensemble of an enormous number of off-the-shelf models. The last aspect is to alleviate the hunger for annotated data. Despite the almost unlimited growth in data, the corresponding high-quality manual annotations lag far behind due to the high cost of labeling, especially in domains such as biology, medicine, and chemistry. Insufficient annotations have hindered the rapid development of supervised deep learning models. Therefore, self-supervised learning (SSL) or unsupervised learning (USL) was used to improve the robustness of feature representations without requiring additional supervision (i.e., expensive manual annotation). Moreover, Diffusion models provide a powerful generative framework for capturing data distributions and producing high-quality samples.

2.3.1 Neural Architecture Searching

Neural Architecture Search (NAS) is a deep learning technique that automates the optimal design of networks in architecture and engineering. Traditionally, designing neural network architectures has required specialized knowledge and extensive trial and error. Instead, NAS formulates the architecture search space as a set of possible combinations of layers, skip connections, operations, and hyperparameters. The mathematical core is to relax the search space so that continuous weights represent the choice between operations. Then the output is a mixed operation over the candidates. Therefore, a joint optimization of architecture choices and network parameters is performed via gradient descent. A final discrete architecture is derived by retaining the operation with the highest weight. Moreover, NAS proposes different searching strategies to explore the search space. It employs performance-estimation strategies to identify optimized archi-

2.3. Techniques

tectures for visual tasks, such as image classification, segmentation, and object detection [163]. Early NAS works explored various search spaces defined by units, operations, and their connections for CNNs [873, 874, 422, 875, 531, 611, 826, 742, 148, 470, 739]. Later, NAS further explored the design of more efficient and scalable ViT architectures suitable for vision tasks [82, 64, 391, 596] by combining tokenization and attention blocks. Recent advancements in NAS have identified optimal combinations of CNNs and ViTs, leading to more powerful and efficient models [216, 579, 363, 777, 468, 771]. Nevertheless, NAS requires a delicate trade-off among search space expressiveness, search strategy cost, and estimation accuracy. While gradient-based methods are fast, they may favor narrow, continuous spaces. Evolutionary strategies are more flexible but computationally expensive, requiring thousands of model evaluations.

2.3.2 Neural Architecture Compression

In this subsection, we exclude some techniques (e.g., matrix decomposition and knowledge distillation) that, though applicable to model compression, are not dedicated solutions due to their broader applicability in other applications [498]. Instead, we mainly introduce two dedicated directions for model compression: pruning and quantization.

Pruning

Pruning techniques aim to sparsify neural networks by removing redundant components with minimal impact on performance [736, 249]. Currently, pruning is the most common method of reducing network complexity and overfitting [120]. The pruning methods began with unstructured pruning in the early stages for CNNs. Inspired by Dropout and DropConnect operations, early unstructured pruning methods removed neurons or connections whose magnitude or importance ranking fell below a predefined threshold. The variants [266, 775, 1, 464, 801, 605, 15] alternated between weight updates and pruning by dynamically adjusting thresholds. More recent structured pruning techniques have introduced different levels based on pruning granularity. From fine to coarse, they are kernel-level, block-level, and filter-level pruning [341]. The kernel-level pruning [11, 411, 495, 463, 147] refers to removing entire convolutional filters, thereby achieving channel-wise sparsity. Additionally, block-level pruning [348, 392, 430, 805] aims to trim multiple filters grouped by their similar inherent sparsity patterns. The filter-level pruning [452, 250, 473, 616, 821] focused on removing entire layers of blocks based on more sophisticated rules. In ViTs, pruning techniques

trim unimportant attention heads to improve efficiency during deployment. The redundancy of attention heads can define the importance scores to estimate the influence of each head [475, 168, 522, 263, 863, 614].

Quantization

Quantization aims to lower the number of bits required for storing weights or intermediate features by mapping representations to a small set of discrete values [639, 721, 759, 372]. Early quantization methods have been explored for convolutional neural networks. They can be divided into different categories: vector quantization that adopted code-book and quantization code for restoring the original data [298, 198, 695, 354], multi-bit quantization that substituted floating-point weights with low bit(s) directly [146, 664, 539, 396, 538, 741, 688], and hashing quantization that adopted a random weight sharing strategy to reduce the model redundancy substantially [91, 271, 781]. Recently, there has been growing interest in quantization schemes for ViTs [615], aiming to compress the inputs and weights by low-bit representations [523, 575, 442, 660].

Comparative Analysis

Pruning reduces parameter count and FLOPs, directly improving model size and theoretical computation. Quantization reduces memory bandwidth and enables integer acceleration on dedicated hardware. In practice, pruning is favored for aggressive FLOP reduction (e.g., edge devices with strict latency limits), while quantization is essential for maximizing throughput on commodity integer processors (e.g., mobile CPUs). They might be combined sequentially as a pipeline, i.e., pruned first and then quantized, to maximize efficiency.

2.3.3 Self-supervised/unsupervised learning

Self-supervised learning (SSL) [259, 645, 637] and unsupervised learning (USL) [568, 621, 54, 324, 149, 731] aim to leverage abundant unlabeled visual data [136] for learning robust representations. These approaches have been widely adopted in CNNs and ViTs to enhance generalization. USL and SSL share similar feature-learning pipelines. Therefore, we focus more on the more popular SSL approaches. In the early stages, SSL was applied to CNNs to learn rich visual representations for downstream tasks [511, 377, 857, 584, 792]. Later, contrastive learning, one of the most important frameworks in SSL, has been prominent and developed by VAEs [326, 546, 150, 582], SimCLR [87],

2.3. Techniques

SwAV [640, 54], BYOL [206, 514], and MoCos [244, 93]. Recent works [81, 534, 366, 399, 789, 16, 84, 720, 479] also explored self-supervised schemes for ViTs. DeiT [626] introduced a distillation token-based approach to SSL that enables efficient training on smaller datasets. Likewise, DINO [55] proposed a self-distillation scheme that integrates knowledge distillation with SwAV, focusing on learning representations via clustering without requiring labels. BEiT [26] applied random masking of image patches (tokens) from BERT in NLP [320] to ViTs. Moreover, MAE [242] applied a masking technique, forcing ViTs to learn contextual relationships among patches by reconstructing masked image patches. In addition, some works that were originally developed for CNNs have been adapted for ViTs. MoCo v3 [95] adopted momentum-based contrastive learning to obtain robust image representations between different augmented views of the same image for ViTs using a memory bank. MoBY [719] combined MoCo v2 and BYOL for Swin Transformers [441]. PeCo [155] improved masked image modeling by optimizing image tokenization with a perceptual codebook. We briefly summarize the effectiveness of SSL Methods across domains. The efficacy of self-supervised learning (SSL) paradigms varies significantly across application domains, dictated by data characteristics and task objectives. In natural image understanding, contrastive learning methods (e.g., SimCLR, MoCo) excel by learning invariant representations from geometric and photometric augmentations, providing strong features for downstream classification and detection. In contrast, masked image modeling (MIM) methods (e.g., MAE, BEiT), which learn to reconstruct masked content, have shown superior performance on dense prediction tasks such as segmentation because they inherently model patch-level context and semantics. In medical imaging, where annotations are scarce and textures are subtle, MIM often outperforms contrastive learning because its reconstruction objective forces the model to learn fine-grained anatomical structures, thereby improving transfer to segmentation and detection tasks. In remote sensing, multimodal SSL that leverages geographic and temporal metadata alongside images has proven more effective than unimodal approaches. Meanwhile, for video data, contrastive methods that exploit temporal consistency (e.g., by contrasting frames from the same video clip) are dominant for action recognition. This domain-specific performance highlights that no single SSL method is universally optimal. Hence, selecting or designing a pretext task must align with the inherent structure and annotation constraints of the target data.

2.3.4 Knowledge Distillation

Knowledge Distillation (KD) is a machine learning technique that transfers knowledge from a source model (the teacher model) to a destination model (the student model). The general process of KD involves training the student network to mimic the teacher's intermediate or final outputs based on a loss function that measures mutual similarity. The transfer facilitates effective knowledge propagation from the teacher to the student. KD has been widely applied across various domains, including model compression and ensemble learning. As for model compression, KD can derive smaller but more efficient student models out of larger and more complicated teacher models for CNNs [46, 18, 549, 254, 86, 454, 620, 145, 1, 854, 859, 212, 302] and ViTs [485, 675, 477, 626, 304]. As for ensemble learning, KD can fuse information from multiple teacher models into a single but more powerful student model [185, 493, 13, 215, 7, 501, 61, 385, 317, 500]. Generally, KD is a versatile bridge between other techniques. It can recover from accuracy loss caused by aggressive pruning or quantization, distill knowledge from ensemble foundation models via NAS into a single model, or serve as the final fine-tuning step for a model trained via SSL.

2.3.5 Diffusion Models

The diffusion Model (DM) methodology uses CNNs and ViTs as building blocks to approximate data distributions. It defines a framework that learns to generate new data by adding and removing noise from the inputs. Specifically, diffusion models are iterative generative models based on *Markov chains*. Every chain begins with true data and ends with pure noise. Besides, each state corresponds to one noisy intermediate data point with increasing noise levels. A diffusion model consists of two stages: a forward diffusion stage and a reverse diffusion stage. In the forward diffusion stage, the input data is converted into pure noise by gradually adding Gaussian noise. In the reverse stage, a diffusion model learns to progressively recover the original input data from the pure noise using CNNs or ViTs as backbone architectures. The fundamental paradigm shift from Generative Adversarial Networks (GANs) to DMs represents a mathematical evolution in generative modeling. GANs [202], introduced by Goodfellow et al., formulate generation as a min-max game between a generator and a discriminator. The generator learns to produce realistic samples, while the discriminator attempts to distinguish real from generated data. The training objective seeks a Nash equilibrium that minimizes the Jensen–Shannon divergence between the

2.3. Techniques

real and generated distributions [201]. However, GANs rely on an indirect adversarial mechanism: the generator and discriminator jointly learn implicit representations, and convergence is achieved when the two distributions become indistinguishable. The core idea is to use two mutually close approximations to the unknown real distribution. Despite their success, GANs suffer from well-documented limitations, including training instability (e.g., vanishing gradients when the discriminator becomes too strong), mode collapse (where the generator produces only a limited variety of samples), and difficulty in capturing complex multi-modal distributions. Diffusion models [124] emerged from principles of non-equilibrium thermodynamics and adopt a fundamentally different paradigm. Instead of adversarial training, they employ a denoising objective grounded in rigorous probabilistic theory. Diffusion models can be formulated within the framework of stochastic differential equations (SDEs) [586], unifying popular variants such as DDPM [260], DDIM [583], and NCSNs [585] as special cases. By optimizing a well-behaved surrogate likelihood bound (e.g., the variational lower bound or a simplified mean-squared error), DMs achieve stable training and comprehensive coverage of the data distribution without mode collapse, directly addressing the core weaknesses of GANs. Originally developed for data generation, DMs [483] have demonstrated performance comparable to that of GANs and VAEs [326]. Recently, they have been adapted for diverse computer vision tasks (e.g., image segmentation, object detection), offering enhanced robustness by modeling additive noise. We will elaborate on the applications of DMs in section 2.4. For more information on applications, please refer to the survey [237].

2.3.6 Deep Unfolding Techniques

Iterative optimization methods play a central role in deep learning, serving as the mathematical foundation for training, inference, and estimation. Despite interpretability and theoretical guarantees, conventional iterative optimization algorithms suffer substantial computational latency due to the complexity of hyperparameter tuning. Conversely, deep learning models offer powerful data-driven modeling capabilities but lack the structure, transparency, and efficiency needed for optimization-driven inference. Therefore, deep unfolding has recently emerged as a remedy to bridge these two paradigms by systematically transforming iterative optimization algorithms into structured, trainable ML architectures [574]. The hybrid framework, also known as the Deep Unfolding Network (DUN), can integrate prior domain knowledge into network design. Therefore, DUNs need relatively few trainable parameters in theory. The evolution of DUNs

demonstrates their role as a versatile meta-technique. Initially, DUNs were primarily incorporated into CNNs to tackle low-level tasks, such as compressed sensing [335] and image denoising [733], leveraging their efficiency in local feature extraction. Later, CNN-based DUNs have attracted attention in high-level vision, such as concealed object segmentation [240] and image super-resolution [811]. A significant advancement occurred as researchers began to integrate ViTs (or Vision-Language models) into the unfolding framework to capture long-range dependencies, which is crucial for tasks such as image restoration [796] and video processing [134]. The most recent trend focuses on designing hybrid modules within the DUN skeleton, combining convolutional layers for local processing and attention blocks for global interaction into a single, optimized pipeline [846]. This progression underscores DUNs as a flexible framework that can harness and enhance the advantages of underlying architectural components (CNNs, ViTs, or their hybrids) and other techniques (e.g., Diffusion [238]) to improve performance and interpretability across diverse inverse problems. Generally, DUNs are a prime example of hybridizing domain knowledge with data-driven learning. Their interpretable structure makes them particularly attractive for trustworthy deployment in scientific and medical domains. For a comprehensive exploration of deep unfolding methodologies, please refer to the detailed survey [135, 574].

2.3.7 Section Summary

Some techniques have different intrinsic characteristics. For instance, SSL and DM represent different points on the spectrum of data utilization: SSL condenses information from real data into representations, while DMs expand the data distribution through synthesis. In practice, these techniques are rarely used in isolation. A common pipeline might involve: 1) SSL is the workhorse for pre-training general-purpose representations, especially in label-scarce domains. 2) DM can use SSL-enhanced representations to synthesize high-fidelity samples for data augmentation or to provide robust generative priors as fake labels. 3) NAS can discover an efficient backbone for a target task. 4) Compressing the model via pruning and quantization for deployment, and 5) Employing KD to fine-tune the compressed model and recover performance from the original large model as a teacher. 6) DUNs inject strong domain priors into optimization for structure transparency. Understanding their complementary strengths and formal interconnections is crucial for building efficient, robust, and deployable vision systems.

2.4 Applications

Due to their ability to automatically fit complex data, deep learning models have been widely applied to computer vision tasks and have achieved good performance. In this section, we introduce the numerous applications of different model types and reveal trends in architecture selection based on dataset properties.

2.4.1 Classification

Classification is the most fundamental task in computer vision. It aims to predict class labels at multiple levels: image-level (class presence), pixel-level (semantic segmentation), and instance-level (instance segmentation). As depicted in Figure 2.7, Image-level classification identifies the existence of classes but lacks spatial localization. Semantic segmentation refines this by partitioning the image into disjoint regions corresponding to semantic classes. Instance segmentation further distinguishes individual instances within each semantic class.

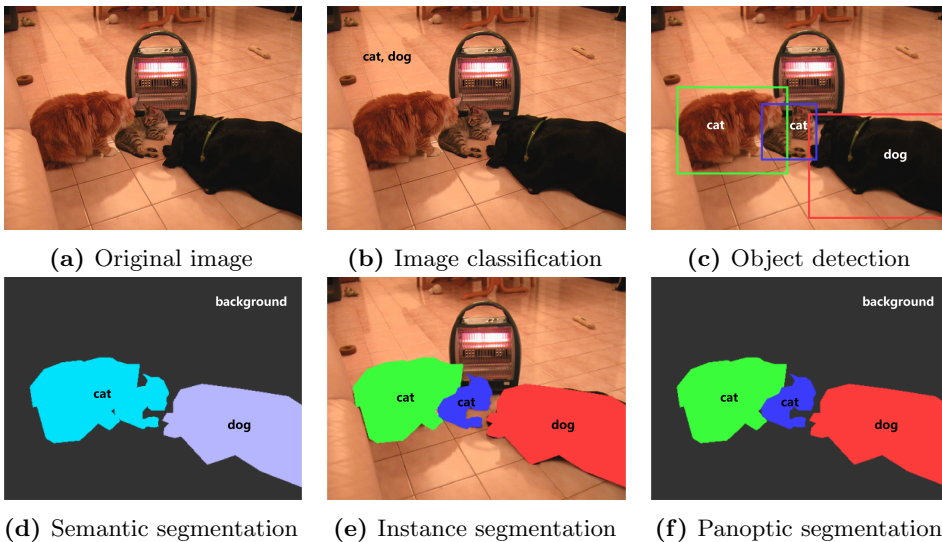


Figure 2.7: Classification hierarchy from coarse-grained to fine-grained inference: classification at image level, detection or localization at part/region level, semantic segmentation, instance segmentation, and panoptic segmentation all at pixel level.

Image-level classification

Large-scale datasets like ImageNet [138] have driven the widespread adoption of CNNs for image classification. Their ability to automatically extract complex information from data has surpassed traditional image analysis algorithms. Inspired by LeNet-5 [351], AlexNet [340] was a pioneering CNN-based classification architecture that achieved superior performance on ImageNet. The automation sparked the enthusiasm of subsequent researchers for designing more complex architectures, driven by the importance of stacking additional layers. VGG [577] designed deeper networks by stacking more convolutional and fully connected layers sequentially, gaining improvement in classification tasks. At the same time, GoogLeNet [608] was designed with wider networks using an inception module inspired by NiN [410]. Since replacing all fully connected layers, GoogLeNet has outperformed VGG with around one-eighth the number of parameters. Several improvements to the inception module [610, 288] advanced the accuracy of image-level classification. As depth increases, CNNs can suffer from vanishing or exploding gradients. ResNet [248] introduced a skip connection to interact with arbitrary pairs of layers at different depths and enabled training a deep CNN up to 1000 layers. The core innovation lies in parameter-free residual functions, which bypass intermediate layers to stabilize gradient flow and mitigate vanishing gradients. The residual function is almost necessary for deep learning models and has been improved by the following research [607, 827, 99, 716, 277, 758]. In addition, capsule networks [559, 256, 220, 2] adopted steerable convolutions and introduced equivariance to geometric transformations in image classification.

Later, visual transformers became popular alternatives to CNNs for image-level classification due to their self-attention mechanisms, which capture long-range dependencies. Vision Transformer [156] proposed the first ViT framework for image classification. The successive ViTs focus on tokenization improvement, multi-scale feature extraction, self-attention simplification, and training efficiency enhancement (see more details in section 2.2.2). Most recently, hybrid architectures have gained popularity by combining the strengths of both CNNs and ViTs [815, 384]. There are two main branches, i.e., enhancing CNNs with the attention (Attn-CNNs) mechanism and Transformers with Convolution (Con-Trans). Attn-CNNs introduce self-attention mechanisms into hierarchical feature representations to capture global context information at different scales and model generic object relations to understand complex scenes (refer to subsection 2.2.3). Con-Trans incorporates convolutional operations into ViTs to build multi-scale features and alleviate the data-hungry problem during

2.4. Techniques

training (refer to subsection 2.2.3). Some significant work on image classification tasks is summarized in the Table 2.1, including CNN-based, ViT-based, and hybrid models.

Pixel-wise classification

Pixel-wise classification is more conceptually difficult than image-level classification because disjoint spatial partitions are needed to locate the present entities. Depending on the partition granularity, semantic and instance segmentation are defined. More specifically, semantic segmentation assigns each semantic class to a single disjoint partition, which may contain multiple instances. Instance segmentation pushes the concept of segmentation one step further, identifying clues to individual semantic objects and separating disjoint semantic partitions into non-overlapping instance regions at finer granularity. However, many semantic classes (uncountable regions such as sky, sea, grass, etc.) cannot be further divided into individual entities, unlike other semantic classes (countable entities such as cat, dog, person, etc.). Therefore, panoptic segmentation [329] unifies uncountable "stuff" classes and countable "thing" classes into one segmentation framework to deal with all possible semantic entities. CNNs have developed specialized frameworks for segmentation tasks, whereas ViTs have enabled general-purpose frameworks for all pixel-level classification tasks. Therefore, we will introduce task-specific and generic architectures separately. We organize task-wise models in the Table 2.2 for the semantic, instance, and panoptic segmentation, respectively. Moreover, some representative universal models are listed in Table 2.3.

A. Semantic Segmentation

The success of CNNs in image-level classification motivated researchers to explore the capabilities of pixel-level labeling on semantic regions. The key advantage of CNNs, which enables them to outperform traditional methods, is their ability to learn appropriate feature representations from a particular dataset in an end-to-end manner, rather than relying on handcrafted features that require extensive domain expertise and manual parameter tuning to adapt to the specific scenario. However, CNNs for image-level classification struggle to achieve good partitions since their sequential/linear architectures are not designed to extract spatial information at different scales. Therefore, CNNs for semantic segmentation [672] have to find rich details about locations and appearances to infer the pixel-wise affiliation at different depths. With residual skip connections, the following research focuses on designing architectures that integrate local and global information across multiple scales. FCN

Model			Dataset	Year	Publication	Top-1 Acc(%)	GFLOPs	#Param (M)
CNN		AlexNet[340]	ImageNet-1K	2012	NeurIPS	63.3	1.45	62
		VGG[577]		2015	ICLR	75.6	15.5	144
		GoogLeNet[608]		2015	CVPR	68.3	1.5	6.8
		Inception-v3[610]		2016	CVPR	77.2	11.4	23.8
		ResNet-50[248]		2016	CVPR	76.1	4.1	25.6
		ResNet-101[248]		2016	CVPR	77.4	7.9	44.7
		Inception-v4[607]		2017	AAAI	80.1	13.4	55.8
		PolyNet[827]		2017	CVPR	81.3	20.8	94.7
		DPN[99]		2017	NeurIPS	80.7	11.7	61.7
		ResNeXt[716]		2017	CVPR	79.6	8.0	83.7
		DenseNets[277]		2017	CVPR	77.8	7.7	28.9
		CliqueNet[758]		2018	CVPR	76.0	8.5	14.4
		MobileNetV3[264]		2019	ICCV	75.2	0.22	5.4
		Capsule-Net[559]	CIFAR-10	2017	NeurIPS	92.1	0.23	4.2
		Capsule-EM[256]		2018	ICLR	88.5	0.17	0.8
		CapsNet[220]		2019	NeurIPS	92.2	0.14	3.2
		Star-caps[2]		2019	NeurIPS	91.2	0.51	0.3
Image Classification	ViT	Vision Transformer[156]	ImageNet-1K	2021	ICLR	79.8	17.6	86.6
		CaiT-S[627]		2021	ICCV	83.3	13.9	48.0
		XCiT-L[6]		2021	NeurIPS	86	417.9	189
		DeiT-B[626]		2021	ICML	81.8	17.6	86.6
		DINO[55]		2021	ICCV	78.2	17.6	85
		Swin-B[441]		2021	ICCV	83.3	15.4	88.0
		Performer[113]		2021	ICLR	76.5	10.2	87
		PiT-B[252]		2021	ICCV	84	12.5	73.8
		CrossVit[66]		2021	ICCV	82.8	9.5	44.3
		SeT-33[604]		2022	PAMI	83.3	7.7	40.3
		KVT-PB[663]		2022	ECCV	82.6	12.5	74
		DAT-B[707]		2022	CVPR	84.8	22.6	100
		TaT[413]		2022	CVPR	72.4	2.7	13.8
		DW-B[544]		2022	CVPR	83.8	17.0	91
		VSA[817]		2022	ECCV	83.9	16.0	88
		CrossFormer++[676]		2023	PAMI	85.0	21.8	96
		Dual-ViT[763]		2023	PAMI	85.7	18.0	73
Hybrid		SENet-R50[268]	ImageNet-1K	2018	CVPR	78.4	4.4	30.5
		FANet[533]		2019	NeurIPS	77.6	7.2	18.0
		SAN[836]		2020	CVPR	78.2	3.3	20.5
		LeViT[203]		2021	ICCV	82.5	2.3	39.4
		ConViT[161]		2021	ICML	82.5	30	152
		ResNeSt-50[807]		2022	CVPR	81.1	5.39	27.5
		MPViT-B[359]		2022	CVPR	84.3	16.4	74.8
		CMT-L[213]		2022	CVPR	84.8	19.5	74.7
		Swin-ACMix-S[506]		2022	CVPR	83.5	9.0	51
		Dilate-B[312]		2023	TMM	85.6	31.1	48
		Conformer-B[515]		2023	PAMI	84.1	46.6	83.3
		SwiftFormer-L3[567]		2023	ICCV	83.0	4.0	28.5
		EfficientViT-B2[47]		2023	ICCV	82.7	2.1	24
		SwiftFormer-L3[567]		2023	CVPR	83.0	4.0	28.5

Table 2.1: A comparative analysis between different models on Image-level classification in terms of parameter complexity and Top-1 (%) Accuracy on ImageNet-1K. For a direct comparison, we consider models with input of size 224×225 .

2.4. Techniques

[444] is a milestone in semantic segmentation, enabling CNNs to be trained end-to-end. It replaced the fully connected layers with 1×1 convolutional layers, inspired by NIN [410], to generate pixel-wise predictions. Additionally, it introduced deconvolution [795, 793] as an up-sampling operation to learn how to restore spatial information from hierarchical feature maps, rather than relying on fixed interpolation methods. However, FCN adopted complicated models (AlexNet, VGG, GoogLeNet, ResNet, etc.) as encoders for feature extraction but used a very simple decoder for inference, limiting its performance. The following research explored integrating feature maps from multiple paths into decoders to capture object location and appearance cues across different scales. U-Net [551] and its variants [20, 510, 63, 529, 299, 10, 151] employ a ladder-shaped architecture that facilitates bilateral information flow from encoder layers to their corresponding decoder counterparts. This design effectively combines low-level local cues with high-level global context. To merge information from multiple paths, feature pyramid structures [532, 555, 36, 68, 406] adopted different down-sampling and up-sampling strategies to obtain rich information flows more flexibly, like dilated convolution [770, 661, 510, 79] expands receptive fields with dilation rates to collect statistics from wider scope at the cost of memory consumption, and spatial pyramid pooling [435, 79, 837, 21, 80, 694, 657, 706] to capture contextual information from different receptive field sizes for semantic regions of varying scales. The multi-path structures allow the insertion of pixel-wise loss functions at arbitrary depths for deep supervision. In addition, traditional probabilistic models, like Conditional Random Fields [338, 77, 845, 58, 78], random walk [35, 309], were adapted into deep learning to capture more representative features to support pixel-level inference. Most recently, CNNs [656] integrated with knowledge distillation techniques to enhance segmentation performance.

Some ViTs have been specifically applied to semantic segmentation and use self-attention to model long-range pixel interactions. The ViTs [847, 712, 595, 6, 671, 51, 800] follow the standard ViTs that divide an input image into patches by a tokenization process and generate better feature representations with self-attention. Then, an additional semantic segmentation module is required to produce pixel-wise semantic masks. In addition, many hybrid models have achieved impressive performance on semantic segmentation. Attn-CNNs [785, 678, 769, 806, 287, 184, 494] adopt an attention mechanism to capture long-range context and predict semantic masks. Con-Trans [73, 592, 632, 674, 47, 314, 294] replace the patch tokenization process with CNNs to better locate meaningful semantic regions, and use self-attention to obtain enhanced semantic feature maps.

B. Instance Segmentation

Unlike semantic segmentation, instance segmentation requires finer-grained location information to divide semantic regions into individual entities further. The purpose is to reason about spatial relationships among various instances to better understand the scene. Currently, it remains challenging for CNNs to represent each entity with location and semantic information. Like object detection, instance segmentation can be divided into one-stage and two-stage methods. As for two-stage methods, CNNs developed instance segmentation upon object detection approaches, which detect and locally segment sequentially: SDS [229], MCG [12], SharpMask [517], MultiPath [790], Instance-FCN [125], Mask R-CNN [245], PANet [433], MaskLab [76], Mask Scoring R-CNN [284], Cascade Mask R-CNN [50], HTC [74], BlendMask [70], CondInst [622] and Refinemask [804]. As for one-stage instance segmentation, CNNs build grid structures and use each grid point to represent nearby potential instances: YOLACT [39], TensorMask [94], SOLOs [679, 680], PointRend [331]. Moreover, Deep Snake [513] offers a different perspective on anchor-free instance segmentation by adapting active shape models to deep learning. Few transformer-based approaches have focused on instance segmentation rather than generic segmentation. QueryInst [178] adopts shared queries to represent all possible instances within the dataset and uses cross-attention to find instance-wise responses. In addition, some CNNs with attention mechanisms offer novel perspectives on modeling long-range object relations. They enhanced instance segmentation tasks [360, 516, 707, 319, 506, 861] with an attention mechanism to highlight informative channels or locations, thereby improving the representation of potential candidate entities.

C. Panoptic Segmentation

Panoptic segmentation [329] further extends the partition criteria and introduces two auxiliary classes, "things" and "stuff" tags, corresponding to "countable" and "uncountable" concepts in natural language. The added intermediate classes divide the original semantic classes into two groups, and panoptic segmentation employs two selection branches: one for semantic segmentation (handling 'stuff' classes) and another for instance segmentation (handling 'thing' classes). This approach unifies semantic and instance segmentation tasks within a joint framework. Therefore, panoptic segmentation must define flexible architectures that can select between semantic and instance segmentation to assign each pixel a semantic label or an instance ID based on a given auxiliary "things" or "stuff" tag. Initially, panoptic segmentation models were

2.4. Techniques

heavily based on CNNs, leveraging their ability to extract hierarchical spatial features. It is challenging for CNNs to jointly represent the concepts of "stuff" and "things". The early CNNs [519, 520, 708], such as Panoptic FPN [328], UPSNet [724], DeeperLab [752], AdaptIS [580], Panoptic-DeepLab [104], EfficientPS [481], and LPSNet [262], are ensembles of existing instance segmentation (like Mask R-CNN) and semantic segmentation architectures (like DeepLab) and fuse their results directly. Later, CNNs adopted unified architectures that concentrated on determining the most effective ways to partition "things" and "stuff" together. DR1Mask [69] proposed a novel dynamic convolutional module called dynamic rank-1 convolution to merge high-level context information with low-level detailed features. Panoptic FCNs [390, 389] generated kernel weights for each entity that fused dual information from position and kernel heads. Then, each generated kernel can locate and predict both "things" and "stuff" entities directly on high-resolution feature maps from the encoder. In addition, ViTs have adopted convolution operations for panoptic segmentation [88, 22, 877]. Panoptic Segformer [400] decoupled queries into "stuff" and "things" groups and inserted extra location information into "things" queries only. Then, it proposed a unified mask decoder to predict panoptic segmentation for two query groups. kMaX-DeepLab [774] replaced the computationally intensive computation of key-query correlation with a simple k-means clustering, thereby improving the efficiency of panoptic segmentation. Panoptic-partformer[382, 383] formulated panoptic segmentation and part segmentation as a unified mask prediction and classification problem. Moreover, hybrid architectures have attracted research interest in panoptic segmentation [382, 281]. CNNs can integrate attention mechanisms to model long-range object relations. Axial-DeepLab [655] introduced position-sensitive self-attention into CNNs and designed axial attention to reduce computational complexity. EPSNet [60] introduced cross-layer attention to generate dynamic prototypes to represent different entities and predict semantic and instance masks separately. Max-DeepLab [654] was the first end-to-end model. It proposed path-transformer architectures and directly predicted class-labeled masks for "stuff" and "things" entities. CMT-DeepLab [772] built upon the Clustering Mask Transformer and viewed both "stuff" and "things" entities as cluster centers, generating pixel affiliation with cross-attention mechanisms. Mask-DINO[369] extended DINO [808] with a mask branch and unified panoptic segmentation with object detection.

D. Universal Image Segmentation

Developing a generic framework to partition images into multiple perceptual levels

Model			Dataset	Year	Publication	metric (%)	GFLOPs	#Params (M)
Semantic Segmentation (mIoU)	CNN	FCN[444]	ADE20K	2015	CVPR	29.4	178	134
		U-Net[551]		2015	MICCAI	33.9	65	31
		RWN[35]		2017	CVPR	37.2	210	141
		G-CRF[58]		2017	ICCV	37	92	37
		DeepLab-v3+[78]		2018	PAMI	44.1	255	63
		HRNet-v2[657]		2020	PAMI	49.2	695	65.9
	ViT	Light-Ham[193]		2021	ICLR	51.5	71.8	61.1
		Segmenter[595]		2021	ICCV	53.6	380	307
		XciT[6]		2021	NeurIPS	48.4	200	32
		SegFormer-B5[712]		2021	NeurIPS	51.8	84.0	183
		SegViT-v2*[800]		2022	IJCV	55.7	308.8	234
		TransKD*[432]		2024	ITSS	50.7	12	80
	Hybrid	NL*[678]		2018	CVPR	44.7	300	54
		DANet[184]		2019	CVPR	45.3	250	69.0
		OCRNet[784]		2021	IJCV	48.0	260	71.0
		HMANet*[494]		2021	TGRS	44.2	180	55
		MaxViT-L*[632]		2022	ECCV	51.1	245.4	212
		Semask[294]		2023	ICCV	53.5	356	212
Instance Segmentation (mAP)	CNN	one-stage	MSCOCO	2019	ICCV	29.8	45.1	30.8
				2018	CVPR	37.3	71.6	44.7
				2019	ICCV	40.5	172.2	50.7
				2020	NeurIPS	41.7	190.3	63.7
				2020	CVPR	36.3	284.9	46.7
				2020	CVPR	30.3	24.8	21.5
	two-stage	Mask R-CNN[245]		2017	ICCV	38.5	266	44.4
		PANet[433]		2018	CVPR	42.0	286.8	49.1
		HTC[74]		2019	CVPR	41.2	324.6	58.3
		BlendMask[70]		2020	CVPR	41.3	256	42.4
		CondInst[622]		2020	ECCV	39.1	185	38.5
		Refinemask[804]		2021	CVPR	41.5	360	61.2
Hybrid	CenterMask-Lite[360] Conformer[516] DAT-B[707] MaxViT-B[632] Mask-Transfuser[319] Biformer-B[861]	2020	CVPR	38.3	24.6	96.9		
		2021	ICCV	40.74	457.7	56.9		
		2022	CVPR	45.8	1003	145		
		2022	ECCV	45.7	346.8	92.5		
		2022	CVPR	41.6	326.7	62.5		
		2023	CVPR	43.7	245.3	58.7		
CNN	Panoptic FPN[328] AdaptIS[580] Panoptic-DeepLab[104] LPSNet[262] Panoptic FCN[389] Mask-Pyramid[708]	Cityscapes	2019	CVPR	58.1	275	63.0	
			2019	ICCV	59.0	76.4	32.6	
			2020	CVPR	63.1	144.9	44.3	
			2021	CVPR	61.9	80	18.9	
			2022	PAMI	61.4	213.0	62.5	
			2024	Sensors	65.2	215	49.2	
			ViT	kMaX-DeepLab[774] Panoptic-Segformer[400] Center-Trans[22] Panoptic-partformer[382] Pix2Seq-D[88] PanSR[877]	2022	ECCV	68.4	285.0
2022	CVPR	65.1			205.0	62.6		
2023	Electronics	65.3			428	72.0		
2022	ECCV	64.7			249	56		
2023	CVPR	64.9			1120	78.0		
2024	arXiv	67.1			305	54.8		
Hybrid	Axial-DeepLab[655] EPSNet[60] Max-DeepLab[654] CMT-DeepLab[772] Mask DINO[369] BEE-Net[281]	2020	ECCV	62.1	55.6	178.5		
		2020	ACCV	60.8	98.7	24.3		
		2021	CVPR	63.5	214.0	360.0		
		2022	CVPR	67.8	238	48.1		
		2023	CVPR	64.6	987.3	72.5		
		2024	ICRA	65.0	305.7	45.3		

Table 2.2: A comparative analysis between different models on segmentation

2.4. Techniques

Model			Year	Publication	ADE20K mIoU(%)	COCO mAP(%)	COCO PQ(%)	#Params (M)
Universal Segmentation	ViT+Hybrid	K-Net[822]	2021	NeurIPS	50.6	40.1	48.3	56.3
		Mask2Former[105]	2022	CVPR	56.1	50.1	57.8	216
		SEEM[876]	2023	NeurIPS	-	46.3	56.1	40.0
		OneFormer[293]	2023	CVPR	58.1	49.2	58.0	223
		DaTaSeg[208]	2023	NeurIPS	54.0	-	53.5	-
		DFormer[653]	2023	arXiv	48.3	44.4	52.5	44
		ViT-P[566]	2025	arXiv	58.6	49.5	58.0	309

Table 2.3: A comparison of universal segmentation models.

(e.g., scene, object, part) has recently become an active research area [710]. Compared with the task-specific framework, the generic framework must define a common entity representation and a universal pixel-level classification process shared across semantic, instance, and panoptic segmentation tasks within a single model. As for CNNs with attention mechanisms, K-Net [822] adopted a group of learnable kernels as the common entity representation, where each kernel can dynamically update meaningful groups conditioned on the input image. Therefore, K-Net allows contextual information to flow among kernels via multi-head attention and establishes a one-to-one mapping between dynamic kernels and entities, generating one mask per potential "things" or "stuff" partition. As for ViT-based approaches, many researchers are developing unified frameworks that use learnable embeddings (also called queries) and cross-attention to integrate entity representation, long-range entity relation inference, and pixel-wise classification. MaskFormer [106] introduced a set of mask embeddings associated with a global binary prediction for each entity. Mask2Former [105] proposed masked cross-attention for a new architecture capable of addressing universal image segmentation. OneFormer [293] proposed a task-conditioned joint training strategy that enables training on ground truths of each domain (semantic, instance, and panoptic segmentation) within a single multitask training process. Integrated a task token into the task-conditioned prediction head, OneFormer can support multi-task training and inference based on the given task. DFormer [653] viewed the universal image segmentation task as a denoising process using a diffusion model. By adding Gaussian noise at various levels to ground-truth masks, DFormer used the noisy masks as inputs and gradually learned to predict denoising masks using a diffusion-based decoder. SEEM[876] proposed a promptable and interactive model for segmenting everything everywhere all at once in an image. It used a text encoder to encode text queries and mask labels into the same semantic space. Datasetseg[208] trained a universal multi-dataset multi-task segmentation model with a shared representation between mask proposals and class predictions for all tasks. ViT-P[566] introduced a novel two-stage

segmentation framework that decouples mask generation from classification for accurate mask classification, especially in the presence of ambiguous boundaries and imbalanced class distribution.

Evolution trends and practical deployment gaps

The evolution of image classification models follows a clear trajectory from pure CNNs to Vision Transformers and, finally, to hybrid architectures, driven by the need to balance local feature extraction with global context modeling. Early CNNs excelled in efficiency and translation equivariance but struggled with long-range dependencies. ViTs introduced global attention for superior context capture, but at a high computational cost and substantial data consumption. Hybrid models have primarily combined the convolutional inductive bias with the attention mechanism’s flexibility, achieving state-of-the-art accuracy with manageable FLOPs. However, key deployment gaps remain: (1) efficiency–accuracy trade-off for lightweight models (MobileNetV3, EfficientViT) that sacrifice accuracy for speed, limiting their use in high-stakes applications; (2) hardware optimization for many hybrid designs that are not yet optimized for mobile or edge deployment due to irregular memory access patterns; (3) generalization under domain shift of models trained on curated datasets (ImageNet) that often degrades in real-world noisy environments. Future work must focus on hardware-aware co-design and dynamic inference to adapt the computational cost to the input.

2.4.2 Object Detection

As a fundamental task in computer vision, object detection is the key block for more complex tasks such as autonomous driving [428, 872, 762, 588], robotics [740, 426, 231], surveillance [59], pose / behavior estimation [375, 269, 842], etc. The core difficulty in object detection is processing two distinct sources of information: spatial information (the location of each target object) and appearance information (the pattern of each object for category classification). Compared to image classification, object detection operates at a regional level, where each object is localized with a bounding box. This approach bridges the gap between semantic segmentation (pixel-level) and instance segmentation (instance-level), simplifying the task by avoiding pixel-wise annotations. Although regional representations require fewer parameters and simpler annotations for training, this comes at the cost of reduced spatial exclusivity across different instance regions. The compromise means that bounding boxes may overlap in crowded scenes, leading to ambiguities in instance separation. In this subsection, we introduce the

2.4. Techniques

development of generic object detection and then elaborate on one specialized sub-task, small object detection. We summarize the key models for generic and small-object detection in Table 2.4 for easier comparison.

Generic object detection

Generic object detection has been an active research area in computer vision for several decades, aiming to detect a broad range of general categories [429]. Although many traditional algorithms have been proposed to address the challenging problem, deep learning techniques have obtained significant attention for learning feature representations automatically from data, further advancing the development of object detection.

At an early stage, CNNs, as powerful tools for automating feature learning, have been a major focus of research aimed at addressing the core difficulty of object detection: how to represent spatial and appearance information. Then, set loss functions are used to solve the many-to-one mapping problem, assigning each ground-truth object to the most likely candidate among many others. Around the representation schemes, CNN-based detectors were organized into two main categories: two-stage frameworks extract different information sequentially, while one-stage frameworks represent features in parallel. The two-stage methods consist of two sub-networks. The former sub-network represents spatial information as regional proposals, e.g., using RPN (region proposal network), and the latter extracts appearance information within each region. The two-stage methods can generate sparse proposals flexibly because RPN over-samples only the potential regions and ignores the rest, thereby achieving high precision on average. Sorted by release time, two-stage CNNs mainly include R-CNN [196], SPP-Net [247], Fast R-CNN [197], Faster R-CNN [545], ION [31], HyperNet [334], R-FCN [126], FPN [414], DCN [127], Mask R-CNN [245], ZIP [371], Light R-FCN [398] and Cascade RCNN [503, 49]. Alternatively, the one-stage method employs a unified regression framework. Each position in the generated feature maps encodes spatial coordinates (e.g., bounding boxes) and appearance information (e.g., class probabilities). Therefore, it utilized a grid structure where each grid point serves as an anchor corresponding to a potential object, resulting in dense predictions. Due to the shared feature maps, the one-stage methods have lower computational complexity in resource-limited scenarios. The one-stage CNNs mainly include MultiBox [164], DetectorNet [609], OverFeat [565], AttentionNet [767], YOLO [540], G-CNN [488], SSD [436], DSSD [181], RON [333], YOLOv2 [541], STDN [853], FSSD [402], RefineDet [820], RFBNet [434], YOLOv3 [542], CornerNet [346], CenterNet [160], FCOS [623] and other improved YOLOs

[446, 409, 224, 279, 38, 295, 649, 647]. However, CNN-based detectors cannot solve the ambiguities in bounding box assignment. Some post-processing steps, such as non-maximum suppression (NMS), must be applied during inference to eliminate false-negative candidates near each true positive.

Later, the attention mechanism was explored in object detection, as query embeddings can be object proposals that efficiently combine spatial and appearance information. Therefore, attention mechanisms can capture long-range interactions among potential object proposals and infer complex contextual patterns across different images. However, the pure ViT-based detectors have obtained limited research interest because their characteristics are less effective for object detection. Specifically, the patch tokenization process cannot provide sufficient and precise object proposals. At the same time, the plain and non-hierarchical architecture cannot generate multi-scale features to handle diverse shapes and sizes [659]. YOLOS [176] adapted its patch tokenization process to provide sufficient *Det-Tok* that are learnable query tokens for object binding. Moreover, CSWin Transformer [154] and ViTDet [387] explored hierarchical structures to generate multi-scale feature maps.

Hybrid architectures that combine CNNs and ViTs have recently emerged as the primary research direction for improving object detection performance [825, 633]. The hybrid frameworks can be grouped in two ways: (a) ViT backbones are incorporated with various CNN detection heads (e.g., Mask R-CNN [245], Cascade R-CNN [49], FCOS [623]) to perform object detection. ViT-FRCNN [30] added a ViT-based backbone to a Faster R-CNN object detector for better generalization ability. Similar to FPN [414], FPT [803] combined a non-local network with multi-scale features, while PVT [673, 674] introduced a high-to-low resolution process into RetinaNet [554] to improve detection performance. Moreover, Swin Transformer [441], SwinV2 [440], ViL [816], and Focal Transformer [747] constructed a local-to-global combination to efficiently extract both local and global dependencies with less computational overhead. (b) CNN backbone with ViT-based decoder [844, 462]. DETR [53], as a pioneer for transformer-based detectors with CNN backbones, defined a small fixed number of static learnable embeddings (called object queries) and adopted cross-attention to learn long-range dependencies between queries and feature maps. By using set loss functions for object prediction, DETR eliminated traditional handcrafted components, such as RPN and NMS post-processing. The following research aims to improve DETR’s convergence speed, computational complexity, and other aspects. ACT [844] adopted a locality sensitivity hashing (LSH) method to cluster the query features for a lower computational burden adaptively. Inspired by deformable convolutions [127],

2.4. Techniques

Model				Dataset	Year	Publication	mAP (%)	GFLOPs	#Params (M)
Generic Object Detection	CNN	one-stage	CenterNet[160]	MSCOCO	2019	ICCV	47.0	-	-
			FCOS[623]		2020	PAMI	44.7	-	-
			RepPoints-v2[96]		2020	NeurIPS	49.4	-	-
			EfficientDet[613]		2020	CVPR	52.6	-	-
			PP-YOLOE[735]		2022	arXiv	50.9	110.1	52.2
			YOLOv10[647]		2024	NeurIPS	29.5	160.4	29.5
	CNN	two-stage	Faster RCNN[545]		2015	NeurIPS	42.7	-	-
			Mask RCNN[245]		2017	ICCV	37.1	63.8	248.0
			DCN[127]		2017	ICCV	37.5	-	49.5
			Cascade RCNN[49]		2018	CVPR	42.8	257.0	87.1
			Dynamic RCNN[809]		2020	ECCV	49.2	-	-
			Sparse RCNN[603]		2021	CVPR	51.5	-	-
	ViT		YOLOS[176]		2021	NeurIPS	42.0	538	127
			CSWin[154]		2022	CVPR	48.7	526	97
			ViTDet[387]		2022	ECCV	58.7	3600	692
			Swin-v2[440]		2022	CVPR	56.4	1043	160
			DINO[808]		2023	ICLR	63.2	860	218
			ViT-Adapter[101]		2023	ICLR	52.1	-	158
	Hybrid		DETR[53]		2020	ECCV	44.9	253	60
			FPT[803]		2020	ECCV	42.6	88.2	346.2
			Focal Transformer[747]		2021	NeurIPS	49.0	533	110
			Deformable DETR[870]		2021	ICLR	46.2	173	40
			Swin-ACMix[506]		2022	CVPR	51.1	754	58
			Next-ViT-L[373]		2022	arXiv	48.0	391	77.9
Small Object Detection	CNN	one-stage	Polardet[839]	DOTA	2021	IJRS	76.6	-	-
			R ³ Det[753]		2021	AAAI	71.9	264.2	37.8
			S ² ANet[222]		2021	TGRS	74.2	287.0	39.7
			GWD[754]		2021	ICML	76.3	336	41.9
			KLD[755]		2022	PAMI	77.3	336	41.9
			KFIoU[756]		2023	ICLR	74.2	287.0	39.7
	CNN	two-stage	O-RCNN[718]		2021	ICCV	76.3	199	41.1
			ReDet[223]		2021	CVPR	76.3	-	31.6
			CFA[218]		2021	CVPR	76.7	-	-
			AOPG[107]		2022	TGRS	75.4	-	-
			ARC[524]		2023	ICCV	77.4	-	74.4
			PKINet-S[48]		2024	CVPR	78.39	70.2	30.8
	ViT+Hybrid		ViTDet[387]		2022	ECCV	74.4	502	103.2
			RVSA[650]		2023	TGRS	81.2	414	114.4
			ViTDet-OB[659]		2023	IGARSS	75.7	-	-
			AO2-DETR[128]		2023	CSVT	77.7	-	-
			Strip R-CNN[783]		2025	arXiv	80.6	52.3	13.3
			LSKNet[397]		2025	IJCV	81.6	161	31.0

Table 2.4: A comparative analysis between different models on detection.

Deformable DETR [870], and DINO [808] introduced deformable self-attention that samples on a sparse set of significant positions across all scales, reducing computational complexity. TSP [606] and SMCA [189] explored accelerating the convergence by improving the cross-attention mechanism. Conditional DETR [472] proposed a spatial

prior, called conditional spatial embedding, to narrow down the valid search range for localizing distinct regions. Efficient-DETR [764] introduced an additional region proposal network into the detection pipeline as a dense prior to improve initialization. Anchor DETR [682] replaced object queries with anchor points to predict multiple objects at one position and proposed a variant of attention for lower computational complexity. Different from the DETR series, Pix2Seq [89] is a generalized Transformer-based framework that uses a quantization and serialization scheme to convert bounding boxes and class labels into a sequence of discrete tokens. Diffusion efficiently represents potential candidates of small objects without handcrafted anchors in a one-stage framework equipped with dynamic anchors [814, 173, 438]. DiffusionDet was the first work to introduce diffusion models into object detection. It used a forward diffusion process to generate normalized box coordinates from Gaussian noise and adopted cross-attention blocks to learn how to predict high-quality boxes by denoising iteratively.

Small Object Detection

Small object detection (SOD)³ [108] is a specialized object detection task that gains increasing popularity due to fast-growing data acquired from drone-view images, remote sensing satellite images, High-altitude surveillance images, and high-throughput microscopy images [458]. In addition to common difficulties in generic object detection, SOD has demonstrated unique challenges: (1) Image acquisition from high altitudes generates high-resolution images due to the wide-angle lens, causing computational difficulty in a detector [49], especially on mobile devices; (2) Fitting large data in detectors with down-sampling operations inevitably deteriorates representation details of small-sized objects; (3) The aerial viewed images are characterized severely uneven information densities because overcrowded objects normally occupy a minor proportion, imposing another request of efficient detection on key regions [421]; (4) Complicated situational factors, such as occlusion, background clutter, and various lighting conditions, further aggravate detection precision. Therefore, generic detectors cannot locate and classify small objects, which are typically densely packed in images, without dedicated adaptation. To tackle the dilemma between information density and efficiency, many works aim to restore missing information about small objects, focus on crowded regions, and extract long-range contextual information among objects. The Focus-and-Detect approach [745, 159, 336] was a two-stage framework that found the crowded regions and restored the missing information with super-resolution techniques

³In this task, the term *small* can be defined by thresholds of area [416] or length [751, 776].

2.4. Techniques

[505, 137, 29, 192]. Another line of one-stage detectors fused feature maps at different scales to enhance the representation of small objects [97, 527, 199, 261, 744]. In addition, attention mechanisms have been introduced into SOD detectors for complicated context modeling [388, 182, 449, 283, 535, 447, 427]. Most recently, diffusion models have been introduced to efficiently represent potential candidates for small objects without handcrafted anchors [677, 658].

Evolution trends and practical deployment gaps

Object detection has evolved from two-stage CNN frameworks (e.g., Faster R-CNN) to one-stage anchor-free models (e.g., YOLO, FCOS) and recently to Transformer-based detectors (e.g., DETR). The shift reflects a trade-off between precision and inference speed, with one-stage models favoring real-time performance and transformers offering end-to-end training without handcrafted components. For small object detection in aerial/remote-sensing imagery, multi-scale feature fusion and attention mechanisms are critical, yet they increase computational cost. Deployment gaps include: (1) scalability on high-resolution images for costly dense prediction; (2) robustness to occlusion and scale variation, especially in crowded scenes; (3) integration with downstream systems, e.g., autonomous driving, requires low latency and high reliability, which current detectors struggle to guarantee simultaneously. Future directions could emphasize efficient multi-scale attention, diffusion-based refinement, and real-world robustness testing beyond standard benchmarks.

2.4.3 Image Generation

Image generation [162, 798, 307] refers to using algorithms to create or synthesize new images from scratch or by modifying existing ones. These algorithms learn from a large dataset in a generative manner, since they have to approximate the distribution of the dataset and output new, realistic, and detailed images from generated samples. As a result, the learning algorithms can capture patterns within the dataset and combine them to create new feature representations for data generation. It is a prominent field in computer vision and artificial intelligence, in which models generate realistic or stylized images that may resemble photographs, art, or imaginative scenes given specific input conditions. The primary goal is to generate images that appear authentic or fulfill a specified criterion, often mimicking real-world data distributions. This section primarily focuses on technique development for generating super-resolution images with enhanced quality and for translating images across different styles.

Image Super-Resolution

Image super-resolution (SR) [71, 419, 361] is an upsampling technique in computer vision that increases the number of pixels for a high-resolution (HR) image based on a given low-resolution (LR) input, as depicted in Figure 2.8. The goal is to analyze the contextual information surrounding observed pixels in an LR image and identify interpolation patterns to fill the unknown details between upsampled pixels in an HR image. SR is valuable in medical imaging, satellite imaging, and surveillance, where high-quality images are crucial for analysis and interpretation. In recent years, we have witnessed remarkable progress in image super-resolution using deep learning techniques. Initially, CNNs brought a breakthrough in SR by automating feature extraction from data and serving as discriminative models that directly learn LR-to-HR mapping functions, using two different architectures: interpolation-based and sub-pixel-based methods [686]. Interpolation-based approaches seek to identify patterns within a local region and interpolate a new pixel as a weighted sum of the surrounding feature vectors. Some SR models [152, 321, 322, 405, 273] combined static upsample interpolation with a stack of convolutions to decompose a learnable interpolation into static upsampling and learnable adjustment. In addition, transposed convolution (a.k.a deconvolution [794, 793]) is another learnable upsample method for SR models [466, 153, 624, 172, 227, 230, 188]. On the contrary, the sub-pixel-based methods [572] learn to encode the local patterns into feature channels and decode the local information from feature channels for new pixel reconstruction. Due to its end-to-end upsampling nature, the sub-pixel method is widely used in SR models [4, 812, 833, 344, 648]. Later, generative CNNs were developed to solve the SR problem. Starting with Generative Adversarial Networks (GANs), CNNs integrated adversarial training to directly learn to generate realistic textures and finer details for HR images. SRGAN [352] integrated sub-pixel modules into GAN-based architectures for better visual quality. Other variants [681, 8, 57, 378] combined different learnable upsampling methods with GANs and designed new architectures to improve SRGAN. Most recently, Diffusion models have been used as alternatives in generative CNNs for SR image generation [561, 548, 370, 787, 684].

Instead of using each pixel as input, ViTs take patches (a set of pixels) as input and introduce an attention mechanism to capture long-range dependencies in images, allowing models to focus on both global and local features for super-resolution. As a discriminative model, TTSR [746] treated LR images and reference images as the query and key, respectively, and adopted a hard-attention module to match relevant

2.4. Techniques



Figure 2.8: Image Super-Resolution learns to compensate for the missing details for better visualization. The image example is cited from SR3 [561]

textures between the reference and low-resolution images using a transformer. IPT [67] proposed a multi-task architecture based on ViTs pre-trained on large datasets, including super-resolution. DPT [843] proposed a dual-path transformer to capture spatial and frequency domain information for improved SR. ESRT [450] introduced a light ViT for SR with less computational complexity. Some research focuses on improving the computational efficiency of local window-based self-attention. SwinIR [403] adapted the Swin Transformer for image SR by focusing on local and global feature extraction. Uformer [685] proposed a LeWin Transformer block that performed

non-overlapping window-based local self-attention to reduce computational cost on high-resolution feature maps. HAT [92] integrated channel attention into window-based self-attention to activate more input pixels for better reconstruction. Additionally, generative ViTs adopted GANs [358, 490, 536, 190] to connect a ViT generator and a ViT discriminator for adversary learning to generate high-quality SR images.

Recently, hybrid methods that combine the strengths of CNNs and transformers have been applied to image SR. The CNNs with attention mechanisms learn to select the most relevant pixels in a non-local scope to reconstruct more coherent interpolations. SelNet [111] proposed a Selection-Unit to attend to relevant information and filter out the irrelevant for the latter SR reconstruction. RCAN [832] integrated channel attention to highlight significant feature channels for HR image generation. DNLN [72] integrated a deformable non-local attention with a multi-branch weighted feature fusion for better capture pixel-wise correlations, long-range self-similarities, and multi-scale semantic combination. In addition, the hybrid GANs that integrate convolution with ViTs [456, 625, 841, 27] have demonstrated outstanding performance on SR image generation. SWCGAN [631] adopted GAN architectures to connect a local attention-enhanced CNN generator and a pure Swin Transformer [441] based discriminator for adversarial training. Similarly, SRTransGAN [24] designed a generator that combined PVTv2 [674] with sub-pixel methods and adopted a pure ViT as the discriminator.

Evolution trends and practical deployment gaps

Image SR has progressed from CNN-based interpolation to GAN-driven perceptual enhancement (SRGAN, ESRGAN) and Transformer-based methods (SwinIR, HAT) that model long-range texture dependencies. CNNs are efficient and easy to deploy, but often produce overly smooth results; GANs improve visual realism but introduce training instability and artifacts. Transformers capture global patterns better but are computationally intensive, especially for high-resolution outputs. Hybrid designs (RCAN, DNLN) combine channel attention with non-local mechanisms to balance quality and speed. Deployment gaps are pronounced in real-time video SR and mobile applications, where memory and latency constraints are severe. Lightweight SR networks sacrifice some reconstruction fidelity for speed, highlighting the trade-off between perceptual quality, inference time, and hardware compatibility.

Image-to-Image Translation

Image-to-image translation (I2I) is a generative task in computer vision that involves learning a mapping between a source distribution and a target distribution for domain

2.4. Techniques

transfer, as depicted in Figure 2.9. It can be applied to various applications, such as style transfer. To train generative models, deep models, i.e., CNNs, ViTs, and hybrid architectures, have evolved from GANs to Diffusion models. CNN-based GANs were initially the primary approaches for I2I tasks due to their ability to automatically capture local patterns from images with convolutional operations and build a mapping between different domains with adversary training: Pix2Pix [290], CycleGAN [860], UNIT [431], DiscoGAN [323] and MUNIT [282]. Alternatively, CNN-based Diffusion models [492, 165] adopted diverse samples through iterative refinement, simplifying the training process of I2I tasks other than the GANs: SDEdit [471], Palette [560], BBDM [362], and DiffI2I [704]. ViTs have enabled global dependency modeling for I2I tasks that involve complex, non-local transformations, such as large structural changes or style modifications. GAN-based ViTs designed both generators and discriminators with pure visual transformers and integrated attention mechanisms into the adversary training: TransGAN [310], Styleformer [699], StyTr2 [141], StyleSwin [799], ViTGAN [210]. Additionally, there are a few Diffusion-based ViTs that find a mapping between two different Markov chains [705]: IPT [67], StyleDiffusion [687], DiT [512, 840], MDT [191], unidiffuser [25], DVT [221], DIFFIT [233]. Moreover, some researchers try to integrate the strengths of CNNs and ViTs into I2I tasks. The hybrid generative models equipped with GANs and Diffusion models learn to map the source and target domains: UVCGAN [625], TransWeather [636], and ITTR [848].

Evolution trends and practical deployment gaps

Image generation has progressed from CNN-based GANs to Transformer-driven generative models and diffusion paradigms. GANs offer high-quality synthesis but still suffer from training instability; diffusion models provide stable training and fine-grained control but are computationally intensive. ViT-based generators (StyleSwin, TransGAN) enable long-range coherence but require substantial memory. Hybrid approaches balance local texture details with global structure. Key deployment challenges are: (1) inference latency for diffusion models that require iterative sampling, limiting real-time use; (2) resource demands on high-resolution generation that are prohibitive on edge devices; (3) generalization to unseen styles/domains that models often overfit to training data distribution. Practical deployment in fields like entertainment necessitates model compression, distillation to lightweight generators, and efficient sampling algorithms to bridge the gap between quality and speed.

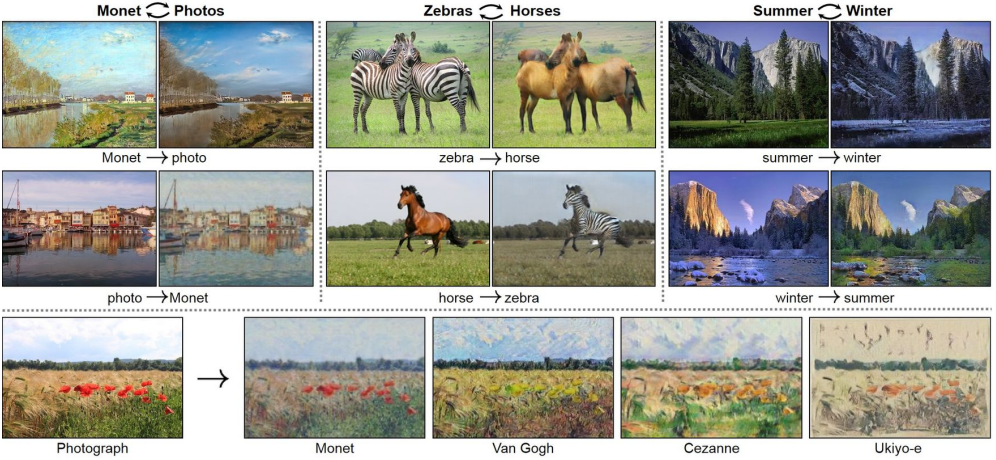


Figure 2.9: Image-to-Image translation learns a mapping between a source distribution and a target distribution for domain transfer. The image example is cited from [860].

2.4.4 Universal Models

The development of universal models paves the way for Artificial General Intelligence (AGI). It is no longer merely a tool for achieving higher accuracy on benchmarks for specific tasks but has become the foundational engine for constructing generalist visual AI. This pursuit of universality, while computationally intensive, is unlocking new frontiers in model capability for important but challenging domains.

Large-Scale Visual Foundation Models

These large-scale visual foundation models have proven effective in specific challenging domains. A compelling research thread involves adapting SAM variants to maneuver the challenges of scarce annotation in weakly-supervised settings. WS-SAM [235] leveraged SAM by using sparse annotations as prompts to generate pseudo segmentation masks of concealed objects. In order to improve mask quality, WS-SAM integrated a multi-scale feature grouping technique with other strategies to tackle the intrinsic similarity challenge. Furthermore, SEE [236] extended the WS-SAM to exploit the SAM for camouflaged object segmentation with incompletely annotated data. Most recently, SCALER [239] presented a unified collaborative framework that jointly optimizes a mean-teacher segmenter and a learnable SAM2 for label-deficient concealed object segmentation. This evolution, from WS-SAM (utilization) through SEE (distillation) to SCALER (collaboration), has demonstrated a clear development trajectory in which a general-purpose foundation model (SAM/SAM2) can be progressively and

2.4. Techniques

deeply integrated to solve domain-specific problems with increasing sophistication and performance.

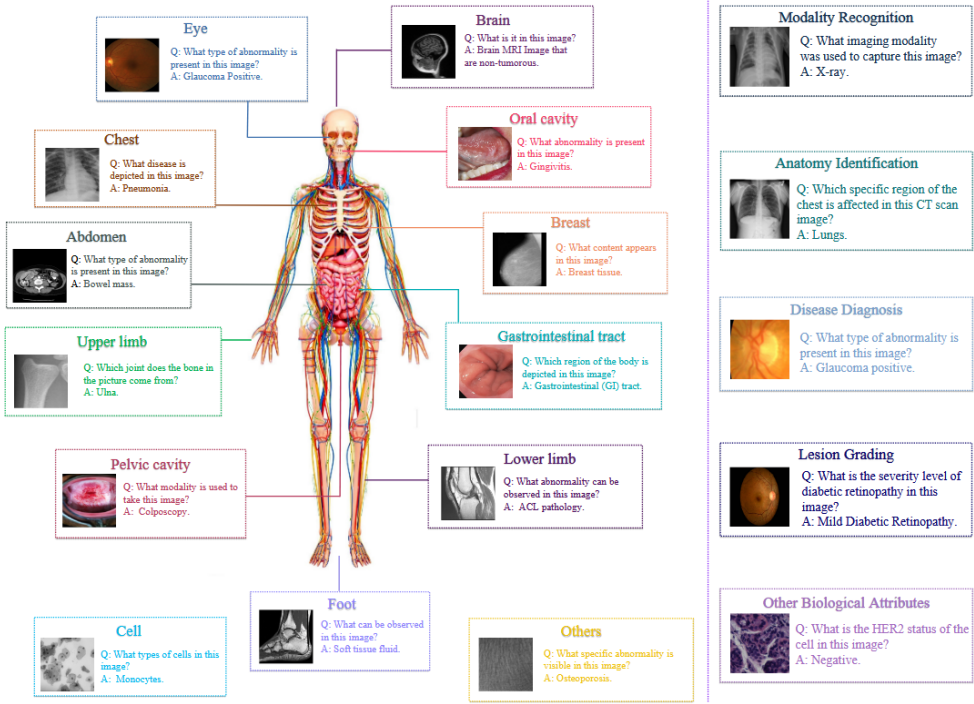


Figure 2.10: Left: Overview of OmniMedVQA dataset[274], which covers the majority of radiology modalities and anatomical regions of the human body. Right: Illustrations of samples from five different medical question types.

Multiple Modal Models

The transition from single-modal to multi-modal foundation models marks a critical phase in the evolution of General Artificial Intelligence. By transcending the confines of research benchmarks, these models demonstrate their paramount importance as versatile core technologies across diverse domains. Their ability to process and synthesize heterogeneous data streams is fundamentally reshaping methodologies and enabling novel solutions to complex real-world problems.

Vision-Language Models

The practical applications of Visual-Language Models (VLMs) are extensively demon-

strated by their zero-shot prediction capabilities across a wide spectrum of visual recognition tasks. The foundational model CLIP established the paradigm, achieving remarkable zero-shot accuracy on diverse image classification datasets, ranging from general benchmarks like ImageNet [138] to fine-grained tasks such as Stanford Cars [339] and Oxford-IIIT Pets [509]. Beyond image-level recognition, subsequent VLMs have been successfully adapted for more complex dense prediction tasks. For instance, models like GLIP [376] and DetCLIP [760] enabled zero-shot open-vocabulary object detection by aligning region proposals with text descriptions. In contrast, methods such as GroupViT [732] and SegCLIP [451] extend this capability to semantic segmentation by matching text embeddings with pixel or segment features. Furthermore, VLMs facilitate zero-shot image-text retrieval and have been evaluated on action recognition datasets such as UCF101 [589], thereby demonstrating their broad utility as general-purpose visual recognition engines.

Emotion Recognition Models

Building on the foundational cross-modal alignment principles of visual-language models, the multi-modal paradigm has been profoundly applied in the field of human-machine interaction. Significant advancements have been made in one important cross-modal application, i.e., emotional analysis, that integrates diverse data sources, such as speech, text, and images. Early explorations focused on novel fusion mechanisms, such as tensor decomposition [669], and brain-like hierarchical perception networks [866] to integrate visual, acoustic, and textual cues. The pursuit of robustness has become a central theme, leading to architectures such as the Robust Adversarial Fusion Transformer (RAFT) [668] for sentiment analysis and to frameworks employing random modality dropout [828] or contrastive learning [709] to mitigate negative information and handle incomplete data. Recent trends emphasize deeper contextual and generative modeling, as evidenced by works on contextual interaction based analysis [666] and methods that integrate audio-visual text generation with contrastive learning [667]. The state-of-the-art further incorporates diffusion models [868] and end-to-end systems [864], while practical deployment is explored through client-server recognition systems [867]. This concentrated research effort, as summarized in key reviews [865], demonstrates how multi-modal foundation models are specifically engineered to tackle the complexities of human emotion analysis across diverse and challenging real-world contexts.

2.4. Techniques

Evolution trends and practical deployment gaps

The rise of large-scale foundation models (DINOv2, SAM) and multi-modal systems (CLIP, VLMs) marks a shift from task-specific models to general-purpose visual intelligence. These models excel in zero-shot transfer and open-vocabulary understanding but require substantial computational and data resources. While VLMs enable semantic grounding across vision and language, their size limits on-device deployment. Trade-offs involve scale vs. accessibility (since foundation models are powerful but require cloud-level resources), and generalization vs. specialization (since they may underperform in niche fields without fine-tuning). Deployment gaps include: (1) memory and latency, which make running billion-parameter models on edge devices infeasible; (2) privacy and data governance, which raise concerns in sensitive domains when centralized training is used; (3) energy consumption, which conflicts with sustainable AI goals. Future efforts could prioritize modular, scalable architectures, federated learning frameworks, and efficient cross-modal distillation to make universal models practical for widespread use.

2.4.5 Medical Data Analysis

Beyond generic vision tasks, deep learning has also revolutionized domain-specific applications, such as the medical field. Medical data analysis is one of the most important Interdisciplinary tasks, since it enables automated decision-making processes, such as disease diagnostics, treatment planning, and prognosis. Formally speaking, each decision-making process is a multi-modal task that fuses information from visual data (e.g., pathology staining and radiological images) and sequential data (e.g., clinical details, biomarkers, protein sequences, and medical history) to support quantitative analysis. Precise and thorough panoptic segmentation of a medical image is key to scene understanding, as it facilitates the representation of each entity separately. It is the first step for multi-modal tasks, but has attracted less attention than instance segmentation, and has recently gained increasing attention in medical image analysis. Moreover, a central challenge in designing medical AI systems lies in the architecture of multi-modal models, which must integrate diverse data sources to extract richer and more robust feature representations.

Image Segmentation

CNN-based segmentation for medical image analysis mainly focused on feature fusion from different branches due to the failure to represent concepts efficiently [140, 725]. Liu et al. [424] combined an auxiliary semantic segmentation branch with the instance

branch to integrate global and local features to enhance nuclei segmentation. PFFNet [423] incorporated the global-level semantic with local-level instance features and proposed a mask quality sub-branch to align the semantic and instance predictions. Panoptic Lintention Network [465] provided a comprehensive understanding of the environment of individuals with visual impairments in healthcare. Additionally, ViT-based instance segmentation models have been the primary options due to their attention mechanisms. Cell-DETR [521] adapted from the DETR for cell instance segmentation and achieved state-of-the-art performance from microscopy imagery. Medical Transformer [635] explored transformer-based solutions and proposed a Gated Axial-Attention model by introducing an additional control mechanism in the self-attention module. Verma et al. [644] explored self-training strategies for panoptic segmentation and implied their potential enhancement for medical image analysis.

Medical Vision-Language Models

Medical Vision-Language Models (MVLMS) [797, 275, 274] are advanced AI systems that combine visual and textual medical data to perform complex tasks such as medical report generation (MRG) and Medical Visual Question Answering (MVQA), as shown in Figure 2.10. These models have to understand medical data and generate corresponding textual descriptions. Beyond merely image captioning, MRG has evolved to generate descriptive reports from input visual and textual data automatically. Initially, MRG models combined different branches to process visual and textual data separately: CNN-RNN architectures [313, 395, 778, 766] and CNN-Transformer architectures [593, 491]. Later, knowledge graphs [425, 831] were adopted to represent heterogeneous data across modalities. Most recently, Transformer-based models [482, 617, 619, 819, 171] have been the most popular solutions. MVQA [272] can answer questions about medical images based on understanding both the visual content of the image and the natural language question. Likewise, MVQA models initially combined two branches: CNN-RNN [504] and CNN-Transformer [646, 280, 401], and later adopted transformer-based architectures [489, 786].

Evolution trends and practical deployment gaps

Medical vision has evolved from CNN-based segmentation to Transformer-enhanced models and multi-modal systems. These advances improve accuracy and enable report generation, but they face unique domain constraints: limited annotated data, high reliability requirements, and multi-modal integration. Hybrid CNN-ViT models balance local anatomical detail and global context, yet they struggle with interpretability and

2.5. Conclusions

robustness to rare pathologies. Deployment gaps are pronounced: (1) clinical validation that models rarely undergo rigorous real-world trials; (2) integration with hospital IT systems, where data privacy and interoperability are major hurdles; (3) computational resources that high-resolution 3D medical images demand GPU clusters, limiting point-of-care use. Moving forward, the field ought to invest in explainable hybrid architectures, federated learning for data privacy, and lightweight, domain-specific compression to transition from research prototypes to trusted clinical tools.

2.5 Discussions and Conclusion

Given the rapid evolution of hardware performance, deep learning models are and will be a hot topic in computer vision due to their competitive performance and tremendous potential compared with other machine learning approaches. To harness the power of deep learning models, as summarized in this survey, many methods have been proposed in recent years. These methods perform well across a range of visual tasks, including image classification, object detection, and image generation. Nevertheless, the potential of deep learning models for computer vision has not yet been fully explored, and both challenges and prospects remain.

2.5.1 Discussions

Thanks to automatic data fitting and generalized inference, deep learning models have transformed many computer vision tasks. Despite their successes, they still have considerable room for improvement, especially in more complex real-world scenarios. As deep learning evolves, it is crucial to incorporate novel ideas into deep learning models, especially in interpretability, data efficiency, multi-modal understanding, and scalability, making it a powerful tool for future AI applications. To address the aforementioned challenges, future research should focus on the following directions.

Improving Data Efficiency and Annotation Scarcity

The performance of all deep vision models, from CNNs to large-scale foundation models, remains intrinsically linked to the volume and quality of annotated data—a major bottleneck in data-scarce domains. While self-supervised learning (SSL) paradigms, exemplified by DINO, have made significant strides in learning transferable representations from uncurated data, they often demand vast computational resources. The integration of generative models, particularly Diffusion Models, offers a transforma-

tive complementary path. Future research could move beyond using these models in isolation and develop synergistic frameworks in which diffusion models generate high-fidelity, task-aware synthetic data or feature augmentations to robustify the training of discriminative CNNs, ViTs, and hybrids. Furthermore, unifying the principles of SSL with the generative capabilities of diffusion processes could lead to next-generation pre-training paradigms that learn compressed, semantically rich world models from minimal supervision, drastically reducing annotation dependency for downstream adaptation.

Enhancing Model Interpretability and Trustworthiness

As architectures grow in complexity—from attention maps in ViTs to dynamic pathways in hybrid or Mamba-based models—their opacity deepens, eroding trust in safety-critical applications like medical diagnostics. Merely visualizing attention weights or activation maps provides post-hoc insight but seldom explains causal reasoning. A promising frontier lies in embedding explainability directly into the model’s architecture and training objective. Deep Unfolding Networks (DUNs) provide a foundational blueprint in which each network module corresponds to a step in an interpretable optimization algorithm. Extending this principle, future architectures could be designed to have inherently interpretable latent spaces, or by integrating differentiable probabilistic graphical models that enforce causal structure. For large foundation models, developing rigorous concept-based explanation methods that align internal representations with human-understandable concepts is crucial. Ultimately, trustworthiness will be achieved not by adding external explanation tools, but by co-designing models whose decision-making processes are transparent and aligned with domain-knowledge constraints by construction.

Achieving Real-World Robustness and Generalization

A significant gap persists between benchmark performance and reliability in dynamic, open-world environments, where models have to face unforeseen distribution shifts, adversarial perturbations, and novel object configurations. Architectural hybrids of CNNs and ViTs aim to balance local invariance and global context for better generalization, but they lack formal robustness guarantees. The strategic incorporation of multi-modal knowledge, particularly from Vision-Language Models (VLMs), offers a powerful regularizing force. Semantic constraints from language can ground visual predictions and help reject nonsensical adversarial outputs. Future research could focus on developing tightly coupled, efficient vision-language systems where a compact, task-specific vision model (CNN/ViT/hybrid) is continuously guided by the semantic and

2.5. Conclusions

commonsense knowledge of a (potentially frozen) large VLM. Furthermore, simulation-to-real and test-time adaptation techniques must be advanced to enable models to continuously self-correct and adapt to novel environments, leveraging principles derived from foundational pre-training, moving from static models to adaptive visual agents.

Exploring Design Paradigm for Efficiency

The pursuit of efficiency must evolve from applying isolated compression techniques post-hoc to embracing a holistic efficiency-by-design paradigm across the entire stack, encompassing everything from training dynamics to inference latency. While Neural Architecture Search (NAS) and compression techniques such as pruning and quantization are essential, they must be deeply integrated and hardware-aware. For hybrid CNN-Transformer or State Space Model (SSM) based models, this means developing NAS frameworks that can co-optimize architectural partitioning (e.g., deciding where to use local convolution or global attention) alongside compression strategies and deployment target constraints (e.g., memory bandwidth, parallel computing). The next generation of efficient models will likely be dynamic and conditional, activating different computational subgraphs based on input complexity. Research could therefore develop unified frameworks that jointly search for optimal architectures, their optimal compression profiles, and their dynamic inference policies, yielding a portfolio-optimal spectrum of models for diverse real-world constraints rather than a single point solution.

From Static Deployment to Adaptive Edge Ecosystems

The final translational challenge is deploying capable models in heterogeneous, resource-constrained, and dynamic edge environments. Current deployment typically involves distilling a large model into a smaller, static network, which becomes brittle under changing conditions. The future lies in adaptive edge ecosystems. The adaptation requires models, particularly efficient hybrids, to be endowed with context-aware adaptive inference capabilities, i.e., dynamically adjusting their depth, width, or even architectural pathways in response to real-time metrics such as battery level, computational load, or perceived task difficulty. Furthermore, the rise of foundation models presents a new opportunity: edge devices could run a highly compressed, task-specific model while occasionally querying a cloud-based foundational model for complex, out-of-distribution inferences, creating a collaborative hierarchy. Advancing federated learning frameworks to support the continuous, privacy-preserving fine-tuning of these adaptive models across devices will be critical. The ultimate goal is to transition from deploying static models to cultivating intelligent, collaborative, and self-improving

visual perception networks at the edge.

Ethical Considerations for Deep Learning Models

Beyond technological advances, the widespread deployment of deep learning systems in real-world applications necessitates consideration of ethical issues, since decisions influenced by AI can have profound consequences for individuals and communities in high-stakes environments. Key concerns include algorithmic bias, which can perpetuate or even amplify societal inequalities; privacy risks arising from data collection and model inference; and a lack of transparency that undermines trust and accountability. As large-scale foundation models become increasingly integrated into daily workflows, ethical considerations collectively implicate their deployment in sensitive domains such as medical diagnostics and surveillance, thereby making addressing ethical dimensions a fundamental prerequisite for responsible use. International organizations ⁴ assemble guidelines for responsible use that developers and users of deep-learning frameworks should take notice of and include in their line of work. Promising research directions for mitigating ethical risks include developing fairness-aware algorithms, adopting privacy-preserving learning paradigms, e.g., federated learning and differential privacy. In addition, and very importantly, is the endeavor to enhance model interpretability to ensure that decisions are both reliable and accountable. As researchers continue to push the boundaries of model performance, we suggest that parallel efforts would be directed toward understanding and addressing the ethical dimensions of deploying these systems in practice. We encourage the community to engage in interdisciplinary discussions that bridge technical robustness and ethical responsibility, ensuring that deep learning technologies evolve into positive and productive forces that serve society equitably and responsibly.

2.5.2 Conclusion

In conclusion, this survey has articulated the trajectory of visual recognition from dedicated CNNs to general-purpose Transformer-based foundation models, highlighting the convergent trend toward hybrid architectures that leverage complementary inductive biases. We have contextualized these core advancements within a broader package of efficiency, distillation, and generative enhancement techniques. The path forward, however, is not merely one of incremental architectural refinement. The next leap in computer vision will be driven by integrative research that crosses the boundaries

⁴1. UNESCO. 2. AI4People.

2.5. Conclusions

among data efficiency and generation, performance and interpretability, and isolated model design and adaptive system deployment. By addressing the interconnected challenges of data scarcity, opacity, fragility, inefficiency, and rigidity through a holistic lens, the field can advance toward building robust, trustworthy, and universal visual intelligence that operates reliably in the complexity of the real world. The convergence of architectural innovation, learning paradigms, and systems-aware co-design will define the next era of discovery in computer vision.

Chapter 3

VGMEA: Interpretable Cross-Attention mechanisms in Semantic Segmentation via Variational Gaussian Model Mixture

Preparing for arXiv pre-print.

Chapter Summary

The rise of attention mechanisms in computer vision has improved state-of-the-art performance and advanced network design. Originating in Natural Language Processing (NLP), attention mechanisms, including self-attention and cross-attention, effectively capture long-range dependencies between pairs of elements without regard to distance. Generally speaking, most existing attention modules can be roughly reformulated as either K-Means or K-NN clustering. K-Means style attention, maintaining clustering centers as global queries, favors building explicit feature spaces for closed-form vision tasks (e.g., semantic segmentation). In contrast, K-NN style attention, by generating dispensable clustering centers as dynamic queries, facilitates implicit feature spaces for open-form vision tasks (e.g., object detection and instance segmentation). In this chapter, we propose a variational Gaussian model mixture embedded attention (VGMEA) module to interpret cross-attention as a generalized probabilistic K-Means clustering approach. The VGMEA module also integrates geometrical regularization into cross-attention to improve the cross-class decision boundary based on non-local spatial dependencies.

Keyword: Probabilistic Inference Variational Bayes Attention Mechanism K-means Clustering

3.1 Introduction

The ability to effectively capture *long-range* dependencies across multiple scales is considered a crucial factor in improving the performance of deep neural networks across various tasks, from academia to industry. Traditionally, recurrent operations [558, 205] are the most widely used approach for modeling long-range dependencies in sequential data (e.g., speech and language). In contrast, large effective receptive fields formed by deep stacks of convolutional operations [186, 558] are the dominant approach to modeling long-distance interactions for spatial and temporal-spatial data (e.g., images and videos). However, a single recurrent or convolutional operation can only capture short-term relationships within a local neighborhood, either spatially or temporally. Therefore, capturing long-range dependencies can be approximated only by a chain of repeated operations, in which short-range signals are progressively propagated through each node.

Apparently, a long chain of repeating local operations has several limitations. First, approximating one operation through a chain will cause optimization instability that needs to be carefully addressed by some tricks, including dropout [258], residual shortcut [248]. Second, propagating signals step-wise along the chain is computationally inefficient. Finally, their perplexing computational graphs are too heavily weighted to be used as modules in other networks to model multi-scale dependencies, e.g., when messages must be communicated back and forth between distant positions at different scales. Therefore, attention mechanisms are operations designed to alleviate the aforementioned limitations. Those operators are more efficient in capturing long-range dependencies between pairwise positions.

Attention mechanisms have been a key component widely applied to improve the performance of deep neural networks on challenging visual datasets [138, 166, 416, 849, 123]. According to different purposes [214], attention mechanisms can be classified as channel attention [268, 665, 528], spatial attention [806, 53, 156], temporal attention [749, 5] and branch attention [97, 807]. Among these variants, some researchers [496, 44] explored the unified patterns to guide new attention mechanisms. All this work supports the idea that attention can enhance informative features, suppress irrelevant ones, and improve the learned feature representation by precisely modeling correlations between pairwise positions.

Compared to the significant enthusiasm for designing attention mechanisms, only a few works focus on their interpretability. Some relevant work [678, 379, 715, 773, 663] has partially explained why attention modules can highlight informative channels

3.1. Introduction

and/or regions. However, there is still no systematic framework for categorizing these variants. The attention mechanism can be revisited as mapping a *query* and a set of *key-value* pairs to an output. The output is computed as a weighted sum of the *values*, where the weight of each *value* is computed by a metric function between the *query* and the corresponding *key*. If regarding the *query* as a cluster center, then the attention mechanism can be understood as computing distances from the center to corresponding *keys* in one vector space while outputting a center of *values* in another vector space. However, allowing *keys* and *values* in different feature spaces actually hinders a systematic top-down viewpoint from interpreting all attention variants. Considering *values* are linear transformations from *keys* (as most variants adopted), we make a rational constraint that *values* would be mapped onto the same vector space as *keys*. Therefore, an attention operation can be analogized as one step of updating/shifting the **old** center (i.e., *query*) to a **new** center, depicted in Figure 3.1. Hereafter, one attention module, i.e., a stack of attention operations, can be interpreted as a clustering process with corresponding steps.

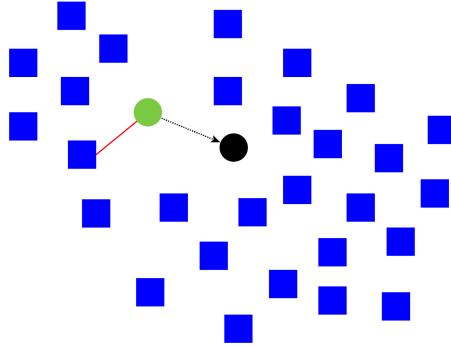


Figure 3.1: It is a straightforward visualization of an attention mechanism that maps a query and a set of key-value pairs to an output. Blue squares denote the input key-value pairs. The green circle is the input query, while the black circle is the output. In addition, the red edge between a blue square and the green circle represents the vector distance. Therefore, the attention mechanism can be interpreted as shifting the green circle forward to the black circle determined by all pairwise distances.

The top-down viewpoint is beneficial for applications since it can provide theoretical evidence to select suitable modules for different scenarios. Depending on the implementation, attention modules can be reformulated as two clustering processes: K-means and k-nearest neighbors (k-NN). Assume that there is a feature space that contains all input feature vectors. The K-means style attention [806, 379, 784, 654, 105, 773] aims to capture globally shared information across all input feature vectors by intentionally

smoothing¹ out their local/individual details. Subsequently, the smoothed part of all feature vectors will span an embedded subspace, which can distill the common information that exists among those features. Therefore, we argue that K-means attention is a suitable tool for closed-form classification problems, e.g., semantic segmentation, because all common information can be captured by an adequate number of centers progressively during the training stage. In contrast, the k-NN style attention [642, 678, 66, 870, 663, 855] focuses on finding subtle differences among a set of feature vectors that are already close to each other. In other words, all old centers only compute vector distances against their corresponding k closest neighbors within relatively narrow ranges, and arbitrary new unique centers will be determined conditioned on the inputs. Therefore, we argue that k-NN attentions are efficient tools for non-closed-form classification problems, e.g., object detection/instance segmentation, since the number of objects can vary across inputs, while delicate details must be well distinguished between objects within the same classes.

In this chapter, we are mainly interested in K-means attention and focus on analyzing the learnt embedded feature spaces of deep learning models in semantic segmentation problems. We rethink the K-means attention modules from a variational Bayes (VB) perspective [629] and propose a novel attention module, namely *variational Gaussian mixture model embedded attention* (VGMEA). The graphical model of VGMEA, depicted in Fig. 3.2b, is a generalization of the classical Gaussian mixture model (GMM) [467] in computer vision (shown in Fig. 3.2a) since more regularization terms can be added through the variational parameters ϕ_n . Intuitively, the VGMEA is a flexible linear superposition of Gaussian components, aimed at providing a richer class of density models with latent variables. The latent variables can be interpreted as attention maps that compute the response at a position as a correlation with *all Gaussian centers*. The set of positions can be in space, time, or space-time, implying that our models apply to images, sequences, and video problems. Unlike the other attention modules that compute derivatives with respect to parameters for backpropagation, we apply the VB algorithm to update distribution parameters using a momentum-based moving average. In this setting, the VB algorithm encodes all the available information about the model parameters in their posterior distribution. Given the current parameters, VB minimizes the evidence lower bound (ELBO) to alternatively approximate the posterior distribution of latent variables and update variational parameters. After convergence, the output can be computed as the new centers weighted by the final attention maps.

¹removing, so-called *denoising*, highly varying information, and keeping common features.

3.1. Introduction

- First, we interpret attention mechanisms as shifting the old centers among *keys* toward new centers among *values*. Moreover, K-means attention modules are reformulated as structured probabilistic graphs that use variational Bayes algorithms to iteratively update their parameters. Our model can effectively distill global shared information from feature maps and communicate these messages back and forth between distant positions at different scales.
- Second, we build our K-means attention as a light-weighted module for deep neural networks, and set up a momentum-based moving average strategy for model parameters' maintenance and normalization. Complicated regularization terms, e.g., total variation, are incorporated to control message passing among features of different scales.
- Third, extensive experiments are conducted on four challenging semantic segmentation datasets, including PASCAL Context, Cityscapes, MS-COCO, and ADE20K, demonstrating that the proposed model can interpret many other state-of-the-art methods as specific instances. The proposed model is generic in multi-scale attention learning and achieves competitive performance compared with state-of-the-art methods.



Figure 3.2: The directed acyclic graphs represent the mixture of Gaussian models. The box (plate) denotes a set of N i.i.d. observations and latent variables. Each node of x_n in blue is an observation. Meanwhile, (μ, C) denotes a set of K model variables, where μ denotes $\{\mu_k\}$ and C denotes $\{C_k\}$. (a) Classical GMM corresponds to an observation with a latent variable z_n and a variational variable ϕ_n . z_n comprises a K binary vector that represents the correlations between position n and K centers. It is determined by a Dirichlet, ϕ_n is a K -Dimensional real vector on a decision surface that can integrate geometrical regularization. (b) Our VGMEA corresponds to an observation with a latent variable z_n and a variational variable ϕ_n . z_n comprises a K binary vector that represents the correlations between position n and K centers. Moreover, ϕ_n is a K real vector on a decision surface that can integrate geometrical regularization.

There are several advantages of using VGMEA models: (a) VGMEA can be easily combined as **non-parametric** modules in the end-to-end training of existing neural networks, due to self-updating parameters via the VB algorithm, adding no extra derivatives to the computational graphs; (b) since decoupled from the computation graphs, VGMEA supports flexible inner designs, e.g. adopts non-differentiable distributions or changes the number of iterations, that will not affect the original model architectures; (c) VGMEA is light-weighted and achieve their best results with only few parameters; (d) VGMEA supports high resolutions of the variable input sizes with relatively low computational complexity. We conduct extensive comparisons with baselines to demonstrate that different modules can be generically interpreted as specific instances of our proposed framework. Moreover, comprehensive experiments across multiple public datasets clearly demonstrate the performance of the derived methods. To summarize, the contribution of this chapter is the following: The remainder of the chapter is organized as follows. Section 3.2 introduces all related work. In Section 3.3, we illustrate the proposed framework and its interpretation, and in Section 3.4, we present our model in detail. In Section 3.5, we demonstrate the performance of the model, including comparison experiments, experimental evaluation, and analysis. Finally, the conclusion and future follow-up research are presented in Section 3.6.

3.2 Related Work

3.2.1 Variational Bayesian methods

Variational Bayes (VB), also called Variational Inference or Variational Approximation, is a family of graphical models for approximating intractable distributions arising in Bayesian inference. The node set typically consists of observed variables (i.e., data) and unobserved variables (i.e., parameters and latent variables), while the edge set represents various sorts of dependencies among all variables. VB can be seen as an extension of the expectation-maximization (EM) algorithm that upgrades maximum a posteriori (MAP) estimation to fully Bayesian estimation. Instead of performing point estimation on each parameter, VB computes an analytical approximation to the entire posterior distribution of unobserved variables. The approximated distribution, called the variational distribution, is usually a factorization of the true posterior distribution. Each factor can be restricted to a family of distributions with a simpler form at the cost of cutting off some edges. To increase approximation precision, extra variational

3.2. Related Work

parameters are subsequently attached to factors, model the missing dependencies, and are absorbed into each corresponding factor [195]. VB optimizes the evidence lower bound (ELBO) to alternatively recognize the variational distribution and update parameters. Although VB is scalable to large data sets and applicable to inferring complex models with a large number of parameters, it has not been widely used in the deep learning community because a large number of trials is often required to find a suitable factorization. If the variational distribution has been determined, VB produces locally optimal but exact analytical solutions at a faster rate than sampling methods, e.g., Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC). Our VGMEA model introduces a set of extra variational parameters to support complicated regularization terms, e.g., total-variance, to force attention maps to be smooth and compact. Moreover, the proposed model adopts local variational methods [37] to ensure that the posterior distribution is the conjugate of the prior distribution. In other words, the Bayesian update of our approach can be safely reformulated as an incremental mini-batch update, which better aligns with deep neural networks and achieves fast and stable convergence. Besides, VB can derive a lower bound for the marginal likelihood of the observed data ², which is typically used as a metric to select among models. The general idea is that a higher marginal likelihood of a model indicates a better fit to the data and, subsequently, a greater probability of generating the data [443].

3.2.2 Clustering Algorithms

Clustering methods aim to group a set of objects so that objects in the same cluster are more conceptually similar to each other than to those in other clusters. A feature vector represents each object to describe its task-specific attributes, while the conceptual similarity is quantified as a spatial distance between feature vectors by a given metric. Based on the distance function, all feature vectors are mapped onto a feature space, depicted in Figure 3.3b. Clustering methods are simple yet general algorithms that can be applied to arbitrary feature spaces to distinguish groups of concepts. According to different viewpoints, there are mainly two categories of clustering approaches: k-nearest neighbors (k-NN) and K-means-style clustering methods.

K-NN style clustering is a family of bottom-up algorithms with two hyperparameters. One hyperparameter controls the search scope, while the other determines the fusion thresholds. Given two parameters, the approach directly uses the original feature vectors and computes a center for each object among corresponding nearby

²sometimes called the evidence, i.e., the marginal probability of the data by performing an integral over unobserved variables

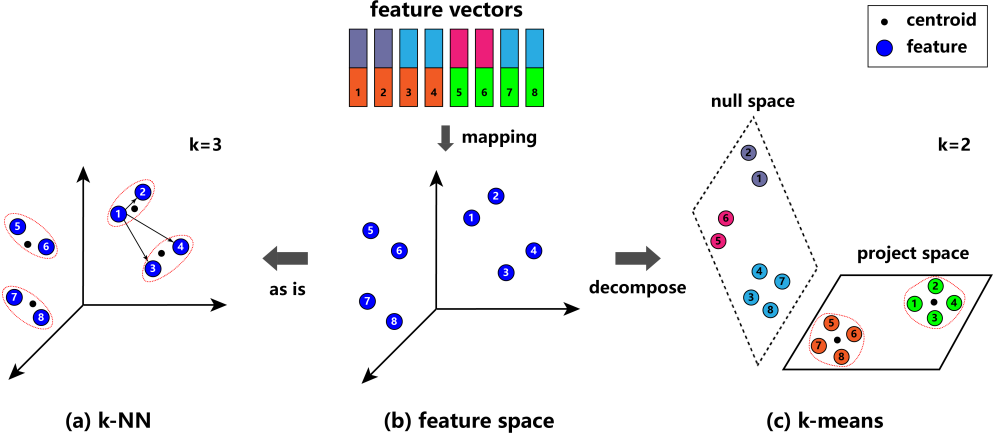


Figure 3.3: A visualization of clustering algorithms. There are 8 feature vectors, divided into 2 parts, each with an index. The lower part represents common information, and the upper part represents individual information. Different colors of features indicate greater distance in the feature space. Therefore, the first 4 feature vectors are from one class and the rest from the other. Moreover, every 2 feature vectors from one object. (a) k-NN clustering directly uses the original feature vectors and assigns each point to its 3 nearest neighbors. Through subtle differences between the individual information, 2 dots belong to one center; (b) All feature vectors are mapped as blue dots onto feature space; (c) K-means clustering divided each feature vector into two parts. The lower parts (i.e., common information) are projected onto an embedding space. In contrast, the upper parts are discarded into a null space. 2 global centers for 2 classes are obtained via significant differences between the common information.

objects, depicted in Figure 3.3a. In other words, k-NN style clustering highlights subtle differences among spatially close features and computes adaptive centers for each input feature. Notably, some centers can be very close to each other. When pairwise center distances are below a given fusing threshold, those centers will be regarded as duplicates and fused. Therefore, the effective number of centers depends on the input features.

In contrast, K-means style clustering is a family of top-down algorithms with a single hyperparameter: the number of centers. Given the number of centers, the method decomposes the original feature space into a null space and an embedding space. The embedding space contains common information, whilst the null space contains the neglected (also called noise). Subsequently, centers are computed within the embedding space, depicted in Figure 3.3c. Therefore, K-means clustering targets significant differences among the common parts and shares centers for all input features.

In summary, K-means style clustering can distill globally shared information, but k-NN style clustering can focus on locally subtle information. Although they are interchangeable in many application scenarios, we argue that K-means style clustering is more suitable for semantic segmentation, and K-NN style clustering for object

3.2. Related Work

detection/instance segmentation. Indeed, K-means can increase the number of centers for object-level clustering, but it is inefficient and will fail when objects are dense [725]. Likewise, k-NN can expand its search scope to find common information, but its variable clusters might be under- or over-clustering of semantic concepts.

3.2.3 Attention Mechanism

Attention mechanisms, first as self-attention, originated from natural language processing (NLP) for sequential data (e.g., speech, language) [642], and then were generalized as a non-local operator [678] for spatial (e.g., image) and temporal-spatial data (e.g., video). Thanks to capturing *long-range* dependencies between pairwise positions, self-attention has achieved promising results on vision-related tasks, including image retrieval [74, 175], image classification [33, 601, 508, 285, 757, 453, 547], and visual tracking [871]. However, self-attention uses a single feature map as input and computes only long-range interactions within a single scale, thereby failing to capture broader dependencies across feature maps at different scales. Therefore, cross-attention was proposed by CrossViT [66], and includes self-attention as a specific instance. Cross-attention takes two feature maps from multiple scales as inputs and computes long-range dependencies across scales. This extension has further improved the performance on visual tasks, including image super-resolution [486, 296, 129], semantic segmentation [184, 287, 448], and instance segmentation [177, 109, 209]. However, these attention modules need to produce large attention maps, which, in turn, results in high computational complexity. The bottleneck lies in the fact that both the generation of the attention map and its use are computed across all positions.

There are many improvements towards addressing the issue. EncNet [806] defined k code-words and computed attention as Euclidean distances between input feature maps and k code-words. EMA [379] reformulated attention as a reconstruction problem and computed a more compact basis set via the expectation maximization algorithm. OCRNet [784] simplified attention to a pure clustering with given learned segmentation masks. Deformable attention [870] sampled at k learnable offsets, and computed correlations between the center and its k selected neighbors. Mask2former [105] defined k global queries, and computed attention as Cosine similarity between k queries and the input *key* feature maps. Max-DeepLab [654, 773] conducted attention as a hard assignment of k -means clustering. K -NN attention [663] adopted top k most similar positions. Our model is a probabilistic model that can interpret K-means attention modules as specific instances. Moreover, multi-head attention can be explained as

independent block matrix computation, and the mini-batch training strategy can work well with the Bayesian update in our model.

3.3 Variational Gaussian Mixture Embedded Attention

In this section, we introduce the variational Gaussian mixture embedded attention (VGMEA) to interpret the attention mechanism in a probabilistic way. A VGMEA is an unifying framework that can capture long-range information across scales, distill global concepts, and compute a residual to highlight semantically oriented information. Given multi-scale feature maps $X = \{X_s\}_{s=1}^S$ that extracts from an input image $I \in \mathbb{R}^{H \times W \times 3}$ by a backbone network, where $X_s \in \mathbb{R}^{C \times N_s}$ is a feature map at scale s with $N_s = H_s \times W_s$ spatial positions. We define the VGMEA as a mapping $F : X \rightarrow Y$, where $Y \in \mathbb{R}^{C \times N}$ are enhanced and semantically oriented feature maps distilled from semantic concepts.

VGMEA updates the following global information.

Semantic concept centroids:

$$\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^K \in \mathbb{R}^{K \times C}$$

Semantic correlation matrix:

$$\mathbf{C} \in \mathbb{R}^{K \times K}$$

Scale-aware attention maps:

$$A = \{A_s \in \mathbb{R}^{N_s \times K}\}_{s=1}^S$$

VGMEA outputs enhanced semantic-oriented feature maps $Y \in \mathbb{R}^{C \times N}$:

$$Y_s = A_s \cdot \mathbf{C}^{-1} \cdot \boldsymbol{\mu} + X_s \quad (3.1)$$

where S , C , N , and K are the number of scales, channels, positions, and bases, respectively. Noted that the multi-scale attention maps represent scale-aware concept assignment.

3.3. Variational Gaussian Mixture Embedded Attention

3.3.1 Problem Formulation and Notation

Denote $\Theta = \{\Phi, m, \beta, W, \nu\}$ as a set of learnable parameters in the VGMEA module with an initial value $\Theta_0 = \{\Phi_0, m_0, \beta_0, W_0, \nu_0\}$. Besides, $(\boldsymbol{\mu}, \mathbf{C})$ is the distilled global concepts and $Z = [Z_1, Z_2, \dots, Z_K]$ is a latent random field to estimate the multi-scale attention maps A for concept assignment. Specifically, each position of Z is a 1-of- K vector $Z_k \in \{0, 1\}$ and satisfies $\sum_{k=1}^K Z_k = 1$. Then, the probabilistic framework of the proposed *variational Gaussian Mixture model with geometric regularization* is defined as follows:

$$\begin{aligned}
 P_{\Theta}(X, Z, \boldsymbol{\mu}, \mathbf{C}) &= P(X|Z, \boldsymbol{\mu}, \mathbf{C}) \cdot P_{\Theta}(Z) \cdot P_{\Theta}(\boldsymbol{\mu}, \mathbf{C}) \\
 P(X|Z, \boldsymbol{\mu}, \mathbf{C}) &= \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(X_n | \boldsymbol{\mu}_k, \mathbf{C}_k^{-1})^{Z_{n,k}} \\
 P_{\Theta}(Z) &= F(\Phi) \cdot \prod_{n=1}^N \prod_{k=1}^K \delta(\phi_{n,k})^{Z_{n,k}} \cdot \delta(-\phi_{n,k})^{1-Z_{n,k}} \\
 P_{\Theta}(\boldsymbol{\mu}, \mathbf{C}) &= \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | m, (\beta \cdot \mathbf{C}_k)^{-1}) \mathcal{W}(\mathbf{C}_k | W, \nu)
 \end{aligned} \tag{3.2}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution, $\mathcal{W}(\cdot)$ is a Wishart distribution. In addition, $P_{\Theta}(Z)$ is a Categorical distribution with a normalization term $F(\Phi)$ and $\delta(\cdot)$ is a sigmoid function. Based on auxiliary Φ , we can regulate the random field Z as a smooth manifold and enforce $P(X|Z, \boldsymbol{\mu}, \mathbf{C})$ generating robust label affiliation A which has continuity within semantic regions and sharp boundaries between.

3.3.2 Variational Bayesian Inference with Geometric Regularization

Set $Q_{\Theta}(Z, \boldsymbol{\mu}, \mathbf{C}) = Q_{\Theta}(Z) \cdot Q_{\Theta}(\boldsymbol{\mu}, \mathbf{C})$ as a lower bound of posterior distribution. We define $T \in \{Z \times \boldsymbol{\mu} \times \mathbf{C}\}$ as an infinitesimal variable in a Descartes product space. For simplicity, we represents $\tilde{T} = T \setminus S = \{\boldsymbol{\mu} \times \mathbf{C}\}$.

Variational Inference to Approximate Posterior Distributions

Denote the evidence lower bound (ELBO) as **K**ullback–**L**eibler divergence with an

additive geometric regularization to measure the gap:

$$\text{KL}(Q||P) = \int_T Q_\Theta(Z) Q_\Theta(\boldsymbol{\mu}, \mathbf{C}) \ln \frac{P_\Theta(X, Z, \boldsymbol{\mu}, \mathbf{C})}{Q_\Theta(Z, \boldsymbol{\mu}, \mathbf{C})} dT + \lambda \text{Reg}(\Phi) \quad (3.3)$$

Noted that it is safe to ignore the geometric term when approximating the posterior functionals because $\text{Reg}(\Phi)$ remains a constant during variational inference. Using the definition of expectation and rearranging the order of integration, we obtain the following two equations.

$$\begin{aligned} \text{KL}(Q||P) &= \int_Z Q_\Theta(Z) \cdot \int_{\tilde{T}} Q_\Theta(\boldsymbol{\mu}, \mathbf{C}) \ln \frac{P_\Theta(X, Z, \boldsymbol{\mu}, \mathbf{C})}{Q_\Theta(Z, \boldsymbol{\mu}, \mathbf{C})} d\tilde{T} dZ \\ &\propto \int_Z Q_\Theta(Z) \cdot \ln \frac{P_\Theta(Z) \cdot \exp\{\mathbb{E}_{\tilde{T}}[\ln P(X|Z, \boldsymbol{\mu}, \mathbf{C})]\}}{Q_\Theta(Z)} dZ \end{aligned} \quad (3.4)$$

where $\exp$ is an exponential function, and

$$\begin{aligned} \text{KL}(Q||P) &= \int_{\tilde{T}} Q_\Theta(\boldsymbol{\mu}, \mathbf{C}) \cdot \int_Z Q_\Theta(Z) \ln \frac{P_\Theta(X, Z, \boldsymbol{\mu}, \mathbf{C})}{Q_\Theta(Z, \boldsymbol{\mu}, \mathbf{C})} dZ d\tilde{T} \\ &\propto \int_{\tilde{T}} Q_\Theta(\boldsymbol{\mu}, \mathbf{C}) \cdot \ln \frac{P_\Theta(\boldsymbol{\mu}, \mathbf{C}) \cdot \exp\{\mathbb{E}_Z[\ln P(X|Z, \boldsymbol{\mu}, \mathbf{C})]\}}{Q_\Theta(\boldsymbol{\mu}, \mathbf{C})} d\tilde{T} \end{aligned} \quad (3.5)$$

In order to minimize the gap, we have to solve the following equations:

$$\begin{aligned} \frac{P_\Theta(Z) \cdot \exp\{\mathbb{E}_{\tilde{T}}[\ln P(X|Z, \boldsymbol{\mu}, \mathbf{C})]\}}{Q_\Theta(Z)} &= 1 \\ \frac{P_\Theta(\boldsymbol{\mu}, \mathbf{C}) \cdot \exp\{\mathbb{E}_Z[\ln P(X|Z, \boldsymbol{\mu}, \mathbf{C})]\}}{Q_\Theta(\boldsymbol{\mu}, \mathbf{C})} &= 1 \end{aligned} \quad (3.6)$$

Therefore, we obtain optimal functionals, where Con is a normalization constant for probability functions.

$$Q_\Theta(Z) = P_\Theta(Z) \cdot \exp\{\mathbb{E}_{\tilde{T}}[\ln P(X|Z, \boldsymbol{\mu}, \mathbf{C})]\} \cdot \text{Con} \quad (3.7)$$

$$Q_\Theta(\boldsymbol{\mu}, \mathbf{C}) = P_\Theta(\boldsymbol{\mu}, \mathbf{C}) \cdot \exp\{\mathbb{E}_Z[\ln P(X|Z, \boldsymbol{\mu}, \mathbf{C})]\} \cdot \text{Con} \quad (3.8)$$

3.3. Variational Gaussian Mixture Embedded Attention

Hereafter, we insert the previous probability functions into Eq.3.7 and Eq.3.8. Then we recognize the approximate posterior function $Q_{\Theta}(Z)$ as a **Categorical** distribution:

$$\begin{aligned}
Q_{\Theta}(Z) &= \prod_{n=1}^N \prod_{k=1}^K \underbrace{e^{\{\phi_{n,k} + \mathbb{E}_{\tilde{T}}[\ln \mathcal{N}(X_n | \boldsymbol{\mu}_k, C_k^{-1})]\}}}_{\rho_{n,k}}^{Z_{n,k}} \cdot \text{Con} \\
&= \prod_{n=1}^N \prod_{k=1}^K \rho_{n,k}^{Z_{n,k}} \cdot \text{Con} \\
&= \prod_{n=1}^N \prod_{k=1}^K \underbrace{\left(\frac{\rho_{n,k}}{\sum_{k=1}^K \rho_{n,k}} \right)}_{r_{n,k}}^{Z_{n,k}} \quad (3.9) \\
&= \prod_{n=1}^N \prod_{k=1}^K r_{n,k}^{Z_{n,k}}
\end{aligned}$$

Hence, we have the standard result

$$\mathbb{E}_Z[Z_{n,k}] = r_{n,k} \quad (3.10)$$

Noted that each quantity $r_{n,k}$ is a part of the softmax functions and represents a soft affiliation, which plays the role of responsibilities. Note that the optimal solution for $Q_{\Theta}(Z)$ depends on moments evaluated with respect to the distributions of other variables. Therefore, the variational update equations are coupled and must be solved iteratively. At this point, we shall find it convenient to define three statistics of the observed data set evaluated from the responsibilities, given by

$$\begin{aligned}
N_k &= \sum_{n=1}^N r_{n,k} \\
\bar{X}_k &= \frac{1}{N_k} \sum_{n=1}^N r_{n,k} X_n \\
S_k &= \frac{1}{N_k} \sum_{n=1}^N r_{n,k} (X_n - \bar{X}_k)(X_n - \bar{X}_k)^T
\end{aligned} \quad (3.11)$$

Moreover, the variational posterior distribution $Q_{\Theta}(\boldsymbol{\mu}, \mathbf{C})$ is recognized as a **Gaussian-Wishart** distribution:

$$\begin{aligned}
 Q_{\Theta}(\boldsymbol{\mu}, \mathbf{C}) & \propto \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | m, (\beta C_k)^{-1}) \cdot \mathcal{W}(C_k | W, \nu) \cdot \prod_{n=1}^N e^{\{\mathbb{E}_Z[Z_{n,k} \cdot \ln \mathcal{N}(X_n | \boldsymbol{\mu}_k, C_k^{-1})]\}} \\
 & \propto \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | m, (\beta C_k)^{-1}) \cdot \mathcal{W}(C_k | W, \nu) \cdot \prod_{n=1}^N e^{r_{n,k}} \cdot \mathcal{N}(X_n | \boldsymbol{\mu}_k, C_k^{-1}) \\
 & = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | m_k, (\beta_k C_k)^{-1}) \cdot \mathcal{W}(C_k | W_k, \nu_k)
 \end{aligned} \tag{3.12}$$

where m_k , β_k , W_k and ν_k are parameters for the posterior distribution. Based on the momentum statistics, the optimal parameters can be updated as

$$\begin{aligned}
 \beta_k &= \beta_0 + N_k \\
 m_k &= \frac{\beta_0 m_0 + N_k \bar{X}_k}{\beta_0 + N_k} \\
 W_k^{-1} &= W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{X}_k - m_0)(\bar{X}_k - m_0)^T \\
 \nu_k &= \nu_0 + N_k
 \end{aligned} \tag{3.13}$$

Geometric Regularization in Concept Space

In addition, our variational probabilistic model is flexible to integrate geometric regularization terms Φ for the regularity of the responsibility $r_{n,k}$. For simplicity, we integrate total variance (TV) as an example. Denote the likelihood function as $L(\Theta, \Theta^*)$. Under the expectation conception, all hidden random variables become constants. Therefore, the likelihood function only contains geometric regularization parameters for optimization.

$$\begin{aligned}
 L(\Theta, \Theta^*) &= \int_T Q_{\Theta^*}(Z, \boldsymbol{\mu}, \mathbf{C}) \ln \frac{P_{\Theta}(X, Z, \boldsymbol{\mu}, \mathbf{C})}{Q_{\Theta^*}(Z, \boldsymbol{\mu}, \mathbf{C})} dT + \lambda \text{Reg}(\Phi) \\
 &= \sum_{n=1}^N \sum_{k=1}^K r_{n,k} \cdot \phi_{n,k} + \ln \delta(-\phi_{n,k}) + \lambda \text{Reg}(\phi_{n,k}) + \text{Con.} \\
 &= \sum_{n=1}^N \sum_{k=1}^K r_{n,k} \cdot \phi_{n,k} + \ln \delta(-\phi_{n,k}) + \lambda \cdot |\nabla \delta(\phi_{n,k})| + \text{Con.}
 \end{aligned} \tag{3.14}$$

3.3. Variational Gaussian Mixture Embedded Attention

Where Con. contains other irrelevant model parameters. Then we compute the derivative and set it to zero:

$$\begin{aligned}
\frac{\partial L(\Theta, \Theta^*)}{\partial \phi_{n,k}} &= \frac{d}{d\phi_{n,k}} [r_{n,k} \cdot \phi_{n,k} + \ln \delta(-\phi_{n,k}) + \lambda \cdot |\nabla \delta(\phi_{n,k})|] \\
&= r_{n,k} - \delta(\phi_{n,k}) + \lambda \cdot \delta'(\phi_{n,k}) \cdot \underbrace{\text{div}\left(\frac{\nabla \phi_{n,k}}{|\nabla \phi_{n,k}|}\right)}_{\kappa_{n,k}} \\
&= r_{n,k} - \delta(\phi_{n,k}) + \lambda \cdot \delta'(\phi_{n,k}) \cdot \kappa_{n,k} \\
&= 0
\end{aligned} \tag{3.15}$$

It is a difficult differential equation to solve. However, we can follow the routine of the level-set method that sets $\frac{\partial L(\Theta, \Theta^*)}{\partial \phi_{n,k}} = -\frac{d\phi_{n,k}}{dt}$. Then we have a momentum-based update equation:

$$\begin{aligned}
\frac{d\phi_{n,k}}{dt} &= \delta(\phi_{n,k}) - r_{n,k} - \lambda \cdot \delta'(\phi_{n,k}) \kappa_{n,k} \\
\rightarrow \phi_{n,k}^{t+1} &= \phi_{n,k}^t + \nabla t \left[\underbrace{\delta(\phi_{n,k})}_{\text{sigmoid}} - \underbrace{r_{n,k}}_{\text{softmax}} - \underbrace{\lambda \delta'(\phi_{n,k}) \kappa_{n,k}}_{\text{smoothness}} \right]
\end{aligned} \tag{3.16}$$

where $\nabla t \geq 0$ is a scaling factor to update the manifold surface and λ is a weight for smoothness regularization. Noted that the geometric regularization introduces smoothness into the softmax responsibility function and generates light-tailed distributions. It is beneficial to generate piece-wise smoothing label affiliation maps for $r_{n,k}$, i.e., consistent assignment within each semantic region with sharp interval boundaries.

3.3.3 Cross-Attention as Concept Distillation

Modern vision transformers implement a cross-attention module as a series of blocks. Each block adopts a residual structure and updates the query once. Hence, the entire cascade cross-attention module performs a unified formulation for iterative concept refinement. Given query features $\mathbf{Q} \in \mathbb{R}^{N_q \times C}$, key features $\mathbf{K} \in \mathbb{R}^{N_k \times C}$, and value

features $\mathbf{V} \in \mathbb{R}^{N_k \times C}$, the iterative residual formulation can be reformulate as:

$$\begin{aligned}\nabla \mathbf{Q}^{(l)} &= \text{softmax} \left(\frac{\mathbf{Q}^{(l)} \mathbf{K}^\top}{\sqrt{d_k}} \right) \cdot \mathbf{V} \\ \mathbf{Q}^{(l+1)} &= \mathbf{Q}^{(l)} + \nabla \mathbf{Q}^{(l)}\end{aligned}\tag{3.17}$$

where l represents the block index. This operation can be interpreted as a mean-shift process that sequentially computes the query shifts $\nabla \mathbf{Q}$ and updates the query using a momentum update with even weights 1.

The proposed VGMEA generalizes the mean-shift process as a joint probabilistic model and interprets it through EM algorithms:

$$\begin{aligned}\textbf{E-step (Assignment): } r_{n,k}^{(l)} &= \mathbb{E}_Z[Z_{n,k}] \\ \textbf{M-step (Update): } \mu_k^{(l+1)} &= \mu_k^{(l)} + \eta \bar{X}_k \\ \mathbf{C}_k^{(l+1)} &= \mathbf{C}_k^{(l)} + \eta \bar{S}_k\end{aligned}\tag{3.18}$$

where $\mathbb{E}_Z[Z_{n,k}]$ is the expectation of the approximated posterior distribution. η is the momentum update factor. $\bar{X}_k$ and $\bar{S}_k$ are sampled mean and correlation estimated from responsibility $r_{n,k}$, respectively.

Method	Assignment	Update	Correspondence
Non-local*[678]	$r_{i,j} = \Upsilon \left(e^{\mathbf{x}_i^T \mathbf{q}_j} \right)$	$\mu_j = W_z \cdot \bar{X}_j + \mathbf{q}_j, \mathbf{C}_j \equiv I$	$\eta = W_z, \bar{X}_j = \sum r_{i,j} \cdot g(\mathbf{x}_i), \phi_i \equiv 0$
EncNet[806]	$r_{i,j} = \Upsilon \left(e^{-s_k \ \mathbf{d}_{i,j}\ ^2} \right)$	$\mu_k = h(r_{i,j} \cdot \mathbf{d}_{i,j}), \mathbf{C}_j \equiv I$	$\eta = 0, \bar{X}_j \equiv 0, \phi_i \equiv 0$
EMANet[379]	$r_{i,j} = \Upsilon \left(e^{-(\mathbf{x}_i - \mathbf{q}_j)^T \mathbf{C}_j^{-1} (\mathbf{x}_i - \mathbf{q}_j)} \right)$	$\mu_j = \bar{X}_j, \mathbf{C}_j = \mathbf{S}_j$	$\bar{X}_j = N_j^{-1} \sum r_{i,j} \cdot \mathbf{x}_i, \phi_i \equiv 0$ $\mathbf{S}_j = N_j^{-1} \sum r_{i,j} \cdot D^2(\mathbf{d}_{i,j})$
OCRNet[784]	$r_{i,j} = \mathbf{m}_{i,j}$	$\mu_j = \bar{X}_j, \mathbf{C}_j \equiv I$	$\bar{X}_j = \sum \mathbf{m}_{i,j} \mathbf{x}_i, \phi_i \equiv 0$
kMax-DeepLab[773]	$r_{i,j} = \arg \max(\mathbf{q}_j \mathbf{k}_i^T)$	$\mu_j = \mathbf{q}_j + \bar{X}_j, \mathbf{C}_j \equiv I$	$\eta = 1, \bar{X}_j = \sum r_{i,j} \cdot \mathbf{v}_i, \phi_i \equiv 0$
MaskFormer[106]	$r_{i,j} = \Upsilon(q_j \mathbf{k}_i^T)$	$\mu_j = \mathbf{q}_j + \bar{X}_j, \mathbf{C}_j \equiv I$	$\eta = 1, \bar{X}_j = \sum r_{i,j} \cdot \mathbf{v}_i, \phi_i \equiv 0$
Mask2Former[105]	$r_{i,j} = \Upsilon(\mathbf{q}_k \mathbf{k}_i^T + \mathcal{M}_i)$	$\mu_j = \mathbf{q}_j + \bar{X}_j, \mathbf{C}_j \equiv I$	$\eta = 1, \bar{X}_j = \sum r_{i,j} \cdot \mathbf{v}_i, \phi_i = \mathcal{M}_i$

Table 3.1: Special Interpret SOTA models as VGMEA Instances. (1) Non-local[678] is implemented as the cross-attention and $g(\cdot)$ denotes transformation function. (2) $\Upsilon(\cdot) = \text{softmax}(\cdot)$, $\mathbf{m}_{i,k}$ denotes a semantic segmentation prediction from a sub-network. (3) $h(\cdot)$ denotes Batch Normalization with ReLU activation and $\mathbf{d}_{i,k} = \mathbf{x}_i - \mathbf{q}_k$. (4) $\kappa(\mathbf{x}, \mathbf{y}) = f_1(\mathbf{x})^T f_2(\mathbf{y})$ where $f_1(\cdot)$ and $f_2(\cdot)$ are two transformation functions. (5) $\mathcal{M}_i$ is the binarized output of the resized mask prediction, $D^2(\mathbf{x}) = \mathbf{x} \cdot \mathbf{x}^T$.

To demonstrate generalization, we analyze a few state-of-the-art models as special

3.4. Implementation Details

cases of VGMEA and organize them in Table 3.1. We observe that all methods set the correlation matrix $\mathbf{C}$ to the identity matrix, except EMANet, suggesting that semantic centroids are orthonormal. Furthermore, none of the methods can support geometrical constraints to further sharpen cross-class decision boundaries, except Mask2Former. More specifically, Non-local[678] is a generalized K-means clustering algorithm that allows transformations between different feature representations. However, it cannot be stacked to update semantic concepts μ due to its high computational complexity. In addition, EncNet[806] defined a global semantic codebook and computed cross-attention by averaging the residuals between the input features and the codebook. The averaged residuals are interpreted as context encoding to guide semantic segmentation, suppressing absent classes. Nevertheless, EncNet cannot extract sample statistics $\bar{X}$ from the input feature maps and therefore fails to adapt the context encoding to the input. In contrast, EMANet[379] only considers sample statistics and sample correlation, resulting in vulnerability to the batch size and data noise. In contrast, OCRNet [784] is an amortized K-means clustering algorithm that relies heavily on predicted semantic segmentation masks to compute the statistics $\bar{X}$. It used $\bar{X}$ as dynamic queries to compute cross-attention. However, its bottleneck is the quality of mask prediction and the batch size. kMax-DeepLab[773] follows the conventional K-means clustering algorithms and adopts hard assignment to compute sample statistics $\bar{X}$. MaskFormer[106] is a simplified version that does not use geometrical constraints to ensure cross-class decision boundaries. Likewise, Mask2Former [106] adopts amortized constraints represented as binarized mask predictions. In summary, our VGMEA is a generalized probabilistic framework that can accommodate a wide range of cross-attention mechanisms.

3.4 Implementation Details

3.4.1 Online Bayesian Updates for Batch Processing

Deep learning pipelines process data in mini-batches. However, traditional EM requires full-dataset iterations, making it incompatible with stochastic optimization. Therefore, we reformulate VGMEA with the Bayesian formula to support incremental/iterative updates.

Given a dataset of M independent batches $X = \{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_M\}$ and any mini-

batch $\mathcal{B}_i$, the probabilistic model is equal to a chain of conditional distributions:

$$\begin{aligned}
 P_{\Theta}(X, Z, \boldsymbol{\mu}, \mathbf{C}) &= P_{\Theta}(X \setminus \mathcal{B}_i, \mathcal{B}_i, Z, \boldsymbol{\mu}, \mathbf{C}) \\
 &= P_{\Theta}(\mathcal{B}_i, Z, \boldsymbol{\mu}, \mathbf{C} | X \setminus \mathcal{B}_i) \cdot P(X \setminus \mathcal{B}_i) \\
 &= P_{\Theta}(\mathcal{B}_i | Z, \boldsymbol{\mu}, \mathbf{C}, X \setminus \mathcal{B}_i) \cdot P_{\Theta}(Z, \boldsymbol{\mu}, \mathbf{C} | X \setminus \mathcal{B}_i) \cdot P(X \setminus \mathcal{B}_i) \\
 &\propto P_{\Theta}(\mathcal{B}_i | Z, \boldsymbol{\mu}, \mathbf{C}, X \setminus \mathcal{B}_i) \cdot P_{\Theta}(Z, \boldsymbol{\mu}, \mathbf{C} | X \setminus \mathcal{B}_i) \\
 &\approx P_{\Theta}(\mathcal{B}_i | Z, \boldsymbol{\mu}, \mathbf{C}, X \setminus \mathcal{B}_i) \cdot P_{\Theta}(Z | X \setminus \mathcal{B}_i) \cdot P_{\Theta}(\boldsymbol{\mu}, \mathbf{C} | X \setminus \mathcal{B}_i) \\
 &= P_{\Theta}(\mathcal{B}_i | Z, \boldsymbol{\mu}, \mathbf{C}, X \setminus \mathcal{B}_i) \cdot Q_{\Theta}^{(t-1)}(Z) \cdot Q_{\Theta}^{(t-1)}(\boldsymbol{\mu}, \mathbf{C})
 \end{aligned} \tag{3.19}$$

where $Q_{\Theta}^{(t-1)}(Z)$ and $Q_{\Theta}^{(t-1)}(\boldsymbol{\mu}, \mathbf{C})$ are the approximated variational posterior distributions based on an arbitrary subset of the dataset. Thanks to the conjugate priors $Q_{\Theta}^{(t-1)}(Z)$ and $Q_{\Theta}^{(t-1)}(\boldsymbol{\mu}, \mathbf{C})$, the generative distribution can be approximated as a VGMEA at any bipartite dataset. Formally, the VGMEA can adapt the EM algorithm into an online update scheme to compute the posterior distribution:

$$\begin{aligned}
 P_{\Theta}(X \setminus \mathcal{B}_i, Z, \boldsymbol{\mu}, \mathbf{C} | \mathcal{B}_i) &= \frac{P_{\Theta}(X \setminus \mathcal{B}_i, \mathcal{B}_i, Z, \boldsymbol{\mu}, \mathbf{C})}{P(\mathcal{B}_i)} \\
 &\propto P_{\Theta}(\mathcal{B}_i | Z, \boldsymbol{\mu}, \mathbf{C}, X \setminus \mathcal{B}_i) \cdot Q_{\Theta}^{(t-1)}(Z) \cdot Q_{\Theta}^{(t-1)}(\boldsymbol{\mu}, \mathbf{C}) \\
 &\approx Q_{\Theta}^{(t)}(Z) \cdot Q_{\Theta}^{(t)}(\boldsymbol{\mu}, \mathbf{C})
 \end{aligned} \tag{3.20}$$

with conjugate priors

$$\begin{aligned}
 P_{\Theta}(\mathcal{B}_i | Z, \boldsymbol{\mu}, \mathbf{C}, X \setminus \mathcal{B}_i) &\sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \\
 Q_{\Theta}^{(t-1)}(\mu_k, C_k) &\sim \mathcal{N}\left(m_k^{(t)}, (\beta_k^{(t)} C_k)^{-1}\right) \cdot \mathcal{W}\left(W_k^{(t)}, \nu_k^{(t)}\right) \\
 Q_{\Theta}^{(t-1)}(Z_{n,k}) &\sim \text{Cat}(\phi_{n,k})
 \end{aligned} \tag{3.21}$$

Therefore, we can formulate the standard incremental update rules for the VGMEA.

3.4. Implementation Details

Let $(N_k^{(\mathcal{B})}, \bar{X}_k^{(\mathcal{B})}, S_k^{(\mathcal{B})})$ be sufficient statistics at mini-batch $\mathcal{B}$:

$$\begin{aligned} N_k^{(\mathcal{B})} &= \sum_{i \in \mathcal{B}} r_{i,k} \\ \bar{X}_k^{(\mathcal{B})} &= \frac{1}{N_k^{(\mathcal{B})}} \sum_{i \in \mathcal{B}} r_{i,k} X_i \\ S_k^{(\mathcal{B})} &= \frac{1}{N_k^{(\mathcal{B})}} \sum_{i \in \mathcal{B}} r_{i,k} \left(X_i - \bar{X}_k^{(\mathcal{B})} \right) \left(X_i - \bar{X}_k^{(\mathcal{B})} \right)^T \end{aligned} \tag{3.22}$$

Therefore, the model parameters $\Theta = \{\mathbf{m}_k, \beta_k, \mathbf{W}_k, \nu_k\}$ can be updated as:

$$\begin{aligned} \beta_k^{(t)} &= \beta_k^{(t-1)} + N_k^{(\mathcal{B})} \\ m_k^{(t)} &= \frac{\beta_k^{(t-1)} m_k^{(t-1)} + N_k^{(\mathcal{B})} \bar{X}_k^{(\mathcal{B})}}{\beta_k^{(t-1)} + N_k^{(\mathcal{B})}} \\ (W_k^{(t)})^{-1} &= (W_k^{(t-1)})^{-1} + N_k^{(\mathcal{B})} S_k^{(\mathcal{B})} + \frac{\beta_k^{(t-1)} N_k^{(\mathcal{B})}}{\beta_k^{(t-1)} + N_k^{(\mathcal{B})}} \epsilon_{\mathcal{B}}^{(t-1)} \\ \nu_k^{(t)} &= \nu_k^{(t-1)} + N_k^{(\mathcal{B})} \end{aligned} \tag{3.23}$$

where $\epsilon_{\mathcal{B}}^{(t-1)} = \left(\bar{X}_k^{(\mathcal{B})} - m_k^{(t-1)} \right) \left(\bar{X}_k^{(\mathcal{B})} - m_k^{(t-1)} \right)^T$. We have reformulated the EM algorithm to support mini-batch style based on the variational inference with Bayesian update formula. It implies the existence of a firm mathematical foundation for adapting large probabilistic models to deep learning models without losing generality.

3.4.2 Adaptions to Deep Learning

For efficiency and stability, we have adapted VGMEA for deep learning models.

Streaming update with Smoothing

To mitigate batch-size effects, we employ streaming averaging to update the sufficient

statistics:

$$\begin{aligned}
N_k^{(t+1)} &= N_k^{(t)} + N_k^{(\mathcal{B})} \\
\bar{X}_k^{(t+1)} &= \alpha \bar{X}_k^{(t)} + (1 - \alpha) \bar{X}_k^{(\mathcal{B})} \\
S_k^{(t+1)} &= \alpha S_k^{(t)} + (1 - \alpha) S_k^{(\mathcal{B})} + O(\alpha^2) \epsilon_S^{(t)}
\end{aligned} \tag{3.24}$$

where $\alpha = \frac{N_k^{(t)}}{N_k^{(t)} + N_k^{(\mathcal{B})}}$, $O(\alpha^2) = \frac{N_k^{(t)} N_k^{(\mathcal{B})}}{(N_k^{(t)} + N_k^{(\mathcal{B})})^2}$ and $\epsilon_S^{(t)} = \left(\bar{X}_k^{(t)} - \bar{X}_k^{(\mathcal{B})} \right) \left(\bar{X}_k^{(t)} - \bar{X}_k^{(\mathcal{B})} \right)^T$. Noted that our streaming method can always achieve stable states with arbitrary mini-batch sizes, since $(1 - \alpha) \rightarrow 0$ and $O(\alpha^2) \rightarrow 0$ when $N_k^{(t+1)}$ accumulate to large numbers with training process.

In addition, we introduce an extra momentum smoothing parameter $\gamma = 0.99$ to stabilize the early-stage training for convergence acceleration:

$$\begin{aligned}
N_k^{(t+1)} &= \gamma N_k^{(t+1)} + (1 - \gamma) N_k^{(t)} \\
\bar{X}_k^{(t+1)} &= \gamma \bar{X}_k^{(t+1)} + (1 - \gamma) \bar{X}_k^{(t)} \\
S_k^{(t+1)} &= \gamma S_k^{(t+1)} + (1 - \gamma) S_k^{(t)}
\end{aligned} \tag{3.25}$$

Orthogonality Constraint

From the perspective of semantic segmentation, enforcing orthogonality among concept centroids is well-motivated both theoretically and practically. The core justification stems from the fundamental requirement that distinct semantic concepts should be linearly independent in the feature space. The precondition means that the feature representation of one class should not be representable as a linear combination of the features of other irrelevant classes. Linear independence is a prerequisite for a model to make unambiguous and confident predictions.

Orthogonality is a stronger and more desirable condition than mere linear independence. By enforcing $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j \approx 0$ for $i \neq j$, we ensure that the learned semantic concepts are maximally separable and reside in independent ranges of the feature space. Orthogonality directly mitigates feature entanglement, a common issue where models struggle to distinguish between visually similar classes. The orthogonality constraint acts as a powerful regularizer, guiding the learning process towards a feature space where the decision boundaries between classes are sharp and well-defined.

Computational advantages primarily drive the transition from encouraging lin-

3.4. Implementation Details

ear independence to enforcing strict orthogonality. From a probabilistic standpoint, orthogonality allows us to make a critical simplification: it justifies the assumption of a diagonal precision matrix $\mathbf{C}_k^{-1}$. When concept centroids are orthogonal, the covariances between different feature dimensions are minimized. Consequently, the full Wishart distribution $\mathcal{W}(\mathbf{W}_k, \nu_k)$, which governs the prior over the full precision matrix, elegantly factorizes into a product of independent Gamma distributions for each diagonal element λ_{km} :

$$P(\mathbf{C}_k) = \mathcal{W}(\mathbf{C}_k^{-1} | W, \nu) = \prod_{m=1}^K \text{Gamma}(\lambda_{km} | \alpha_{km}, \beta_{km}) \quad (3.26)$$

This factorization improves computational efficiency in online Bayesian update for posterior inference by replacing Gaussian-Wishart distributions with Gaussian-Gamma distributions. The equivalence benefits the matrix inversion calculation for $\mathbf{W}_k^{-1}$ and $\mathbf{C}_k^{-1}$ in the mini-batch EM algorithm. Compared to updating two full inverse matrices, we can apply the reciprocal function only along all diagonal elements, reducing the complexity from $O(K^3)$ to $O(K)$. In practice, we enforce concept orthogonality via:

$$L_{\text{orth}} = \sum_{i \neq j} |\mu_i^T \mu_j| + \sum_k \|C_k - \text{diag}(C_k)\|_F^2 \quad (3.27)$$

where $\text{diag}(\cdot)$ is a diagonal matrix. The loss function ensures that semantic concepts are disentangled and that covariance matrices remain diagonal, thereby reducing computational complexity. In summary, orthogonality is not merely a mathematical convenience but also a structurally sound inductive bias for semantic segmentation. It ensures the learning of distinct, disentangled semantic concepts while enabling a highly efficient and scalable implementation of the probabilistic model by leveraging the computational simplifications afforded by diagonal covariance matrices.

Amortized Geometric Constraints

The geometric constraints can be a computational bottleneck in the VGMEA model, since iterative level-set Φ calculations typically require more than 15 iterations to converge from scratch for each new input, increasing training time by 50%. Therefore, we adopt amortized variational inference (AVI) for geometric constraints, reaching a balance between accuracy and speed. AVI amortizes the cost by finding an input-dependent initialization rather than solving from $\Phi = \mathbf{0}$. We train a single, lightweight neural network (the ϕ -predictor) to directly predict a coarse $\hat{\Phi}$ from each input feature

as a good start state. Only if the predicted $\hat{\Phi}$ is very close to the optimum, a short, subsequent refinement loop (e.g., 3 iterations instead of 15) is then sufficient to converge. The strategy narrows the "amortization gap" – the difference between the amortized and iterative solutions. The predictor network can be extremely simple – often just several convolutional layers. The setting requires a minimal number of parameters and FLOPs while yielding substantial computational savings by drastically reducing the number of refinement iterations. The forward pass of the predictor is computationally cheaper than a long optimization iteration.

In conclusion, from an engineering standpoint, AVI can serve as a learning-based approximation, a necessary architectural decision for deploying probabilistic models with complex constraints in practical computer vision systems. It successfully transforms an intractable, iterative inference problem into a tractable, feed-forward prediction problem. The minimal sacrifice in theoretical purity and a negligible drop in final performance are overwhelmingly justified by the orders-of-magnitude improvement in inference speed, predictability, and scalability. It makes a powerful but computationally intensive model like VGMEA feasible for improving attention maps, ensuring spatial coherence and alignment with object boundaries.

3.5 Experiment

3.5.1 Dataset and Metrics

Dataset: We conduct comprehensive evaluations on four challenging semantic segmentation benchmarks to assess the generality and robustness of our proposed method. PASCAL Context[484] is a detailed scene understanding dataset built upon PASCAL VOC 2010[167]. It provides pixel-wise semantic labels for the same images, extending beyond the core object classes. The dataset contains over 400 categories, including both "thing" (countable) and "stuff" (amorphous) classes, but we follow the convention of using only 59 classes for semantic segmentation. With 10,103 annotated images, it offers a more comprehensive understanding of an entire scene's context for benchmarking. Cityscapes [123] is a large-scale dataset focused on urban street scenes. It contains high-resolution images (2048×1024) with annotations across 19 semantic categories relevant for autonomous driving. The dataset provides 5000 fine-grained annotations (2,975 images for training, 500 for validation, and 1,525 for testing) and 20,000 extra coarse-grained annotations for training. Its emphasis on precise boundary delineation

3.5. Experiment

makes it ideal for evaluating geometric accuracy. MS-COCO[415] is a large-scale object detection, segmentation, and captioning dataset. For semantic segmentation, we use the train2017, val2017, and test2017 splits, which contain 118,287, 5,000, and 40,670 images, respectively. The dataset features complex daily scenes with 80 object categories. Its large scale and dense, cluttered scenes make it a rigorous test for a model’s robustness and generalization capability. ADE20K[850] is one of the most challenging benchmarks for semantic segmentation, containing 150 finely annotated categories including both objects and stuff. The dataset is split into 20,210 images for training, 2,000 for validation, and 3,352 for testing. Its large number of categories and complex scene compositions make it particularly suitable for evaluating model semantic scalability. To systematically investigate the relationship between interpretability and scalability in cross-attention mechanisms, we select four representative datasets with progressively increasing semantic complexity (ranging from 19 to 150 classes), simulating a continuum of segmentation difficulty.

Metrics: We use two standard metrics to evaluate semantic segmentation performance. The Mean Intersection over Union (mIoU) is the most widely adopted metric for semantic segmentation, providing a balanced measure of region similarity. It emphasizes overall region accuracy but is particularly sensitive to large objects. The Dice coefficient measures the overlap between the predicted and ground-truth regions. It is more sensitive to boundary accuracy and performs better at evaluating the segmentation of small objects than mIoU. To assess the efficiency of the parameters and computations, we define a new metric, RCPR, to compare the relative cost-performance ratio between our model and other baselines. RCPR consists of three items: Performance Retention (PR), Parameter Efficiency (PE), and Computational Efficiency (CE).

$$\text{RCPR} = \text{PR}^\alpha \times \text{PE}^\beta \times \text{CE}^\gamma \quad (3.28)$$

where α , β , and γ represents the importance weights. Moreover, each component is defined as:

$$\begin{aligned} \text{PR} &= \frac{\text{mIoU}_{\text{VGMEA}}}{\text{mIoU}_{\text{baseline}}} \\ \text{PE} &= \frac{\#\text{Params}_{\text{baseline}}}{\#\text{Params}_{\text{VGMEA}}} \\ \text{CE} &= \frac{\text{GFLOP}_{\text{baseline}}}{\text{GFLOP}_{\text{VGMEA}}} \end{aligned}$$

Therefore, RCPR favors light-weight models with high performance and low computational burden, and $\text{RCPR} > 1$ defines a relatively better model than the baseline. In our experiment, we set $\alpha = 2$, $\beta = 1$, and $\gamma = 0.5$ to focus on performance and parameter efficiency rather than computational efficiency.

Training protocol: All networks are implemented with PyTorch. Specifically, CNN-based methods are reproduced with MMSegmentation [121], while ViT-based networks are reproduced with Detectron2 [702]. We employ standard data augmentation techniques, including random scaling in $[0.5, 2.0]$, random horizontal flipping, and random cropping. Training follows the standard protocols for each dataset with a batch size of 4, using the Adam-W optimizer with an initial learning rate of 0.001 and polynomial learning rate decay. For fair comparison, all models are trained under identical training protocols until convergence on 2 NVIDIA TITAN RTX GPUs.

3.5.2 Performance Analysis

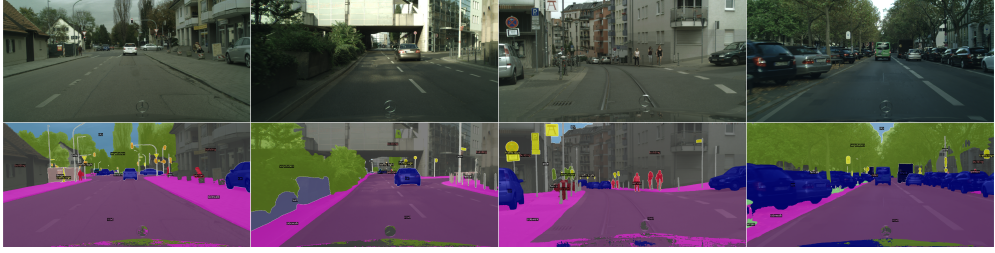
We evaluate our VGMEA framework against established attention mechanisms and typical semantic segmentation networks across four semantic segmentation benchmarks with increasing numbers of semantic classes. Our model is tested in three configurations: (1) VGMEA: without geometric constraints; (2) VGMEA $\dagger$: with amortized variational inference for geometric constraints; and (3) VGMEA $\ddagger$: with full iterative geometric constraints. We compare with two categories frameworks, CNN-based methods: Non-local[678], EncNet[806], EMANet[379], OCRNet[784], PSPNet [837], PSANet [838] and kMaxDeeplab [773]; and ViT-based approaches: MaskFormer[106], Mask2Former[105], SETR[847], HLG[813], Segmenter[595], SegFormer[712] and OneFormer[293].

Quantitative Results

Table 3.2 reports the mIoU and Dice scores on all datasets, along with #Params (parameter counts) and GFLOPS (Giga Floating-Point Operations Per Second). Our VGMEA models achieve competitive performance with significantly fewer parameters than other methods, yet consume more GLOPs.

Cityscapes: The emphasis on precise boundaries highlights the value of our TV regularization. CNN-based VGMEA $\ddagger$ achieves 81.8% mIoU, outperforming all other CNN models using the fewest parameters. Moreover, ViT-based VGMEA $\ddagger$ gains 82.7% mIoU, approaching to Mask2Former (84.1%) and OneFormer (84.6%) with

3.5. Experiment



(a) Cityscapes successful cases



(b) Cityscapes failure cases

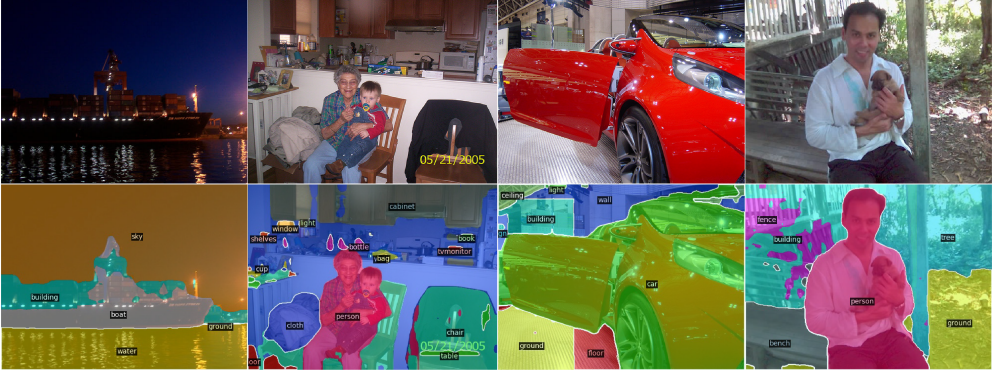
Figure 3.4: Visualization of semantic segmentation for the Cityscapes dataset. (a) Our model can classify pixels into the right semantic classes. (b) There are correlated semantic classes defined in the Cityscapes, resulting in the false responses between ambiguous classes, such as sidewalk vs. road vs. terrain.

significant parameter efficiency. Specifically, we exchange a 1.4% performance drop for a 27% parameter save with respect to Mask2Former, while a 1.9% drop yields a 48% save with respect to OneFormer. In addition, the 1.6% DICE improvement from w/o TV to Full TV demonstrates the critical importance of geometric constraints for improving boundary accuracy in urban scene understanding. In Fig.3.4, there are semantic segmentation results of VGMEA[‡] on the Cityscapes. From the failure cases, we observe that some ambiguously defined classes in Cityscapes are the main reason for the false classification since they violate our assumption on class independence, such as bicycle/motorcycle vs. person vs. rider and sidewalk vs. road.

PASCAL Context: Our CNN-based full TV model achieves 58.4% mIoU and 59.1% Dice, which achieves the best performance over the other CNN networks. Besides, our full TV model has demonstrated competitive results in a ViT-based setting. Specifically, it achieves more than 96% of Mask2Former and OneFormer’s performance with around 70% of the parameters. The amortized TV variant provides an acceptable trade-off between accuracy and computational efficiency, at the cost of more parameters in the mini-network for predicting the initial Φ surface. In Figure. 3.5, we expect that more



(a) PASCAL Context successful cases



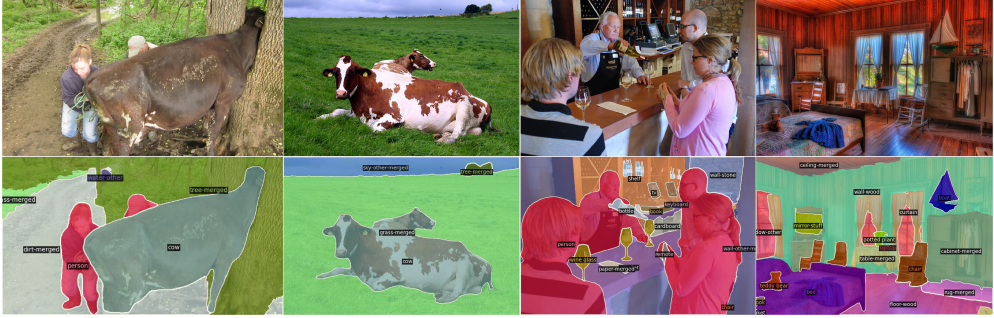
(b) PASCAL Context failure cases

Figure 3.5: Visualization of semantic segmentation from the PASCAL Context dataset. (a) Our model can classify pixels into the right semantic classes. (b) There are more ambiguous semantic classes defined in the PASCAL Context, such as building vs. wall, ground vs. floor.

correlated classes have competed with each other in our feature representation space and caused more perplexed label assignment.

MSCOCO: With 81 categories, our method achieves the top performance 62.4% mIoU among CNN-based models and maintains competitive performance against the ViT-based networks. Despite the efficiency, VGMEA with Full TV achieves 96% mIoU of Mask2Former and OneFormer, and outperforms the others with fewer parameters, demonstrating remarkable scalability to complex, multi-category scenes. In addition, some classes are difficult to distinguish within limited context information, such as a backpack, suitcase, or handbag in outdoor scenarios.

3.5. Experiment



(a) MSCOCO successful cases



(b) MSCOCO failure cases

Figure 3.6: Visualization of semantic segmentation for the MSCOCO dataset. (a) Our model can classify pixels into the right semantic classes. (b) Some classes are difficult to distinguish from limited context information, such as backpack vs. suitcase vs. handbag.

ADE20K: This dataset has diverse scenes with 150 categories, representing the most challenging scenario for model capacity. VGMEA with Full TV attains 46.5% mIoU in CNN-based setting and outperforms the other CNN networks. Moreover, the ViT-based VGMEA with Full TV (52.1%) is slightly lower than Mask2Former (53.9%) and OneFormer (54.7%), yet with around 30% fewer parameters, confirming its effectiveness across varying levels of complexity. In Figure. 3.7, our model learns the degraded feature representation space due to complicated correlation among various ambiguous classes.

Compatibility Analysis

Our VGMEA module demonstrates remarkable compatibility across diverse network architectures, serving as a lightweight drop-in replacement for conventional cross-attention



Figure 3.7: Visualization of semantic segmentation for ADE20K dataset. (a) Our model can classify pixels into the right semantic classes. (b) Some classes are confused by limited context information, such as house vs. tree, table vs. stand.

mechanisms without requiring architectural modifications. In CNN-based baselines, VGMEA seamlessly replaces their original attention modules with our probabilistic attention, reducing parameter count and improving performance. Meanwhile, VGMEA replaces the standard cross-attention layers in the transformer decoder for ViT-based models. As shown on the right side of Figure 3.8, VGMEA has achieved relatively better comprehensive performance than all the other benchmark models, since RCPRs are all above 1. Specifically, on the left side of the Figure 3.8, all VGMEA variants have the superiority of parameter efficiency over all the models (all PE metric values are above 1) while requiring more computational loads on average due to the bottleneck of geometrical regularization terms ($CE < 1$ in most cases). In addition, all VGMEA variants can achieve competitive performance compared to other benchmarks.

3.5. Experiment

CNN-based	Backbone	#Params	Cityscapes 768 x 1024 (19 classes)			PASCAL Context 512 x 512 (59 classes)			MSCOCO 1024 x 1024 (80 classes)			ADE20K 512 x 512 (150 classes)		
			mIoU	Dice	GFLOPs	mIoU	Dice	GFLOPs	mIoU	Dice	GFLOPs	mIoU	Dice	GFLOPs
NonLocal[678]	ResNet-101	67	78.7	79.2	911	45.4	46.2	286	52.4	53.2	1249	42.9	43.6	287
EncNet[806]		53	78.6	79.4	656	51.7	52.8	219	58.1	58.3	876	42.1	43.2	219
EMANet[379]		59	79.1	79.6	742	53.1	54.8	247	59.4	59.8	990	43.3	44.1	247
OCRNet[784]		56	80.1	80.4	693	56.2	57.3	231	60.4	61.1	926	43.9	44.5	231
kMaxDeepLab[773]		73	79.7	80.2	1004	<u>57.8</u>	<u>58.4</u>	320	60.7	61.0	1468	45.3	46.2	324
PSANet[838]		76	78.4	78.6	843	46.2	46.4	279	58.4	58.6	1107	43.8	44.3	274
PSPNet[837]		66	79.4	79.8	769	47.8	48.4	256	58.7	59.1	1021	43.6	43.9	257
VGMEA		42	80.6	80.9	648	56.4	57.1	250	60.8	61.4	982	45.8	46.9	254
VGMEA†		<u>44</u>	<u>81.2</u>	<u>81.8</u>	728	57.2	58.1	319	<u>61.3</u>	<u>61.7</u>	1128	<u>46.1</u>	<u>47.4</u>	324
VGMEA‡		42	81.8	82.5	995	58.4	59.1	358	62.4	62.5	1367	46.5	48.7	367
Transformer-based	Backbone	#Params	Cityscapes 768 x 1024 (19 classes)			PASCAL Context 512 x 512 (59 classes)			MSCOCO 1024 x 1024 (80 classes)			ADE20K 512 x 512 (150 classes)		
			mIoU	Dice	GFLOPs	mIoU	Dice	GFLOPs	mIoU	Dice	GFLOPs	mIoU	Dice	GFLOPs
MaskFormer[106]	ViT-Base	112	81.4	82.3	1027	62.3	63.1	584	<u>64.8</u>	65.1	1378	49.4	50.1	571
Mask2Former[105]		128	83.5	<u>84.1</u>	1385	<u>65.7</u>	<u>66.4</u>	792	67.4	68.6	1712	<u>53.9</u>	<u>54.3</u>	804
SETR[847]		160	79.2	79.8	1215	56.4	57.2	687	64.3	64.9	1559	47.7	48.1	688
HLG[813]		100	79.5	80.6	1183	57.5	58.1	618	64.1	64.3	1498	46.2	46.9	621
OneFormer[293]		180	<u>83.1</u>	84.6	1370	66.8	67.1	684	67.4	<u>68.4</u>	1690	54.7	55.2	678
Segmenter[595]		102	80.6	81.4	1205	59.7	60.4	563	62.3	63.1	1625	48.1	48.6	548
SegFormer[712]		129	82.3	83.6	1148	62.1	63.3	572	62.7	63.4	1632	49.6	50.4	568
VGMEA		91	81.3	81.6	1150	61.2	62.7	678	63.1	63.3	1477	50.8	51.2	690
VGMEA†		100	82.2	82.4	1320	62.8	63.2	775	63.8	64.1	1674	51.7	51.9	782
VGMEA‡		<u>93</u>	82.7	82.9	1520	63.9	64.5	850	64.7	64.9	1845	52.1	52.3	832

Table 3.2: Performance comparison on four datasets. Parameter counts are in millions (M), and GFLOPs denotes gigaflops. The best results are in bold, and the second-best results are underlined. † means using amortized TV (with 3 iterations), and ‡ represents the full TV (with 15 iterations).

In detail, we substitute the encoding layer in EncNet with our light-weighted module, reducing the parameter count from 53M to 42M while maintaining consistent gains in mIoU. Similarly, in EMANet, VGMEA replaces the expectation-maximization attention, achieving comparable performance (105% PR) with higher parameter efficiency (140% PE). For ViT-based models like Mask2Former and OneFormer, VGMEA replaces the standard cross-attention layers in the transformer decoder. This replacement is particularly effective because our method is inherently compatible with query-based mechanisms. VGMEA maintains 98% PR of the original mIoU while increasing PE to 128% and 180%, respectively. In conclusion, the universal compatibility stems from VGMEA’s adherence to standard cross-attention interfaces (query-key-value inputs and residual outputs), making it suitable for integration into any modern segmentation pipeline. Besides, our geometrical regularization can generally increase the performance, but at the cost of a high computational burden.

EM iters	mIoU (%)	Time (ms)
1	51.5	0.8
2	51.6	1.1
3	51.6	1.4
4	51.8	1.6
5	52.1	1.8

Table 3.3: Ablation study on the effects of EM iterations against the inference performance and speed. Experiments are conducted on ViT-Base with the ADE20K VOC dataset.

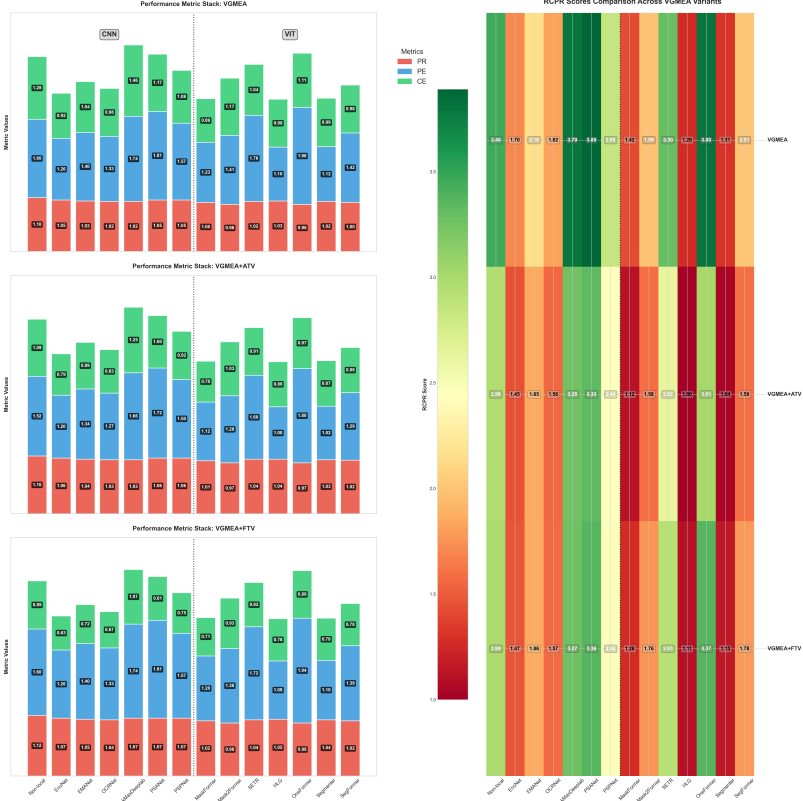


Figure 3.8: Cross-architecture compatibility analysis for three VGMEA variants: VGMEA uses no TV, VGMEA+ATV uses amortized TV, and VGMEA+FTV uses the full TV. PE, PR, CE and RCPR represent **P**arameter **E**fficiency (PE), **P**erformance **R**etention (PR), **C**omputational **E**fficiency (CE) and **R**elative **C**ost-**P**erformance **R**atio (CPR), respectively. For all metrics, values above 1 are better.

Convergence Analysis

On the one hand, we empirically analyze EM convergence in terms of training stability, given the theoretical guarantees of the EM algorithm. On the other hand, we estimate the inference efficiency in terms of EM iterations.

Training Stability: Due to the theoretical guarantees of the EM algorithm, the

3.5. Experiment

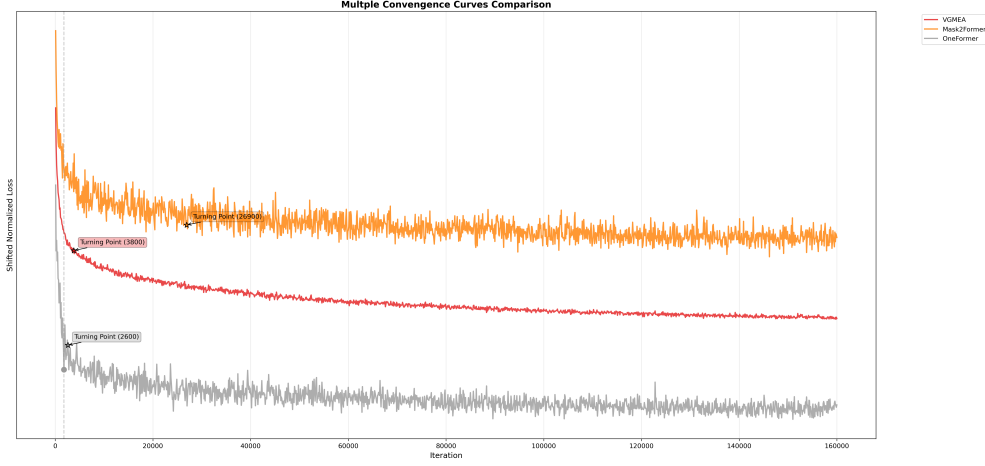


Figure 3.9: Ablation study on the convergence analysis of VGMEA against stacked cross-attention in Mask2Former and OneFormer on the ADE20K dataset.

VGMEA typically converges within 3-5 iterations via the evidence lower bound (ELBO) in a mini-batch. Compared to a deep stack of traditional attention blocks (i.e., 6-12 layers in OneFormer), VGMEA demonstrates smooth convergence curves compared to the oscillatory behavior sometimes observed in deep stacks (Figure 3.9).

Inference Efficiency: Due to the property of the moment-based algorithm, VGMEA can vary the EM iterations in the inference phase without changing the computational graph. That means fewer EM iterations are required for inference without significantly reducing performance, unlike fixed deep stack architectures, where inference typically passes through as many layers as in the training stage. This flexibility is particularly valuable for real-time applications. As shown in the Table. 3.3, we can run the EM algorithm once to achieve fast inference speed while maintaining good performance.

Interpretability Analysis

From a Gaussian mixture model perspective, VGMEA offers straightforward interpretability in the feature space during both the learning and inference stages.

During training, VGMEA learns to form the feature space, where each Gaussian component $\mathcal{N}(\mu_k, \mathbf{C}_k)$ evolves to represent a semantic concept. The centroids μ_k converge to prototypical feature vectors for each category, while the precision matrices $\mathbf{C}_k^{-1}$ capture intra-class variability. The orthogonal constraint ensures that different

concepts occupy nearly independent subspaces, facilitating clear semantic separation. Meanwhile, during inference, the responses of the feature space $r_{kn} = q(z_n = k)$ provide probabilistic assignments of each pixel to semantic concepts. The fact offers several interpretability advantages: (1) uncertainty quantification measured by classification confidence given by the posterior distribution; (2) concept attention visualized by r_{kn} activated in different regions; (3) representation disentanglement that analyzes the personalized input-dependent details separated from the common class-wise features; (4) ontology analysis that can detect the ambiguous semantic conception design in datasets.

3.5.3 Ablation Study

To thoroughly understand the contributions of each component in our VGMEA framework, we conduct comprehensive ablation studies analyzing the individual and combined effects of our key innovations.

Amortized Geometric Constraints

We first investigate the impact of our amortized total variation (TV) regularization scheme, a crucial innovation that balances geometric consistency with computational efficiency.

Experimental Setup: We compare five configurations: (a) No geometric constraints, (b) Full iterative TV regularization (15 iterations), and (c)-(e) Our amortized TV approach ($t \leq 5$ iterations + prediction network). Evaluations are conducted on all four datasets using the DICE metric to highlight the precise delineation of boundaries. From Table 3.4, we observe that the general capability of preserving effectiveness. On average, the amortized TV variants retain more than 90% of the full TV’s boundary delineation while reducing the inference time. In detail, the more iterations VGEMA performs, the higher its DICE score, but at the cost of more inference time. The benefits are consistent across datasets, with similar relative improvements observed across all four. Compared to the VGMEA with no TV terms (decrease $\geq -4.2\%$), the amortized TV with 5 iterations has retained the highest DICE score ($\geq 98.2\%$) but has spent more computation time on level-set evolution ($\geq +1.1$ ms per image). Although they do not completely avoid the TV bottleneck, amortized variants can be a good compromise between performance and inference speed.

3.5. Experiment

Configuration	Cityscapes val set		
	DICE (%)	Time/Image (ms)	Performance Gap
Full TV (15 iter)	82.9	2.5	0 (baseline)
No Constraints	78.3	0.5	-4.6% / -2.0ms
Amortized TV (1 iter)	79.1	0.9	-3.8% / -1.6ms
Amortized TV (3 iter)	80.2	1.5	-2.7% / -1.0ms
Amortized TV (5 iter)	81.1	1.8	-1.8% / -0.7ms

Configuration	PASCAL Context val set		
	DICE (%)	Inference Time (ms)	Performance Gap
Full TV (15 iter)	63.6	2.3	0 (baseline)
No Constraints	59.4	0.6	-4.2% / -1.7ms
Amortized TV (1 iter)	61.2	0.8	-2.4% / -1.5ms
Amortized TV (3 iter)	62.8	1.4	-0.8% / -0.9ms
Amortized TV (5 iter)	62.8	1.7	-0.8% / -0.6ms

Configuration	MSCOCO val set		
	DICE (%)	Time/Image (ms)	Performance Gap
Full TV (15 iter)	64.9	3.8	0 (baseline)
No Constraints	60.7	0.9	-4.2% / -2.9ms
Amortized TV (1 iter)	60.9	1.6	-4.0% / -2.2ms
Amortized TV (3 iter)	62.7	2.3	-2.2% / -1.5ms
Amortized TV (5 iter)	63.3	2.9	-1.6% / -0.9ms

Configuration	ADE20K val set		
	DICE (%)	Time/Image (ms)	Performance Gap
Full TV (15 iter)	52.3	5.4	0 (baseline)
No Constraints	44.9	1.2	-7.4% / -4.2ms
Amortized TV (1 iter)	48.2	3.1	-4.1% / -2.3ms
Amortized TV (3 iter)	49.5	3.9	-2.8% / -1.5ms
Amortized TV (5 iter)	50.8	4.8	-1.5% / -0.6ms

Table 3.4: Ablation study on geometric constraints on all datasets.

Orthogonal Constraint

The orthogonal constraint on concept centroids serves multiple purposes: improving interpretability, enabling computational simplifications, and enhancing semantic separation.

Experimental Setup: We evaluate four conditions: (a) No constraints, (b) Orthogonality only, (c) Diagonal covariance only, and (d) Full orthogonal + diagonal constraints. Measurements include parameter efficiency, semantic separation quality, and computational benefits.

Inter-class Separation quantifies the average dissimilarity between different semantic

concept centroids in the learned feature space. It measures how well the model distinguishes between different categories. Let $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ represent the centroid vectors of concepts i and j respectively, where $i, j \in 1, 2, \dots, K$ and $i \neq j$. The Inter-class Separation $\mathcal{S}_{\text{inter}}$ is defined as:

$$\mathcal{S}_{\text{inter}} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^K \left(1 - \text{cosine}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) \right) \quad (3.29)$$

Intra-class Coherence measures the average compactness of features within each semantic concept. It assesses how tightly clustered the features are around their respective concept centroids. For each concept k , let $\mathcal{X}_k = \mathbf{x}_n | z_n = k$ be the set of feature vectors assigned to concept k . The Intra-class Coherence $\mathbf{C}_{\text{intra}}$ is defined as:

$$\mathbf{C}_{\text{intra}} = \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{|\mathcal{X}_k|} \sum_{\mathbf{x}_n \in \mathcal{X}_k} \text{cosine}(\mathbf{x}_n, \boldsymbol{\mu}_k) \right) \quad (3.30)$$

Where the cosine similarity is calculated as:

$$\text{cosine}(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) = \frac{\boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i\| \cdot \|\boldsymbol{\mu}_j\|} \quad (3.31)$$

Therefore, two metrics provide quantitative evidence for the effectiveness of our orthogonal constraints in creating a well-structured feature space that supports robust semantic segmentation. Specifically, high $\mathcal{S}_{\text{inter}}$ and $\mathbf{C}_{\text{intra}}$ values correlate with better decision boundaries, reduced classification ambiguity, improved generalization performance, and enhanced model interpretability.

We conclude from Table 3.5 that on average orthogonality + diagonal correlation enhances the interpretability (improves inter-class separation $0.39 \rightarrow 0.85$), reduces the parameter count (18.1% by eliminating off-diagonal covariance elements), and accelerates the inference time (by 46.1% due to simplified matrix operations), but may degrade the semantic segmentation precision. That reveals a dilemma: explainable modules are likely to sacrifice performance, especially when the concept designs are ambiguous or complex in the dataset (i.e., terrain vs. sidewalk). Therefore, more prior information needs to be integrated into the design of explainable modules, for instance, a tree-structured conception in which the parent nodes represent the super-categories.

3.5. Experiment

Configuration	Cityscapes val set				
	mIoU (%)	Params (M)	Inter-class Separation $\uparrow$	Intra-class Coherence $\uparrow$	Inference Time (ms)
None	83.2	48.4	0.48	0.44	4.2
Orthogonality	82.8	48.4	0.89	0.85	4.2
Diagonal Correlation	82.9	42.1	0.54	0.45	2.3
Full Constraints	83.1	42.2	0.88	0.87	2.5
Configuration	PASCAL Context val set				
	mIoU (%)	Params (M)	Inter-class Separation $\uparrow$	Intra-class Coherence $\uparrow$	Inference Time (ms)
None	64.1	48.7	0.45	0.46	4.5
Orthogonality Only	63.1	48.7	0.88	0.87	4.5
Diagonal Correlation	62.7	42.2	0.52	0.51	2.4
Full Constraints	63.9	42.2	0.87	0.88	2.2
Configuration	MSCOCO val set				
	mIoU (%)	Params (M)	Inter-class Separation $\uparrow$	Intra-class Coherence $\uparrow$	Inference Time (ms)
None	65.3	52.7	0.36	0.38	5.1
Orthogonality Only	62.3	52.6	0.84	0.82	5.0
Diagonal Correlation	62.8	42.4	0.48	0.44	2.9
Full Constraints	64.7	42.5	0.83	0.82	2.8
Configuration	ADE20K val set				
	mIoU (%)	Params (M)	Inter-class Separation $\uparrow$	Intra-class Coherence $\uparrow$	Inference Time (ms)
None	53.1	57.3	0.28	0.31	5.9
Orthogonality Only	51.8	57.4	0.82	0.83	5.8
Diagonal Correlation	51.7	42.6	0.31	0.32	3.2
Full Constraints	52.6	42.6	0.83	0.84	3.1

Table 3.5: Ablation study on orthogonal constraints on all datasets

Trade-off between efficiency and accuracy

We analyze the overall efficiency-accuracy characteristics of different VGMEA configurations to provide practical guidance for deployment scenarios. There are 6 configurations formed by combinations of 3 TV options (none, amortized, full) with 2 orthogonal constraint options (none, orthogonality).

From Table 3.6, we can obtain optimal configurations for different scenarios on average: (a) the best efficiency chooses No TV + orthogonality for the highest parameter efficiency (60.5% mIoU vs. 40M Param) and speed (2.6 ms per image); (b) best balance adopts amortized TV + orthogonality that offers the optimal balance, achieving 97.2% of best performance with 64.0% of inference speed; (c) best interpretability combines full TV and orthogonality that achieves the second best performance but the complete semantic concept separation; (d) best Accuracy uses full TV + no orthogonality reach peak performance for interpretability-neglect applications. In conclusion, we observe cross-dataset consistency: the configuration of amortized TV plus orthogonality consistently provides the best trade-off between efficiency and accuracy, regardless of dataset complexity or characteristics.

Summary of Ablation Findings

Configuration	Cityscapes val set				
	mIoU (%)	Params (M)	GFLOPs	Memory (MB)	Inference Speed (ms)
No TV, No Orthogonal	79.2	48	485	2843	3.4
No TV, Orthogonal	79.4	40	324	2346	2.6
Amort TV, No Orthogonal	80.6	52	824	3956	4.7
Amort TV, Orthogonal	80.4	48	648	2748	3.5
Full TV, No Orthogonal	81.6	46	1230	3078	5.6
Full TV, Orthogonal	81.3	42	995	2578	4.3
Configuration	PASCAL Context val set				
	mIoU (%)	Params (M)	GFLOPs	Memory (MB)	Inference Speed (ms)
No TV, No Orthogonal	60.1	48	264	2975	3.5
No TV, Orthogonal	59.3	40	217	2453	2.5
Amort TV, No Orthogonal	62.4	50	397	4173	4.8
Amort TV, Orthogonal	62.7	45	319	2875	3.4
Full TV, No Orthogonal	64.1	48	537	3248	5.7
Full TV, Orthogonal	63.5	42	358	2671	4.4
Configuration	MSCOCO val set				
	mIoU (%)	Params (M)	GFLOPs	Memory (MB)	Inference Speed (ms)
No TV, No Orthogonal	60.3	47	1246	3247	3.8
No TV, Orthogonal	59.4	40	1078	2978	2.9
Amort TV, No Orthogonal	61.9	52	1578	4524	5.7
Amort TV, Orthogonal	61.3	48	1429	3314	4.3
Full TV, No Orthogonal	63.8	50	1684	3745	6.8
Full TV, Orthogonal	62.4	42	1367	3126	5.5
Configuration	ADE20K val set				
	mIoU (%)	Params (M)	GFLOPs	Memory (MB)	Inference Speed (ms)
No TV, No Orthogonal	44.3	49	287	3879	4.2
No TV, Orthogonal	43.8	40	224	3598	3.3
Amort TV, No Orthogonal	45.9	54	425	4975	6.8
Amort TV, Orthogonal	45.4	50	324	4015	5.1
Full TV, No Orthogonal	47.3	52	512	4178	7.4
Full TV, Orthogonal	46.5	42	367	3785	5.8

Table 3.6: Trade-off between efficiency and accuracy analysis on all datasets.

Our ablation studies demonstrate that:

1. Amortized TV provides 97.2% of full TV’s benefits with 64.0% faster inference speed.
2. Orthogonal constraints improve semantic separation by 46.0% while reducing parameters by 18.1%.
3. The optimal configuration (Amortized TV + Orthogonality) achieves the best efficiency-accuracy balance across all datasets.
4. Our approach offers a flexible configuration suitable for diverse deployment scenarios from real-time to accuracy-critical applications.

These results validate the design choices in VGMEA and provide practical guidance for researchers and practitioners seeking to leverage our method in various applications.

3.5. Experiment

3.5.4 Limitations

Despite the compelling advantages of our VGMEA framework, several limitations merit careful consideration and suggest promising directions for future research.

Limited Feature Space Adaptability of EM Iteration

The Expectation-Maximization algorithm, while providing theoretical guarantees and stable convergence, operates within a fundamentally static feature space framework. Unlike deep cross-attention stacks that enable hierarchical feature transformation through multiple layers, our EM-based approach performs iterative refinement within a fixed feature representation space. This limitation manifests in two critical aspects.

First, the inability of feature space migration restricts the modules evolving from abstract semantic concepts to dataset-specific instantiations. While traditional attention mechanisms can progressively transform and specialize features through stacked layers (e.g., $\mathbf{Q}^{(l+1)} = f(\mathbf{Q}^{(l)}, \mathbf{K}, \mathbf{V})$), our EM iterations primarily optimize assignment probabilities and centroid positions within the original feature space:

$$\boldsymbol{\mu}_k^{(t+1)} = \boldsymbol{\mu}_k^{(t)} + \eta \frac{\sum_n r_{kn}^{(t)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t)})}{\sum_n r_{kn}^{(t)}}$$

Second, this static feature space assumption may limit generalization across diverse domains. As shown in Table 3.7, while VGMEA performs excellently on structured datasets such as Cityscapes, its performance gain diminishes on more diverse datasets such as ADE20K, where adaptive feature transformation becomes crucial.

In summary, the practical implications are (a) limited hierarchical abstraction that VGMEA operates at a single level of abstraction while cross-attention stacks can progressively abstract features from local patterns to global semantics; (b) reduced adaptability that the fixed feature space may struggle with domain shifts; (c) architectural rigidity that inability to learn the complex feature transformations hinders deep transformers in few-shot. We suggest that future work could explore hybrid approaches where EM iterations are interleaved with feature transformation layers, creating a dynamic feature space that evolves during optimization.

Semantic Over-Constraint through Orthogonality

Dataset	Semantic Classes	VGMEA	Stacked Cross-Attention	Performance Gap
Cityscapes	19	82.7	83.1	-0.4%
PASCAL	59	63.9	66.8	-2.9%
MSCOCO	80	64.7	67.4	-2.7%
ADE20K	150	52.1	54.7	-2.6%

Table 3.7: Performance comparison highlighting feature adaptability limitations via mIoU(%). Our model adopts full TV + orthogonality while the stacked cross-attention modules are from OneFormer.

The orthogonal constraint $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j \approx 0$ for $i \neq j$, while computationally beneficial, imposes an artificial independence assumption that contradicts real-world semantic relationships. This limitation is particularly evident in natural semantic hierarchies that often exhibit inherent relationships (e.g., "vehicle" $\rightarrow$ "car" $\rightarrow$ "wheel") that should be reflected in the feature space geometry. The orthogonal constraint forces these related concepts to occupy perpendicular subspaces, discarding valuable relational information. Additionally, the computational savings from diagonal covariance matrices come at the cost of ignoring informative feature correlations that could enhance segmentation accuracy. In addition, the constraint leads to inefficient utilization of the feature space. To maintain orthogonality among K concepts in D -dimensional space, we require $D \geq K$, potentially forcing expansion into higher dimensions than necessary if some natural dependence were permitted.

Our experiments show that while orthogonality improves separation metrics, the performance gains diminish as the number of concepts increases, especially when concepts are complex or ambiguous. That suggests increasing "semantic tension" as category relationships become more complex. We advise that future research can focus on hierarchical semantic structures to accommodate layer-wise orthogonal constraints. The architectures might preserve computational benefits while accommodating complex semantic relationships. Specifically, a bottom-top structure consists of base-level orthogonality that enforces orthogonality only between semantically distant categories, hierarchical dependence that allows dependent relationships along category hierarchies, and conditional independence that maintains discriminability among categories on different branches.

Computational Bottleneck of TV Regularization

The Total Variation regularization, while effective for boundary preservation, suffers from significant computational challenges that limit practical deployment. The absence

3.6. Conclusions

of a closed-form solution necessitates iterative optimization through level-set methods. Each iteration requires computation of the curvature:

$$\kappa = \operatorname{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right)$$

This process typically requires more than 15 iterations for convergence, accounting for around 50% of total computation time. Furthermore, the level-set evolution approach used for TV minimization is inherently sequential and difficult to parallelize efficiently on modern hardware. We suggest addressing this bottleneck with several promising approaches: (a) train lightweight networks as a quick TV solver to predict optimal TV regularization directly from features; (b) apply multi-grid Methods to accelerate convergence; (c) use learned priors to provide better predictive initialization, reducing convergence iterations.

3.6 Conclusions

In this chapter, we have presented VGMEA, a novel variational Gaussian mixture model embedded attention framework that fundamentally reinterprets cross-attention mechanisms through a probabilistic lens. Our approach establishes a mathematically grounded foundation for attention operations by formulating them as iterative concept distillation processes within a mean-shift framework. The theoretical innovation can unify diverse attention variants under a single probabilistic framework. Through extensive experiments and ablation analysis, we conclude that our VGMEA not only enhances theoretical understanding but also provides practical adaptation to deep learning approaches. It pioneers a viable pathway for integrating sophisticated probabilistic models into deep learning architectures, bridging the long-standing gap between Bayesian methods and modern neural networks through innovations in online Bayesian updates and amortized variational inference. Despite these advancements, we acknowledge several limitations of our current approach on feature space transformation, semantic over-constraint on orthogonality, and the bottleneck of the Total Variation solver. Therefore, we suggest future work to integrate hierarchical orthogonal constraints and develop more efficient solvers for geometrical regularization. In conclusion, we believe that VGMEA represents a significant step toward more interpretable, efficient, and theoretically sound attention mechanisms, and we hope it will inspire further research at the intersection of probabilistic modeling and deep learning for computer vision and

beyond.

3.6. Conclusions

Chapter 4

Weakly Supervised Semantic Segmentation via Contextual Cross-Attention

Based on the following publication:

Zhan Xiong and Fons J. Verbeek. Weakly Supervised Semantic Segmentation via Contextual Cross-Attention. In *9th International Conference on Frontiers of Signal Processing, ICFSP 2024, Paris, France, September 12-14, 2024*, pages 83–92. IEEE, 2024

Chapter Summary

Weakly supervised semantic segmentation (WSSS) using image-level labels faces two challenges: Class Activation Maps (CAMs) lack explicit background modeling and lose spatial information during pooling. However, post-processing for background inference often results in imprecise separation. Moreover, the absence of geometric regularization in standard loss functions hinders accurate boundary delineation. To address these limitations, we propose a novel end-to-end WSSS framework that integrates a conditional probabilistic model with a geometric prior. Our framework is initiated by explicitly modeling the background using Bernoulli distributions trained with contrastive self-supervised learning. We then introduce a variational total variation (TV) term to refine background regions and improve boundary fidelity. Foreground class probabilities are subsequently modeled with Categorical distributions conditioned on the refined background. This conditional foreground modeling is implemented as a novel Contextual Cross-Attention (CCA) module that captures long-range dependencies among foreground regions. Our hierarchical framework offers insights into WSSS representation learning and significantly improves segmentation accuracy on the PASCAL VOC dataset.

Keyword: Conditional Probabilistic Model Image-Level Classification Cross Attention Contrastive Learning Global-Local Prototype Transformation

4.1 Introduction

Semantic segmentation is essential, but its application is often constrained by the request for pixel-level annotations, particularly in fields like biology and medicine, where such annotations require domain expertise. Weakly Supervised Semantic Segmentation (WSSS), which uses weaker supervision such as image-level labels, offers a promising alternative. However, image-level WSSS faces inherent limitations, including **ambiguous background modeling** due to the lack of explicit indicators and **loss of spatial details** from pooling operations (e.g., max or global average pooling). These limitations complicate the discrimination between precise and imprecise partitions.

Existing WSSS approaches address background modeling through discriminative and generative paradigms. Discriminative approaches, which define the background as the complement of foreground classes, employ a one-against-all classification strategy to allocate pixels to the background by exclusion. Therefore, the background cannot be conceptualized as **simple** as other foreground classes, whose representation cannot be directly modeled in a finite feature space but only rationally defined as feature vectors outside the foreground space, depicted in Figure 4.1. Despite their effectiveness, existing models [734, 158, 100, 315, 556, 862] often rely on class activation maps (CAMs) [564] for foreground representation. However, these methods often resort to ad-hoc post-processing for background inference, leading to false background responses. In contrast, generative methods [356, 343] adopt a one-vs-one classification strategy and leverage adversarial learning to simplify segmentation into binary classifications for each foreground class. However, they may undermine spatial exclusivity among classes and slow inference in multi-class scenarios.

Despite emphasizing background modeling, image-level WSSS methods often overlook issues of spatial information loss. Current approaches fail to ensure that loss functions favor accurate partitions, especially when there are fewer foreground classes [802]. The crux comes from the pooling operations that map inaccurate partitions to feature vectors similar to the correct ones, making them indistinguishable to image-level loss functions. Consequently, there is a need for geometrical regularization to align partitions with smooth boundaries of foreground regions, depicted in Figure 4.2. However, existing methods mainly use post-processing techniques such as thresholding or dense conditional random fields (dense-CRF), which are dataset-specific and rely on parameter tuning.

To address these limitations within a discriminative framework, we propose a novel conditional probabilistic model coupled with a variational geometric prior, depicted in

4.1. Introduction

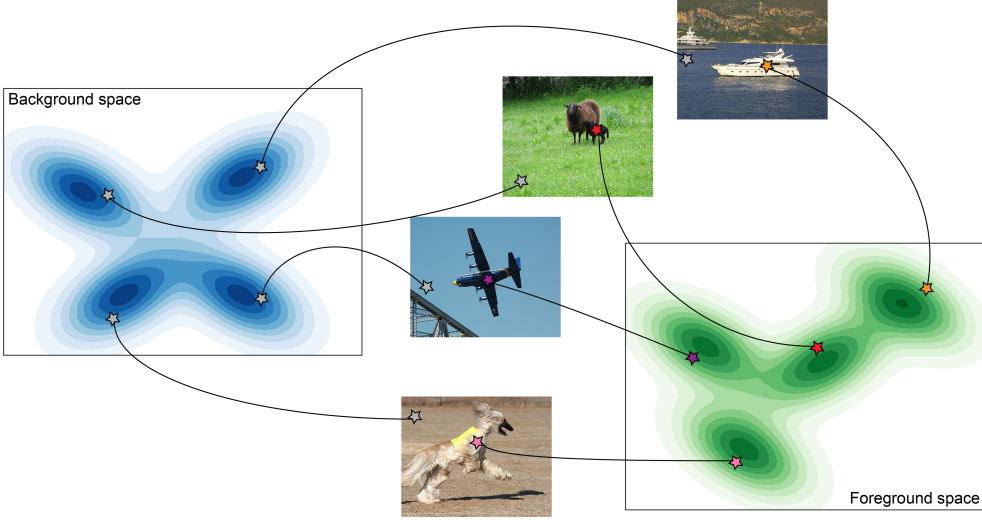


Figure 4.1: As the semantic information of foreground objects differs from that of backgrounds, the representative manifold is divided into two sub-spaces: foreground (green) and background (blue) spaces. All foreground objects share similar high-level representations and can be mapped to the same manifold. However, due to inter-class differences, they reside in different semantic clusters (color stars). All objects outside the foreground space belong to the background class. With different color/texture low-level features, they can also reside in different sub-groups (grey stars).

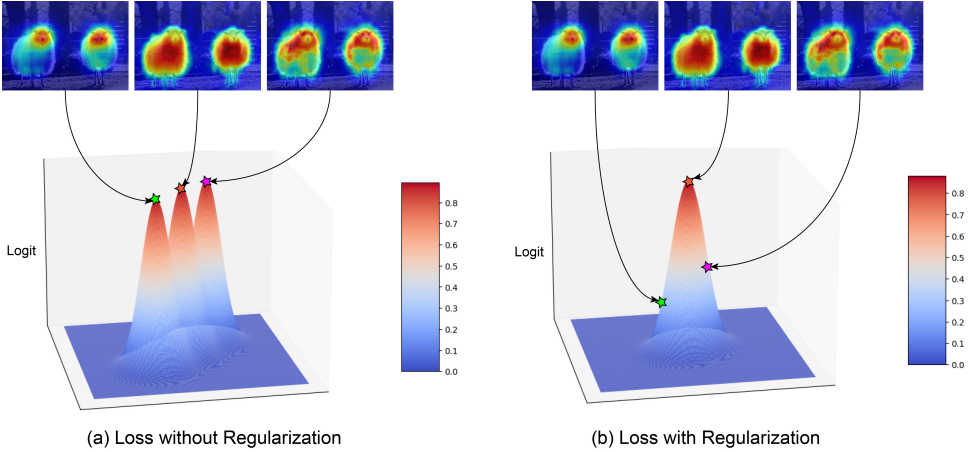


Figure 4.2: An illustration to show the role of geometrical regularization. Three partitions are qualified as bad (green star), good (orange star), and neutral (purple star) from left to right. (a) All partitions have similar logits in a standard image-level loss function. (b) With regularization, they are re-ranked with different logits according to their qualities.

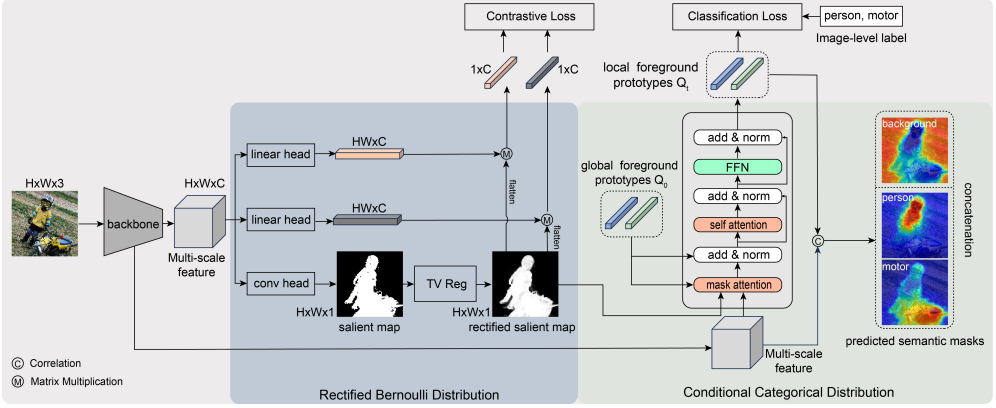


Figure 4.3: An Overview of the proposed framework. It consists of a rectified Bernoulli distribution and a conditional Categorical distribution. Global foreground prototype Q_0 is iteratively transformed into local foreground prototypes Q_t . Note that salient maps denote the complement of the background maps for better visualization. The final predicted semantic masks are obtained by concatenating a rectified background map and conditional foreground classification masks.

Figure 4.3. Our approach reformulates semantic segmentation as a hierarchical model. We first explicitly model the background classification as an independent Bernoulli distribution¹. The Bernoulli distribution is implemented as a sub-network using the sigmoid activation function. To work around the missing background indication information, we introduce foreground-background contrastive self-supervised learning [87, 158, 714] for training instead. Contrastive learning can automatically mine complementary information of background representations across images without requesting image tags. However, the initial background regions generated from self-attention might not be smooth and compact. Therefore, we introduce a variational geometric regularization term [56] to promote segmentation boundary fidelity. We argue that it is empirically rational to only regulate between background and foreground because **each foreground object should have a regular geometrical shape**. Therefore, the union of these regions is still regularly continuous. This regularization encourages coherent region delineations with smooth boundaries, allowing regions to expand / contract to better cover objects. The assumption is empirically proven to be a simple yet efficient way to address spatial information loss with minimal computational overhead. The Bernoulli distribution with the geometrical regularization is called the rectified Bernoulli distribution. For foreground classification, we employ a conditional Categorical distribution within the identified foreground regions, enforcing soft pixel

¹Strictly speaking, it is conditional independence when given previous layers.

4.2. Related Work

assignments among foreground classes. Our conditional Categorical distribution can be formulated as a stack of masked cross-attention modules [105]. Each module captures long-range dependencies among foreground partitions only. We name the conditional Categorical distribution as contextual cross-attention (CCA). The CCA can extract global foreground prototypes from the dataset and gradually transform them into local prototypes based on the input content. Because the local prototypes are adapted to the features of the corresponding input, the soft label assignment can be computed as probabilities by normalizing the pairwise similarities between pixels and local prototypes, and it benefits from reduced variation compared to global ones. Our framework offers a new perspective on the structure of representative embedding spaces under the WSSS paradigm.

In summary, our main contributions are as follows:

- An end-to-end WSSS framework offering a new understanding of representative embedding spaces. Our model uses contrastive learning for explicit background modeling and a novel contextual cross-attention for foreground classification.
- A variational geometric regularization for smooth region delineation, efficiently addressing spatial information loss and improving segmentation masks with minimal computational overhead.
- Our method is a generic approach that can easily extend with other regularization terms and demonstrates superior performance on the benchmark dataset.

4.2 Related Work

Image-level Discriminative WSSS. Discriminative approaches to image-level WSSS often leverage CAMs [564] to identify foreground and background regions. However, CAMs can result in incomplete partitions and imprecise boundaries. Research has focused on expanding CAMs into more robust semantic masks to address this issue. One approach involves global-to-local prototype transformations [62, 856, 734, 83, 862, 100] to mitigate non-discriminative regions. Others integrate complementary information, such as semantic affinities [3], boundary information [550], uncertainty [394], or exploit information bottlenecks [355]. Some efforts mitigate challenges through data manipulation, such as introducing hard out-of-distribution data [357] or strong augmentations [597]. While these methods improve foreground modeling, they often rely on heuristic post-processing for background estimation, potentially limiting precise semantic seg-

mentation.

Contrastive Learning. The core idea of contrastive learning is to group feature vectors based on their similarities within a feature space, enforcing intra-group similarity and inter-group difference. Based on object class labels, this approach creates positive pairs from samples within the same class and negative pairs from those across different classes. Hence, contrastive learning [87] pulls the samples from positive pairs closer together and pushes the samples from negative pairs farther apart. Recent WSSS research uses contrastive learning to extract additional information across images, including pixel-to-prototype contrast [158], knowledge distillation [626], and biased object removal [315]. These complementary cues can assist in accurate semantic classification.

Cross-Attention. Cross-attention computes the similarities between pixels and queries. Some studies [773] implement cross-attention as k-means clustering, while others [663] use a K-NN clustering perspective. We can interpret image-level WSSS as a form of clustering, with object queries serving as cluster centers and their learnable embeddings. In this chapter, we adopt a k-means clustering interpretation and reformulate cross-attention as a conditional probability within our framework to predict pixel-level similarities. Therefore, cross-attention can be a scalable module that transforms cluster centers into different representations.

Regularization. Regularization prevents model underfitting or overfitting by introducing additional information. It is crucial for image-level WSSS because a classifier maps different partitions onto similar vectors. From a Bayesian perspective, one regularization term is a prior distribution reflecting *a priori* knowledge. Most recently, adversarial perturbations [356] have been applied to bridge the train-test error gap. However, existing regularization terms cannot directly guide image-level loss functions in evaluating partitions at the pixel level. To address this, we introduce a total variation (TV) regularization term to encourage partition alignment with object boundaries.

4.3 Method

In this section, we reformulate image-level WSSS as a hierarchical probabilistic framework in an end-to-end manner and provide the solutions of each component approximated by mean-field theory. Given a feature map $X \in \mathbb{R}^{C \times H \times W}$, where C , H and W denote the channel number and spatial dimensions, our model predicts a pixel-level semantic label assignment $S \in \mathbb{R}^{(K+1) \times H \times W}$, where K is the number of **foreground**

4.3. Method

classes. We treat the background as an auxiliary class and split label vectors with two distinct parts: a scalar set $B = \{b_i | b_i \in \{0, 1\}\}$ for the background class, and a conditional 1-of- K vector set $F = \{(1 - b_i) \cdot \tilde{f}_i | f_{i,k} \in \{0, 1\}, \sum_k f_{i,k} = 1\}$ for foreground classes. Here, $i \in \{H \times W\}$ indexes spatial locations, and $f_{i,k}$ is the label of the k -th foreground class at position i . To enhance segmentation boundary fidelity, we introduce an auxiliary term $\tilde{B} \in \mathbb{R}^{1 \times H \times W}$ as the initial binary background map, and the geometrical prior distribution uses $\tilde{B}$ to regulate the background regions.

4.3.1 Probabilistic Framework

We offer a novel perspective on semantic label assignment. In contrast to the conventional manner, we define it as a structural process and formulate it as a conditional probabilistic model:

$$P(S, \tilde{B} | X) = P(B, F, \tilde{B} | X) = P(F | B, X) \cdot P(B, \tilde{B} | X) \quad (4.1)$$

The $P(B, \tilde{B} | X)$ is defined as a rectified Bernoulli distribution to generate binary background maps with compact and smooth boundaries. Moreover, $P(F | B, X)$ denotes a conditional 1-of- K Categorical distribution to assign semantic labels within identified foreground regions.

Rectified Bernoulli Distribution

The rectified Bernoulli distribution is also a conditional probability:

$$P(B, \tilde{B} | X) = P(B | \tilde{B}) \cdot P(\tilde{B} | X) \quad (4.2)$$

$P(\tilde{B} | X)$ represents Bernoulli distributions that generate initial pixel-level background probability maps, and $P(B | \tilde{B})$ represents a geometrical prior that refines the initial predictions into more compact and smooth regions.

Denote a binary classifier as $h(\cdot)$, which inputs the feature map X . Then, the probability of being background is defined as a sigmoid function:

$$p_i = \text{Sigmoid} \left(h(X_i) \right) \quad (4.3)$$

Where i indexes the spatial location. Therefore, we parameterize Bernoulli distributions

as:

$$P(b_i|X) = \begin{cases} p_i, & b_i = 1, \\ 1 - p_i, & b_i = 0 \end{cases} \quad (4.4)$$

Therefore, the initial background probability map $\tilde{B}$ is denoted as the expectation of our Bernoulli distribution:

$$\tilde{B} = P(b_i = 1|X) = \text{Sigmoid} \left(h(X_i) \right) \quad (4.5)$$

That is $h(X)$ activated by the sigmoid function. The binary classifier $h(\cdot)$ can be a simple block consisting of a convolution layer followed by a normalization layer.

However, the missing spatial information in image-level labels leads to the derived $\tilde{B}$ suffering from blurred details, holes, and inaccurate boundaries. We need the geometrical regularization to produce compact and uniform background maps B .

We introduce the total variation (TV) [56] to refine the initial background maps. It removes value variations inside and preserves boundaries, making segments more compact and complete. We defined the geometrical prior as the Gibbs Distribution:

$$P(B|\tilde{B}) = \frac{1}{Z_g} e^{-E(B, \tilde{B})} \quad (4.6)$$

where Z_g is the normalization term and $E(B, \tilde{B})$ is the total variance:

$$E(B, \tilde{B}) = \frac{\lambda}{2} \|B - \tilde{B}\|_2^2 + \|\nabla B\|_2 \quad (4.7)$$

where λ is a scale factor, ∇ is a gradient operator and $\|\cdot\|_2$ is L_2 -norm. According to the heat equation [194], it can be solved as an iterative time-dependent surface:

$$-\frac{\partial B}{\partial t} = \nabla \left(\frac{\nabla B}{\|\nabla B\|_2} \right) + \lambda(\tilde{B} - B) \quad (4.8)$$

Therefore, we can obtain rectified background maps B after t iterations:

$$B_{t+1} = B_t - \nabla t \cdot \left[\nabla \left(\frac{\nabla B_t}{\|\nabla B_t\|_2} \right) + \lambda(\tilde{B} - B_t) \right] \quad (4.9)$$

4.3. Method

Conditional Categorical Distribution

Conditioned on the refined background regions B , we use image-level labels to compute soft foreground assignment. Assume $Q \in \mathbb{R}^{K \times C}$ is an embedding space holding the semantic information for K foreground classes. Because we only infer label assignments within foreground regions, i.e., $b_i = 0$, we can formulate the conditional foreground classification as:

$$\begin{aligned} P(F|B, X) &= \frac{1}{Z_f} \prod_{i=1}^{H \times W} (1 - p_i) \cdot \prod_{k=1}^K \left(e^{X_i^T \cdot Q_k} \right)^{f_{i,k}} \\ &= \frac{1}{Z_f} \prod_{i=1}^{H \times W} \prod_{k=1}^K \left(e^{\log(1-p_i) + X_i^T \cdot Q_k} \right)^{f_{i,k}} \end{aligned} \quad (4.10)$$

where Z_f is the normalization term, T is a transpose operator. Since $\sum_k f_{i,k} = 1$, we can move $1 - p_i$ inside.

Notably, if we denote $\log(1 - p_i)$ as $\mathcal{M} \in (-\infty, 0)$, then the logarithm of the conditional probability is a generalized form of masked cross-attention [105]:

$$\log P(F|B, X) = \sum_{i=1}^{H \times W} \sum_{k=1}^K \text{Softmax}(\mathcal{M}_i + X_i^T \cdot Q_k) \cdot f_{i,k} \quad (4.11)$$

We approximate the foreground probability map F as the expectation of the conditional Categorical distribution and solve it as a process of soft k-means clustering. In order to adapt the iterative process into feedforward neural networks, we unfold the soft k-means clustering as follows:

$$Q_k^t = \text{Softmax}(\mathcal{M} + X^T \cdot Q_k^{t-1}) \cdot X + Q_k^{t-1} \quad (4.12)$$

Therefore, we linked masked cross-attention with soft k-means clustering to efficiently extract semantic information for foreground classes. Consequently, we define an embedding space where each Q_k^0 is a global prototype for the k -th foreground class over the entire dataset. Moreover, Q_k^t is a local prototype that is transformed t times based on the input data. To align with transformers, we extend Eq.(4.12) to a more generalized form, which allows soft k-mean clustering to work in a feedforward manner:

$$Q_t^k = \text{Softmax}(\mathcal{M} + \varphi(X)^T \cdot Q_{t-1}^k) \cdot \psi(X) + Q_{t-1}^k \quad (4.13)$$

where $\varphi(\cdot)$ and $\psi(\cdot)$ are linear projections respectively. In addition, we define our conditional foreground label assignment as the normalized similarities between pixels and local foreground prototypes:

$$F = \text{Softmax}(\mathcal{M} + \varphi(X)^T \cdot Q_t^k) \quad (4.14)$$

Therefore, we named the iterative solution to the conditional Categorical distribution a stack of **Contextual Cross-Attention** modules that compute conditional foreground label assignments via a global-to-local prototype transformation. Thanks to smaller variations in local prototypes, our method can improve the quality of F .

4.3.2 Implementation

In this section, we elaborate on the details to implement three key components in the CCA model: (1) how to exploit the unsupervised contrastive learning to infer the missing background information. (2) how to integrate geometrical terms to enhance the loss function in WSSS. (3) how to convert our probabilistic framework into the well-established cross-attention mechanisms.

Unsupervised Contrastive Learning

It is necessary to train the binary background classification module to predict initial background maps. However, image-level labels provide no background information. Therefore, we adopt unsupervised contrastive learning to mine extra information across images within a dataset. The core idea is that each representation should have intra-class similarity and inter-class difference. Denote $\vec{v}_b$ and $\vec{v}_f$ as the feature representation of foreground and background respectively:

$$\vec{v}_b = \frac{\tilde{B} \otimes X^T}{\|\tilde{B}\|_1}, \quad \vec{v}_f = \frac{(1 - \tilde{B}) \otimes X^T}{\|1 - \tilde{B}\|_1} \quad (4.15)$$

Here, $\tilde{B} \in \mathbb{R}^{1 \times HW}$ and $X \in \mathbb{R}^{C \times HW}$ are flattened. Operators T and $\otimes$ indicate the transpose and matrix multiplication. In addition $\|\cdot\|_1$ denote the L_1 norm. Therefore, $\vec{v}_b$ and $\vec{v}_f$ are mean vectors of each region.

In contrastive learning, we denote the pair of feature vectors from the same class (e.g., background-background or foreground-foreground combination) as the positive pair and vice versa. The core idea behind this is a fact of feature space. The feature representations across images from the same class should be similar to one another

4.3. Method

but significantly different from those of other classes. Therefore, we can train a binary classification by pulling close a positive pair and pushing apart the negative pair[87]:

$$s_{i,j}^{neg} = \text{cosine}(v_i^b, v_j^f), \quad s_{i,j}^{pos} = \text{cosine}(v_i^k, v_j^k) \quad (4.16)$$

where $\text{cosine}(\cdot)$ is the cosine similarity operator that outputs a real value between 0 and 1. Besides, $k \in \{b, f\}$ is either b or f . Finally, i and j are the indices of the image within a mini-batch. Consequently, we can explicitly model the background as a set of feature vectors that do not belong to any foreground class with contrastive learning loss:

$$\begin{aligned} L_{CST} = & -\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \log(1 - s_{i,j}^{neg}) \\ & - \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{[i \neq j]} \log(s_{i,j}^{pos}) \end{aligned} \quad (4.17)$$

where $\mathbf{1}_{[i \neq j]} \in \{0, 1\}$ is an indicator function and it equals 1 if $i \neq j$. In summary, L_{CST} considers the contrastive comparison both within image ($i = j$) and cross-image ($i \neq j$).

Geometrical Prior

We use TV regularization to refine the binary classification maps $\tilde{B}$ generated by contrastive learning, since background regions often exhibit holes or ragged boundaries. According to the implementation detail of level-set methods [270], we define an auxiliary surface $\Phi \in \mathbb{R}^{1 \times H \times W}$, and then the smooth surface $B = \frac{1}{1+e^{-\Phi}}$. In this chapter, we set $\nabla t = 1$ for simplicity, and implement $\nabla \left(\frac{\nabla B}{\|\nabla B\|} \right)$ as Gaussian curvature κ [365]. Therefore, the iterative update rule in Eq.(4.9) has become:

$$B_{t+1} = (1 + \lambda) \cdot B_t - \kappa_t - \lambda \cdot \tilde{B} \quad (4.18)$$

We can adjust the smoothness of the background map while maintaining data fidelity with the scale factor λ . We tried different values for λ and fixed it as $\lambda = 2$. After T iterations, we can obtain smooth background regions. In all experiments, we set $T = 3$. Notably, we chose not to include the geometrical prior in the neural network's objective because adding Eq.(4.7) as an area loss component leads to unstable training. Consequently, we derived Eq.(4.18) and used it iteratively. For correct backward

propagation, the TV regularization module is detached from other modules, and the rectified prediction is directly added to the original binary prediction for regularization.

Contextual Cross-Attention

To align with transformers, we integrate each iteration of contextual cross-attention with a self-attention and a Feed Forward Network (FFN) into a decoder layer [105]. To keep the iterative update rules in the k-means clustering algorithm, we set $T = 6$ decoder layers to capture global semantic information from multi-scale features and gradually transform them into image-specific local prototypes to capture non-discriminative foreground regions.

We use the cross-entropy function to compute classification loss between image label Y and image-specific local prototypes Q_T :

$$L_{CLS} = \text{CrossEntropy}(Y, W \cdot Q_T) \quad (4.19)$$

where W is a linear projection head for classification. Our predicted final semantic segmentation map S is defined as a concatenation of rectified binary background prediction B and conditional foreground label assignment F :

$$S = [B, (1 - B) \cdot F] \quad (4.20)$$

Noted that, $\sum_{k=0}^K S_{i,k} = 1$ satisfies the spatial exclusivity.

4.3.3 Loss Function

In this chapter, the loss function is defined as

$$L = L_{CLS} + \alpha \cdot L_{CST} \quad (4.21)$$

where α is a weight factor to adjust the importance of contrastive learning loss. Because the early background maps generated by contrastive learning are imprecise and noisy, this significantly destabilizes the training process of the entire model. Therefore, we first trained the contrastive learning head on the dataset, then used its weights as the pre-trained backbone, and finally jointly trained it with CCA modules. In all experiments, we set the $\alpha = 0.3$.

4.4 Experiments and Results

In this section, we conduct comprehensive experiments and ablation analyses of our CCA model and demonstrate its advantages in both interpretability and efficiency.

4.4.1 Experimental Setup

Datasets. We evaluate the proposed CCA on the PASCAL VOC 2012 [167]. The PASCAL VOC 2012 contains 20 foreground and 1 background categories. It consists of three splits: 1,464 images for training, 1,449 for validation, and 1,456 for testing. Following previous WSSS methods, we use the augmented 10,528 training images present in SBD [228] along with their image-level labels to train the network.

Evaluation Metrics. To evaluate the segmentation performance of all methods, we use the validation and test set of PASCAL VOC 2012. In addition, we use the PASCAL VOC 2012 training set to evaluate background map performance and conduct ablation studies. The mean Intersection-over-Union (mIoU) ratio is the evaluation metric in all experiments.

Training Details. In our experiments, we adopt ResNet50 [248] pre-trained on the ImageNet [138] as the backbone and use cascade CCA as in Mask2former [105]. We train all the reproduced models using the AdamW [445] optimizer with a polynomial learning scheduler. The augmentation strategy is the same as [83], including random flipping, random scaling, and cropping. All models take a batch size of 16 on 2 NVIDIA GeForce RTX 3090 GPUs.

Inference. To give a fair comparison, all models generate foreground cams and background maps, convert them into pseudo-labels, and adopt a dense-CRF method [338] as the post-process method for refinement. We further use these refined pseudo-labels to retrain a semantic segmentation network DeeplabV2 [78].

4.4.2 Results and Analysis

Improvements on background maps. We first evaluate background maps because they are crucial to locating the foreground regions for subsequent foreground prototype modeling. We compare the performance of generated background maps in correctly separating the background from other foreground classes. Table 4.1 compares CCA with other advanced methods on the PASCAL VOC 2012 training set. Among these

compared methods, WeakTr [862] achieves the best results with a mIoU of 58.1%. Our proposed explicit background modeling obtains a better performance of 58.8%. Furthermore, we report the performance with dense-CRF post-processing. Our performance with dense-CRF improves the mIoU to 65.7% and outperforms all other methods. We interpret the performance gain derived from two aspects: (1) contrastive learning can effectively mine supplementary information across different images to separate background-foreground regions; (2) TV regularization can adaptively enforce smooth boundaries of foreground regions at the pixel level. Therefore, the dense-CRF is less confusing when refining the binary background maps. Figure 4.4 visualizes the results of background maps on the PASCAL VOC 2012 train set. Our rectified background maps effectively capture all semantic regions across diverse scenes, including varying object scales, crowded scenes, and multiple categories. Notably, our model can correctly suppress false responses when there is no background and large-scale objects fully occupy the image. In contrast, other methods treat large-scale partial foreground objects as background, since background is assumed always to exist.

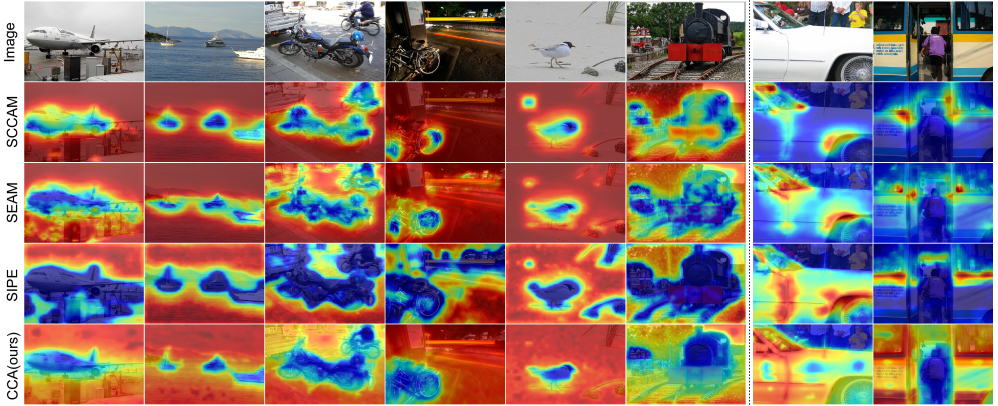


Figure 4.4: Visual comparison of CAMs. From top to bottom: original images, SCCAM, SEAM, SIPE, and the proposed CCA. All columns on the left side of the dotted line demonstrate the CAMs of the background. With geometric regularization, our CCA can effectively suppress false responses from background regions. The two images on the right side have no background, only large-scale foreground objects (i.e., a car and a bus). Other methods mistakenly classify them as background regions, whereas our CCA correctly classifies them as the car and the bus regions.

Improvements on initial pseudo-labels. To show the importance of background modeling, we focus on the performance of initial pseudo-labels generated from background and foreground activation maps. Table 4.2 shows that our method outperforms all other approaches. We can interpret this as the explicit background modeling

4.4. Experiments and Results

Method	Model	Regularization	Extra labels	Location Maps	+dense CRF
SEAM _{CVPR20} [683]	Thold	-	-	55.4	56.8
SCCAM _{CVPR20} [*] [62]	Thold	-	-	50.9	55.3
EDAM _{CVPR21} [698]	Thold	-	-	52.8	58.2
AdvCAM _{CVPR21} [356]	Thold	AC	-	55.6	62.1
ECS _{CVPR21} [600]	Thold	-	-	56.6	58.6
OCCSE _{ICCV21} [*] [342]	-	-	-	56.0	62.8
AEFT _{ECCV22} [768]	-	-	-	56.9	60.3
CLIMS _{CVPR22} [713]	-	AR	$\mathcal{S}$	57.1	64.2
SIPE _{CVPR22} [*] [83]	Adapt	-	-	56.8	63.7
MCTformer _{CVPR22} [*] [734]	Thold	-	-	57.2	63.3
WeakTr _{arXiv23} [*] [862]	Thold	-	-	58.1	64.5
CCA(Ours)	Const	TV	-	58.8	65.7

Table 4.1: Evaluation of background maps on the VOC 2012 train set. The symbol * represents all reproduced methods. *Thold* - threshold. *Const* - contrastive learning. *Adapt* - adaption of the foreground. *AC* - adversarial climbing. *TV* - total variation. *AR* - area regularization. $\mathcal{S}$ - extra class-related background texts. '-' means not use. The best results are shown in bold.

Method	Backbone	Supervision	Val dataset	+dense CRF
SEAM _{CVPR20} [683]	ResNet38	$\mathcal{I}$	49.8	53.3
SCCAM _{CVPR20} [*] [62]	ResNet101	$\mathcal{I}$	50.9	54.2
EDAM _{CVPR21} [698]	ResNet101	$\mathcal{I}$	51.5	54.4
AdvCAM _{CVPR21} [356]	ResNet101	$\mathcal{I}$	52.3	54.9
ECS _{CVPR21} [600]	ResNet38	$\mathcal{I}$	53.9	57.1
OCCSE _{ICCV21} [*] [342]	ResNet38	$\mathcal{I}$	52.9	56.3
AEFT _{ECCV22} [768]	ResNet38	$\mathcal{I}$	54.9	57.3
CLIMS _{CVPR22} [713]	ResNet101	$\mathcal{I} + \mathcal{S}$	56.8	62.0
SIPE _{CVPR22} [*] [83]	ResNet38	$\mathcal{I}$	56.3	61.1
MCTformer _{CVPR22} [*] [734]	ResNet38	$\mathcal{I}$	55.3	60.7
WeakTr _{arXiv23} [*] [862]	DeiT-S	$\mathcal{I}$	57.8	62.2
CCA(Ours)	ResNet50	$\mathcal{I}$	59.2	63.4

Table 4.2: Comparison of initial pseudo-labels on the VOC 2012 val set. The symbol * represents all reproduced methods. The best results are shown in bold. $\mathcal{I}$ - image-level label $\mathcal{S}$ - extra class-related background texts.

effectively removes irrelevant regions from the foreground and enforces a global-to-local prototype transformation within well-separated foreground regions. Therefore, the embedding feature space can capture more abstract information for global representations of foreground classes while generating more locally adaptive details for non-discriminative regions in each image.

Improvements on fully supervised segmentation. To further evaluate the performance of our methods, we train fully supervised models using refined pseudo-labels and compare the segmentation results on PASCAL VOC 2012 val and test sets. Following the common practice [356], IRN refined the pseudo-labels [3] and used them to train DeepLabV2 [78]. Table 4.3 presents the performance of semantic segmentation, where our CCA outperforms other methods with 74.3% on the validation set and 74.1% on the test set. Based on our CCA, DeepLabV2 shows greater robustness across various challenging scenarios, including different object scales, multiple objects, and multiple categories. In addition, Figure 4.8 shows the semantic segmentation predictions from our model trained with pseudo-labels.

Method	Backbone	Supervision	Val dataset	Test dataset
SEAM _{CVPR20} [683]	ResNet38	$\mathcal{I}$	64.5	65.7
SCCAM _{CVPR20} [*] [62]	ResNet101	$\mathcal{I}$	66.1	65.9
EDAM _{CVPR21} [698]	ResNet101	$\mathcal{I}$	70.9	70.6
AdvCAM _{CVPR21} [356]	ResNet101	$\mathcal{I}$	68.1	68.0
ECS _{CVPR21} [600]	ResNet38	$\mathcal{I}$	66.6	67.6
OCCSE _{ICCV21} [*] [342]	ResNet38	$\mathcal{I}$	68.4	68.2
AEFT _{ECCV22} [768]	ResNet38	$\mathcal{I}$	70.9	71.7
CLIMS _{CVPR22} [713]	ResNet101	$\mathcal{I} + \mathcal{S}$	70.4	70.0
SIPE _{CVPR22} [*] [83]	ResNet38	$\mathcal{I}$	68.5	69.0
MCTformer _{CVPR22} [*] [734]	ResNet38	$\mathcal{I}$	70.3	70.8
WeakTr _{arXiv23} [*] [862]	DeiT-S	$\mathcal{I}$	72.2	72.3
CCA(Ours)	ResNet50	$\mathcal{I}$	74.3	74.1

Table 4.3: Comparison of WSSS on the VOC 2012 val and test sets. The symbol * represents all reproduced methods. $\mathcal{I}$ - image-level label. $\mathcal{S}$ - extra class-related background texts. The best results are shown in bold.

4.4.3 Ablation Study

Effects of explicit background modeling. We conduct an ablation study to verify the effect of explicit background modeling. We select 8 options and demonstrate their mIoUs in Table 4.4. Each option concerns a single background modeling method when using the TV term. We observe that the fixed-threshold method achieves the worst mIoU because no single threshold correctly separates background regions across all images. Likewise, estimating background directly from CAMs [83] is also difficult due to heavy reliance on the quality of foreground classifications. In addition, combining background information from a series of binary classifications (used in adversarial

4.4. Experiments and Results

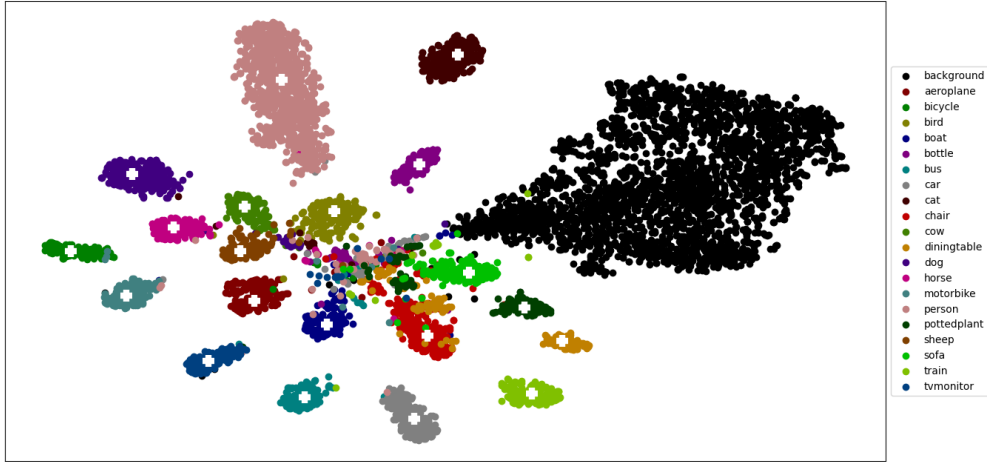


Figure 4.5: Visualization of 21 inter-class groups using t-SNE. Our CCA can capture abstract information to conceptualize each foreground class as a cluster in the feature space.

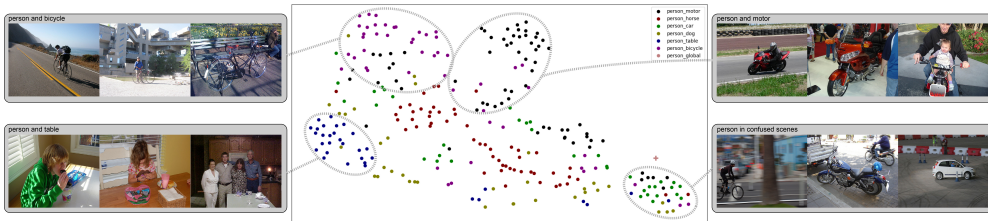


Figure 4.6: Visualization of intra-class deviation in the Person class using t-SNE. Grey dotted circles denote sub-groups. It illustrates the importance of global-to-local prototype transformation in the feature space. Our CCA can transform the abstract global prototype (red plus symbol) into image-specific local prototypes (broadly scattering dots) to capture non-discriminative foreground regions.

learning methods [342]) can achieve performance comparable to that of the contrastive learning method. However, they still face the challenge of spatial exclusivity across different classes, a fundamental requirement for semantic segmentation. Therefore, the presence of multiple responses in binary classifications ultimately degrades the mIoU scores. Our explicit background modeling, with or without the TV term, achieves top-2 scores. We can interpret this superiority as correctly recognizing background regions under different scenarios based on auxiliary cues mined from different images.

Effects of foreground prototype modeling. To understand how our CCA improves the foreground representation quality, we analyze the properties of inter-class separation and intra-class adaptation for our generated feature space via the t-SNE [638] technique.

	Fixed Threshold	Derived from Foreground	Adversary Learning	Contrastive Learning
w/o TV	51.8	52.1	54.3	55.6
w/ TV	55.3	56.4	58.2	59.2

Table 4.4: Ablation study of background modeling options. Each item represents one background modeling method with or without TV. The top 2 best results are in blue and red respectively.

Figure 4.5 visualizes the distribution of all local feature samples, where solid circles denote all local representations and white plus symbols denote global foreground representations. Although a small proportion of confusing samples exists at the center, our generated feature space can still group most samples into their correct categories, demonstrating the ability to distinguish between classes. The visualization supports three important observations: (1) the background class is hard to model within a definite feature space due to the large variation; (2) contrastive learning can extract crucial information to represent image-specific background representations; (3) intra-class deviation causes incomplete coverages using the global prototypes (i.e., large variations of local samples to the white plus symbols). Furthermore, we analyze the global-to-local transformation within the foreground class, Person, that interacts most with other classes. Figure 4.6. visualizes the distributions of local feature samples co-occurring with various classes (e.g., motor, bicycle, car). CCA can adapt a single global prototype (i.e., the plus symbol) into local prototypes to address intra-class deviation. Additionally, we observe that the transformed local prototypes can capture crucial image-specific low-level features and can be partitioned into sub-groups (grey dotted circles). Therefore, our CCA can capture abstract information to conceptualize each foreground semantic group, while transforming these representations into local details to adapt to small intra-class variations across images. Therefore, it can generate more precise partitions to cover non-discriminative foreground regions.

Effects of geometrical regularization. We perform ablation experiments on the PASCAL VOC2012 train set to further validate the effectiveness of geometrical regularization. Figure 4.7. shows that background maps from contrastive learning may have inaccurate separations. We observe that the geometrical regularization term TV can efficiently constrain the size of activated regions and generate smooth boundaries separating background and foreground regions. The first row shows that the TV prior can suppress falsely activated regions and contract the boundaries tightly along the foreground. In addition, the second row demonstrates that TV can expand background regions to cover the confusing parts. Therefore, it can exclude background

4.4. Experiments and Results

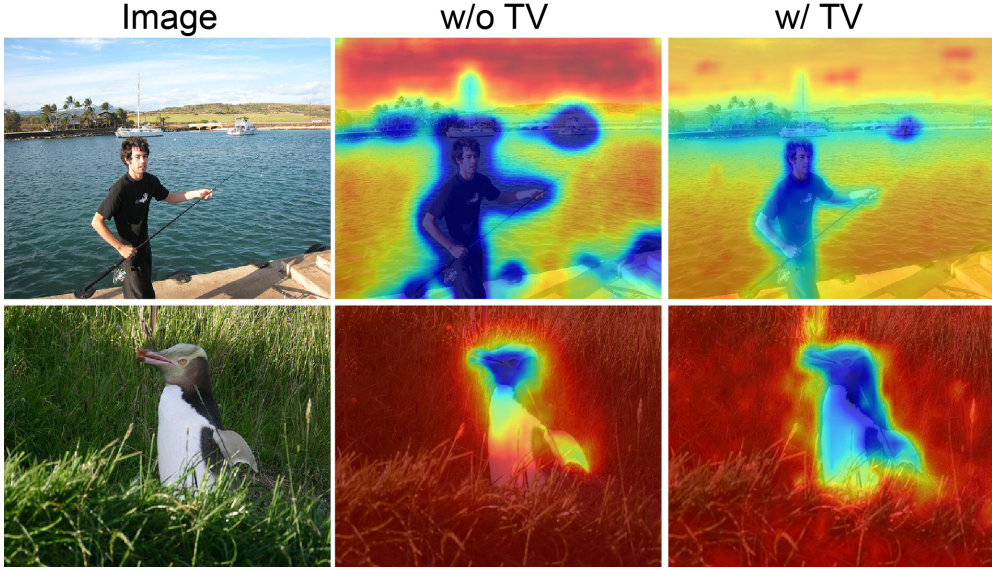


Figure 4.7: Ablation study on the effectiveness of TV term (see in Eq.4.17). TV term can suppress false response background regions (row 1) while expanding confusing regions (row 2). The original images are from the PASCAL VOC 2012 data.

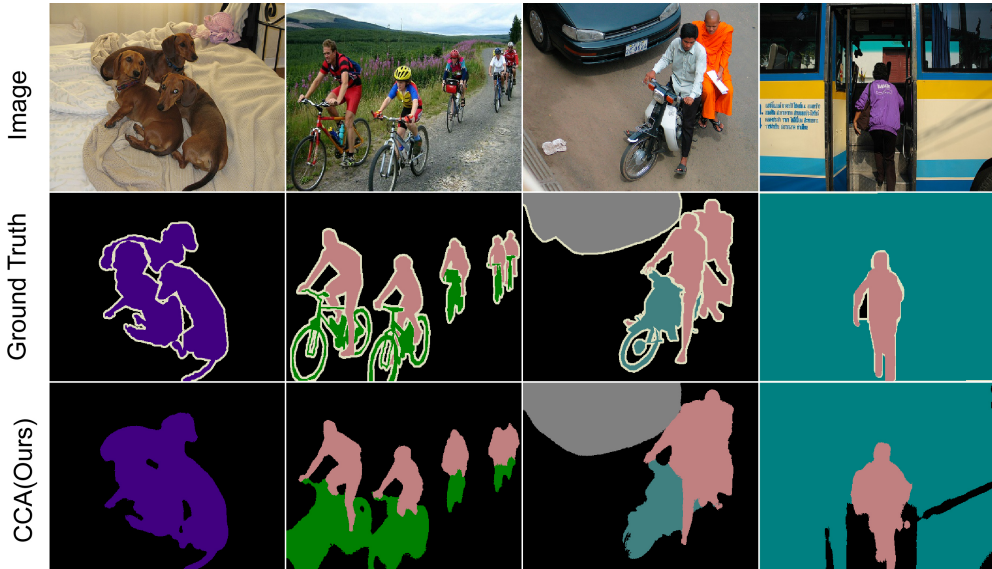


Figure 4.8: Qualitative visualization of semantic segmentation. The original images are from the PASCAL VOC 2012 data.

regions that are in close contact with the foregrounds. Table 4.4 shows quantitative

comparisons between different combinations. As shown, our geometrical regularization term consistently improves performance by suppressing falsely activated regions and expanding non-discriminative background regions.

4.5 Conclusion

This chapter presents a novel conditional probabilistic model with geometric regularization for WSSS. Our CCA approach provides a new understanding of representation learning within the WSSS paradigm. It separates semantic segmentation into two steps: first, extracting background regions, and then classifying foreground classes outside the background. We use Bernoulli distributions to model the background explicitly and train them with contrastive self-supervised learning. Since the initial background maps might be imprecise, we adopt a variational TV term to regulate their boundaries. Subsequently, we define conditional Categorical distributions within identified foreground regions to obtain conditional foreground classification and implement them as Contextual Cross-Attention to capture long-range dependencies. Comprehensive experiments demonstrate the superiority of our CCA over other WSSS methods. Future directions include exploring advanced geometric regularization techniques and extending our framework to fully-supervised and semi-supervised segmentation settings.

4.5. Conclusion

Chapter 5

Advances in Kidney Biopsy Lesion Assessment through Dense Instance Segmentation

Based on the following publication:

Zhan Xiong, Junling He, Pieter Valkema, Tri Q. Nguyen, Maarten Naesens, Jesper Kers, and Fons J. Verbeek. Advances in Kidney Biopsy Lesion Assessment Through Dense Instance Segmentation. *Artificial Intelligence in Medicine*, 164:103111, 2025

Chapter Summary

Renal biopsies are the gold standard for diagnosing kidney disease. Lesion scores made by renal pathologists are semi-quantitative and exhibit high inter-observer variability. Automating lesion classification within segmented anatomical structures can provide decision support in quantification analysis, thereby reducing inter-observer variability. Nevertheless, classifying lesions in regions-of-interest (ROIs) is clinically challenging due to (a) a large amount of densely packed anatomical objects, (b) class imbalance across different compartments (at least 3), (c) significant variation in size and shape of anatomical objects, and (d) the presence of multi-label lesions per anatomical structure. Existing models cannot address these complexities efficiently in a generic manner. This chapter presents an analysis for a **generalized solution** to datasets from various sources (pathology departments) with different types of lesions. Our approach utilizes two sub-networks: dense instance segmentation and lesion classification. We introduce **DiffRegFormer**, an end-to-end dense instance segmentation sub-network designed for multi-class, multi-scale objects within ROIs. Combining diffusion models, transformers, and RCNNs, DiffRegFormer is a computationally friendly framework that can efficiently recognize over 500 objects across three anatomical classes, i.e., glomeruli, tubuli, and arteries, within ROIs. On a dataset of 303 ROIs from 148 Jones' silver-stained renal Whole Slide Images (WSIs), our approach outperforms previous methods, achieving Average Precisions of 52.1% (detection) and 46.8% (segmentation). Moreover, our lesion classification sub-network achieves 89.2% precision and 64.6% recall on 21889 object patches out of the 303 ROIs. Lastly, our model demonstrates direct domain transfer to PAS-stained renal WSIs without fine-tuning.

Keyword: Renal Pathology Dense Instance Segmentation Diffusion Model Regional Transformers Multi-label Lesion Classification

5.1 Introduction

The evaluation of expert pathologists of kidney biopsies remains the gold standard for diagnosing and staging renal diseases [43]. Although biopsies digitalized into Whole Slide Images (WSIs)¹ have facilitated obtaining a visual morphological assessment of different anatomical structures for disease categorization, high-quality diagnostic assessments heavily depend on the correct lesion quantification manually annotated by pathologists across structures within a biopsy. Figure 5.1 shows an example of an annotated region-of-interest (ROI)² within a biopsy that contains hundreds of densely packed tissue objects. The annotation would cost a skilled expert around 2-4 hours for a complete biopsy. Due to the complexity and time-consuming nature of this task, there is a strong need for automated structure annotation and tools for lesion classification to facilitate further quantification, offload annotation time, and reduce intra- / inter-observer variability [9, 591].

Deep learning based instance segmentation algorithms have demonstrated significant capabilities on biomedical datasets [738, 42, 308, 562, 139, 782, 407, 474, 179]. However, developing a generic framework for renal pathology remains a challenge. This challenge can be broken down into four technical gaps and two practical issues. The four gaps (see Figure 5.1) are: (1) densely packed structures, up to 1000, per ROI; (2) considerable variation in size and shape of objects (e.g., arteries can be up to 100 times larger than that of tubuli); (3) class imbalance (e.g., the tubulointerstitial area occupies more than 70% on average in healthy and diseased renal parenchyma [42]); (4) each anatomical structure may present multiple lesions. The two practical issues are: (a) how to fuse multiple datasets with variation in staining to exploit scarce annotations fully; (b) readiness for extensibility, i.e., cost-effective adaptation to new lesion types from expanding datasets in clinical scenarios. Simultaneously addressing these difficulties requires a universal framework that includes efficiency, staining style (domain) transfer [187, 641, 316], and flexibility for continuous learning [830, 140].

Prior studies have demonstrated strong capabilities for lesion classification via dense instance segmentation, but their paradigms lack scalability and adaptability to datasets with potential changes in lesion composition. Some approaches [253, 42, 562, 437, 139] adopted a two-step process: semantic segmentation followed by dataset-specific post-processing to obtain final instance masks, which cannot be scaled to large-scale datasets

¹<https://www.mbfbioscience.com/whole-slide-imaging-analysis/>

²To prevent any confusion with 'RoI' in RCNNs, we will hereafter use the term 'ROI' solely in the context of renal biopsies. To maintain clear distinctions in RCNN discussions, we will adopt the terms 'candidate regions' or 'regional proposals'.

5.1. Introduction

with dense objects of various shapes. Besides, each lesion is assigned to a single semantic class, leading to difficulty with multi-label lesion classification and potential changes in lesion combinations. Specifically, adding or removing lesions requires a complete redesign and retraining of the model. Lastly, increasing the number of lesion types requires adding more segmentation maps for prediction, which is significantly resource-intensive.

Detection-based models with regional convolution neural networks (RCNNs) [301, 308] can efficiently detect dense anatomical structures and separately design lesion classification heads for each class. They can therefore adopt a plug-and-play mechanism and adapt to lesion changes by replacing the corresponding sub-modules and maximally reusing the others. However, these variants are not scalable to multi-class objects with varying scales and shapes because they rely on predefined bounding box anchors, limiting their application to a single class with less variation in scale and shape, e.g., glomeruli.

Recently, transformers have been the prevailing anchor-free approach for multi-class object detection across various scales and shapes. Transformer-based instance segmentation leverages attention mechanisms and learns latent representative embeddings (queries) from global contextual features to process highly varied objects. However, existing models [53, 105, 573] are resource-intensive for dense structures in large-scale datasets. That is due to two shortages: (1) a large number of static queries from one embedding per object; (2) low instance map occupancy from one object per instance map (depicted in Figure 5.2c). Hence, they are limited to processing classes with sparse objects, like glomeruli.

In response to these challenges, we propose a generic and extensible system for dense instance segmentation and lesion classification on large-scale datasets, robust to changes in lesion combinations. Our design has two key components: (a) A novel dense instance segmentation sub-network that recognizes basic anatomical structures, i.e., glomeruli, tubuli, and arteries; (b) a lesion classification sub-network with a set of independent heads that predict lesions for each class. It is crucial to separately modularize lesion classification and dense instance segmentation for the big picture. The segmentation of basic anatomical structures is a universal foundation for all visual assessment systems. However, lesion classification is a task-driven downstream application that may vary depending on the interests of users. Therefore, the robustness of dense instance segmentation to local changes in lesion classification heads is beneficial for the extensibility of our system and for expanding datasets in clinical scenarios.

More specifically, for dense instance segmentation, we propose a novel end-to-

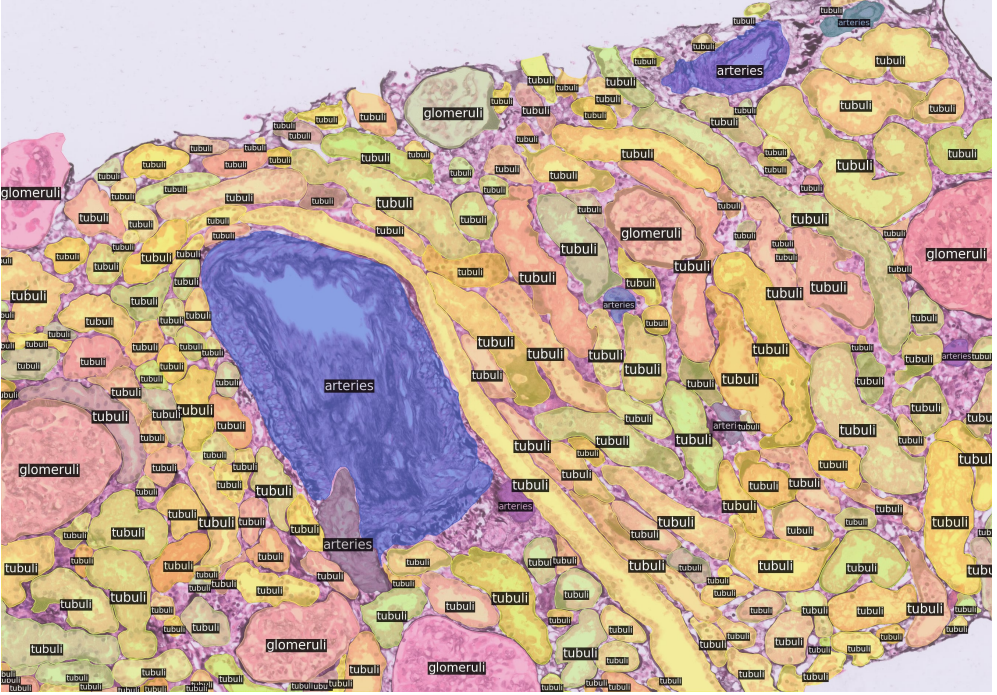


Figure 5.1: An illustration depicts a manually annotated ROI of a kidney biopsy, focusing on **glomeruli**, **tubuli**, and **arteries**. There are three primary challenges encountered in clinical renal biopsies: (1) a large number of objects closely touching each other; (2) the significant variation in size and shape among different instances; (3) the distribution of classes is heavily biased.

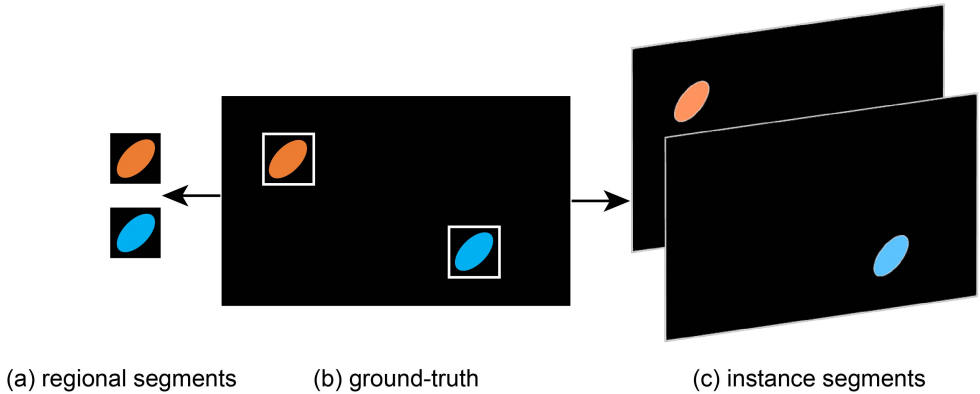


Figure 5.2: Comparison between regional segments and instance segments. (a) cropped instance maps from bounding boxes tightly surrounding each object; (b) ground-truth bounding boxes and instance segments within one image; (c) separated entire instance maps per object.

5.1. Introduction

end approach, named **DiffRegFormer**, which effectively combines the advantages of diffusion models, RCNNs, and transformers to tackle dense objects across multiple classes and scales. Diffusion models [260] generate bounding box proposals from Gaussian noise, and therefore eliminating the need for pre-defined anchors; RCNNs efficiently crop dense instance maps into dense regions which have very high occupancy rate w.r.t. the bounding boxes (see Figure 5.2. (a)); Cross-attention mechanisms with dynamic queries [65, 367, 808, 109, 368] extract long-range contextual features and enables robust representation of objects across varying scales. DiffRegFormer is not just a simple assembly of the aforementioned techniques. Such assemblies may suffice for sparse instance segmentation on datasets like MSCOCO [415]. They, however, fail in the context of dense instance segmentation of kidney biopsy ROIs. The crux lies in the proposals, which are essentially Gaussian noise during the early training stage of diffusion models. Those noisy candidates, consequently, hinder training due to accumulating errors. Our ablation study analyzes these effects in detail (see section 5.4.4) and highlights the importance of our specific designs. Specifically, we introduce the following **key innovations** to address the challenges for dense instance segmentation:

- **Regional features:** Similar to RCNNs, feature maps are cropped using generated proposals and converted to dynamic queries for efficient long-range dependency modeling; this is robust to large-scale variations.
- **Feature disentanglement:** Instead of using shared feature maps, we redesign the model structure to generate separate feature maps for the bounding box decoder and mask decoder; this is crucial to stabilizing the training of the mask decoder.
- **Class-wise balanced sampling:** Unlike conventional sampling methods, we propose a novel sampling approach. The key difference is to select **class-wise balanced** positive samples among the **ground-truth** boxes instead of the proposal ones.

In conclusion, based on these key designs, the proposed model is a generic and extensible approach for large-scale datasets with minimal overhead. To the best of our knowledge, DiffRegFormer is the first end-to-end framework that combines a diffusion-based method with a transformer in an RCNN-style architecture to efficiently process dense objects within ROIs for multi-scale and multi-class tasks. Figure 5.3 depicts a flow chart of the dense instance segmentation sub-network.

Each anatomical structure has a dedicated classification head for the lesion classifier, since lesion classes can vary across datasets. In Figure 5.6, we input the croppings of anatomical structures and resized them to the same size. Each head is trained only with patches of the specific structures and learns to predict multi-label lesions over each class. Our dataset currently only contains sclerotic glomeruli and atrophic tubuli. However, the lesion classifier is extensible to the expanded large-scale dataset at minimal cost, because adding, removing, or replacing heads does not disrupt the other modules.

Our work aims to propose a **generic technical solution** that has flexibility and scalability for automating lesion classification on dense anatomical structures in clinical scenarios. It works as the foundation for further quantification analyses. The main contributions of this chapter are the following:

- We propose the first end-to-end dense instance segmentation that effectively combines diffusion models, RCNNs, and transformers for multi-class, multi-scale objects. More importantly, our model can directly process ROIs of kidney biopsies, avoiding patch splitting.
- We propose novel designs to address the accumulative errors caused by diffusion models at an early stage and stabilize the training process.
- We propose a novel class-wise balance sampling method to improve detection and segmentation performance.
- Instead of the static queries used in transformers, we convert regional features into dynamic queries and model the long-range dependencies between regional features while avoiding performance deterioration on dense objects.
- We compare DiffRegFormer with previous models that can process multi-class and multi-scale objects within ROIs in an end-to-end manner. Our model outperforms the previously published models in evaluating Jones’ silver-stained images.
- We show that DiffRegFormer has the potential of stain (domain)-agnostic detection for PAS-stained images without stain-specific fine-tuning.
- Our lesion classifier can achieve multi-lesion classification. In addition, our plug-and-play strategy can flexibly adapt to large-scale datasets while accommodating potential changes in lesion combinations with minimal overhead.

The remainder of the chapter is organized as follows. In section 5.2, all relevant work is introduced, and we elaborate on all improvements w.r.t. previous research. In section

5.2. Related Work

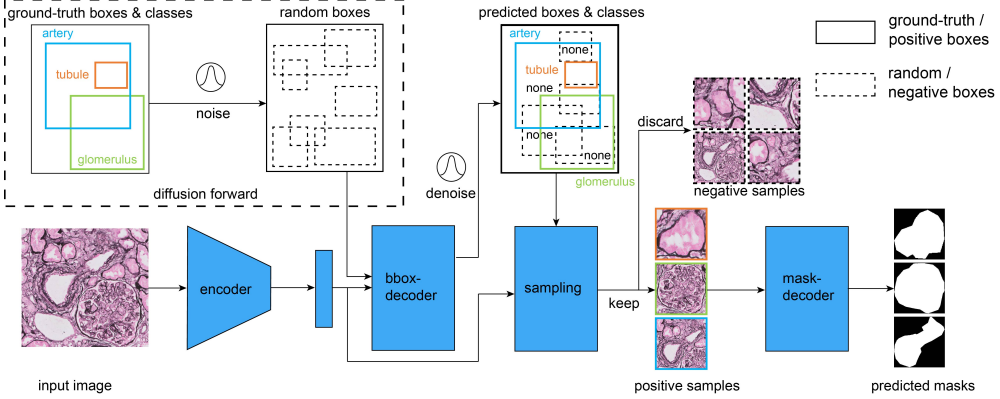


Figure 5.3: Our **DiffRegFormer** is a one-stage anchor-free method. Instead of predefined anchors, we add Gaussian noise to ground-truth boxes and generate a fixed-sized set of random bounding boxes (bbox). With feature maps extracted from the encoder, the bbox-decoder iteratively learns to denoise and predicts class-wise candidate boxes. Due to the uneven distribution of candidates, we propose a sampling module that effectively discards negative samples while maintaining a balanced proportion of positive samples to facilitate fast convergence. Finally, we generate the output instance masks based on the selected positive samples in the mask-decoder. For simplicity of illustration, we choose only one anatomical object per class (artery, tubule, glomerulus).

5.3, we describe each component of our model in detail. In section 5.4, we demonstrate the extensive evaluation of our model, including comparison experiments and ablation studies. In section 5.5, we describe the advantages and limitations of our pipeline and show possible future follow-up research.

5.2 Related Work

Semantic Segmentation with Post-Processing: In the context of lesion classification of dense structures for renal biopsies, semantic segmentation with post-processing has drawn substantial attention. Previous studies [253, 42, 562, 139] treat anatomy-wise lesions (e.g., atrophic tubuli) as semantic classes alongside basic anatomical structures (e.g., arteries, glomeruli). An auxiliary **border** class is often used during training to facilitate splitting semantic masks into distinct instances. This approach has demonstrated notable performance. However, challenges are faced in scalability and adaptability to large-scale datasets with potential changes in lesion composition. These issues are at hand as (1) Reliance on ad-hoc procedures (e.g., thresholding and morphological operations [200]) for border generation reduces generalizability. (2) Due to the inflexibility of dataset change, any modification to lesion types necessitates

model redesign and retraining. (3) Computational cost scales with lesion classes in the dataset. (4) The difficulty in modeling long-range correlations between objects of varying size due to splitting ROIs into tiles. (5) It cannot tackle multi-label lesion classification. In contrast to these issues, our end-to-end framework addresses them by directly modeling dense objects within ROIs using attention mechanisms.

Diffusion Model: Diffusion models, a class of deep generative models [260, 492], learn to approximate complex distributions through iterative denoising. Despite success in data generation [278, 499, 374], natural language processing [380], audio processing [518], and self/weakly-supervised learning [170], their application to dense instance segmentation remains limited. Initial attempts have been to introduce diffusion models to segmentation tasks [85, 209]. The adaptation to dense instance segmentation, however, remains challenging. From our ablation studies, we formulate two primary challenges: (1) The potential dominance of the diffusion modules over feature learning in shared feature maps destabilizes the training of instance prediction modules. (2) Accumulative errors from the diffusion module’s early training stages impair the accuracy of dense object predictions. Our approach introduces innovative strategies to overcome these challenges, enabling the first successful implementation of a diffusion model for dense instance segmentation.

Transformer: Transformer-based approaches introduce an anchor-free paradigm in an end-to-end manner for instance segmentation, leveraging their capability to capture long-range dependencies and aggregate contextual information. In general, the dependency extraction process is called an attention mechanism, and each context feature is treated as a query for object representation. Early transformer-based methods [53, 105, 573] rely on static queries to model all objects within a dataset. These approaches are incompatible with dense object scenarios because the static query set inevitably expands with dataset size. Recent works [367, 808, 109, 368] advancement towards dynamic queries, allowing for temporary modeling of objects within a mini-batch. In this manner, the complexity of processing dense objects is reduced. However, their instance segment representation remains inefficient for dense instance segmentation³ due to the expensive representation of instance segments (Figure 5.2.(c)) with low occupation rate. Our DiffRegFormer addresses this by integrating region proposals with transformers for improved representation efficiency.

³We conducted experiments on our dataset from the official code of both. Compared to our approach, they required more than twice the GPU memory at a batch size of 2 and 500 queries. That is not infeasible for applications with dense objects.

5.3. Methods

RCNN: The Regional Convolutional Neural Network (RCNN) framework [197, 545] effectively generates candidate regions (i.e., proposals) for dense objects. In instance segmentation, RCNN-based methods use mask modules to predict binary instance masks within proposals. Early variants [245, 284, 301, 308] relied on class-agnostic proposals followed by class-specific mask prediction, constrained by predefined anchors. More recent approaches [177] have explored the mapping from proposals to static queries, employing cross-attention mechanisms for instance mask prediction. However, this paradigm faces scalability issues in dense object segmentation due to the computational overhead associated with static queries (see QueryInst in Table 5.2). Our method adopts dynamic queries, balancing computational efficiency and scalability for dense instance segmentation.

In summary, our work combines diffusion modeling and attention mechanisms into an RCNN-style framework that addresses the scalability, adaptability, and efficiency challenges inherent to existing approaches for dense instance segmentation in renal biopsies.

5.3 Methods

In this section, we first introduce the basic conceptions for the diffusion models. Then, details of all core components are elaborated. Finally, implementation information is added for the training and inference stages.

5.3.1 Preliminaries

Diffusion model. Diffusion approaches [583, 548] are a family of deep generative models inspired by the principles of non-equilibrium thermodynamics [581, 278]. Their operation is conceptualized as a T -step sequential process that iteratively transitions from an initial state Z_0 to a final state Z_T , which is pure noise. Diffusion forward involves progressively adding Gaussian noise at each transition. Due to the special properties of the Gaussian distribution [260], it is possible to sample any intermediate noisy state Z_t directly:

$$q(Z_t|Z_0) = \mathcal{N}(Z_t|\sqrt{\bar{\alpha}_t}Z_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (5.1)$$

where Z_0 is the original data, Z_t ($t \leq T$) is the latent state at step t , and $\bar{\alpha}_t$ is the cumulative noise variance schedule [85]. Notably, the diffusion forward process does

not introduce trainable parameters. The objective of a diffusion model is to learn the reverse process. A neural network, $f_\theta(Z_t, t)$, approximates the recovery of the original distribution Z_0 from a noisy state Z_t . Training optimizes this network by minimizing a loss function to denoise at an arbitrary intermediate state t :

$$\mathcal{L} = \frac{1}{2} \|f_\theta(Z_t, t) - Z_0\|^2 \quad (5.2)$$

where Z_0 represents the ground-truth and Z_t is the noisy input. During inference, the model f_θ reconstructs from pure noise Z_T to the predicted data Z_0 through a series of reverse updates [583] with a step-size s : $Z_T \rightarrow Z_{T-s} \rightarrow \dots \rightarrow Z_0$. See [583] for a detailed derivation.

5.3.2 Dense Instance Segmentation Model

Iterative proposal generation in DiffRegFormer requires multiple model executions during inference. That poses a computational burden if the entire model runs on the raw image at each training step. To address this issue, we decouple the image encoder from the remaining modules. This strategy enables the encoder to extract multi-scale feature maps from the raw input image only once. Then, subsequent modules operate iteratively to refine box and mask predictions from an initial state of Gaussian noise, conditioned on these deep features. Moreover, to facilitate the mask-decoder training, we introduce a novel sampling module that randomly selects positive samples from each class, ensuring a balanced distribution of class-wise regional features. That facilitates the learning of object representations at lower frequencies in the dataset.

In summary, the DiffRegFormer comprises four components:

- **Encoder:** Processes the raw input image once to extract multi-scale feature maps for the bbox-decoder and mask-decoder separately.
- **Bbox-decoder:** Utilizes the extracted feature maps to iteratively refine bounding box predictions.
- **Sampling Module:** Stabilizes the training by randomly selecting class-wise balanced positive samples for the mask-decoder.
- **Mask-decoder:** Iteratively refines mask predictions based on the sampled regional features.

5.3. Methods

Encoder

The image encoder extracts high-level feature maps from the raw image for subsequent decoders. To generate multi-scale feature maps, it utilizes a ResNet [248] pre-trained on ImageNet [138] as its backbone, followed by a Feature Pyramid Network (FPN) [414]. After FPN, a stack of 3×3 convolutional operations is applied to produce two separate feature maps, thereby decoupling the bbox-decoder and mask-decoder.

Bbox-decoder

The bbox-decoder builds upon DiffusionDet [85]. As shown in Figure 5.4 (a), a set of random boxes and multi-scaled feature maps is first taken from the encoder as input. Then, dynamic queries (Figure 5.4b) are initialized from a regional pooling operator (RoIAlign [245]) and a feed-forward network (FFN) [642]. These queries undergo iterative refinement through multiple stages. Each stage takes proposal boxes and dynamic queries from the previous stage, generating refined dynamic queries, box predictions, and class predictions for the next stage. Figure 5.4c illustrates one detailed refinement stage. First, new regional features (cropped by RoIAlign) interact with dynamic queries (after a self-attention module) via dynamic convolution [97], emphasizing regions likely containing objects while suppressing others. Next, the enhanced regional features are processed by FFNs to produce refined box predictions, box-wise classifications, and queries, respectively.

Sampling

Conventional RCNN models use intersection over union (IoU) to classify a proposal box as positive if its IoU with a ground-truth box is ≥ 0.5 [245]. This approach, however, is ill-suited to diffusion methods, particularly in early training stages when they predominantly produce noise. That can complicate convergence or even cause failure due to error accumulation. Further, conventional IoU-based methods are susceptible to class imbalance, biasing the learning process towards frequently occurring objects and neglecting rare instances. We therefore propose a novel sampling method to address these issues. We first replace proposal boxes with ground-truth boxes to ensure a sufficient supply of positive samples. That is justified because the bbox-decoder aims to predict boxes that closely approximate the ground-truth. We then divide the ground-truth bounding boxes into groups by class. Within each group, we randomly select up to n boxes as positive samples without repetition. The core idea is that objects of each class that have similar shapes and appearances can be mapped to a cluster. Consequently, it allows us to learn instance feature representations with a

small sample size n . The rule especially applies to the tubular class, which accounts for most of our dataset. This sampling strategy offers three distinct advantages:

1. **Rare Instance Learning:** The mask-decoder can effectively learn representations of infrequently occurring instances whenever they appear in the input, mitigating the risk of being overwhelmed by more common classes.
2. **Randomize Frequent Instance:** Learning on a randomized subset of frequent instances prevents model overfitting.
3. **Training Stability:** Employing selected ground-truth boxes as positive samples promotes stability and accelerates mask training.

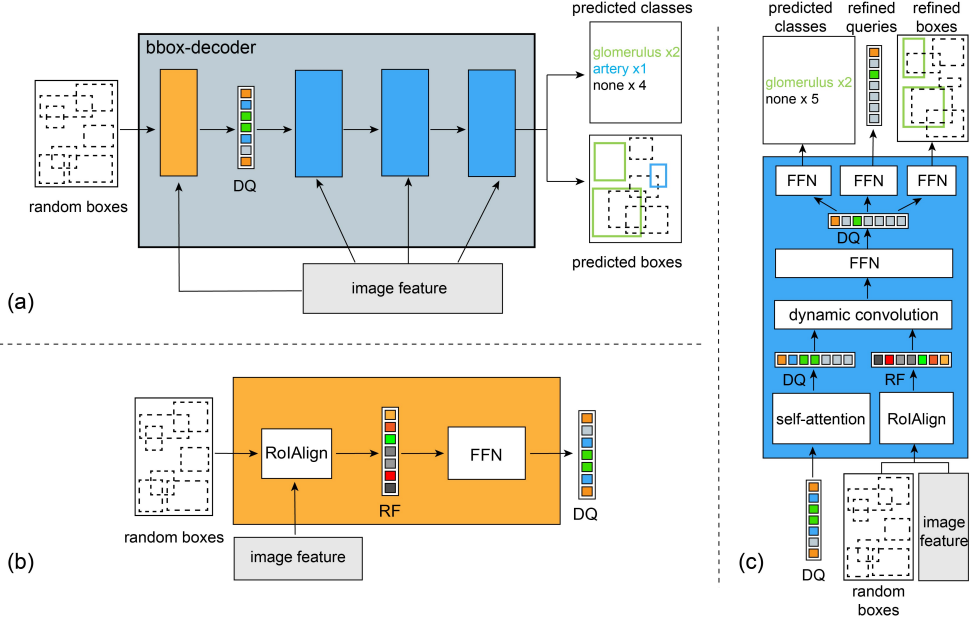


Figure 5.4: The bounding box (bbox)-decoder takes multi-scale feature maps and a set of random boxes as input. Then, the predicted classes and boxes will be output iteratively. (a) The module comprises an initialization head for dynamic queries (orange rectangle) and multiple box refinement heads (blue rectangles). (b) In the orange rectangle in (a), the initial dynamic queries are generated via a RoIAlign pooling operator and a feed-forward network (FFN). (c) Each box refinement module (one blue rectangle in (a)) takes the previous stage's dynamic queries and proposal boxes as input, generating predictions and refined dynamic queries for the next stage. Abbreviations: dynamic queries: DQ, feed-forward network: FFN, regional features: RF.

5.3. Methods

Mask decoder

The mask decoder iteratively refines mask predictions by using regional features cropped from multi-scale feature maps, based on boxes selected by the sampling module. Instead of using static queries [105, 106], we compute cross-attention between dynamic queries and positive regional features. That highlights effective queries and filters, enhancing regional features by focusing on pixels within positive regions, thereby facilitating accurate per-pixel instance score prediction (Figure 5.5).

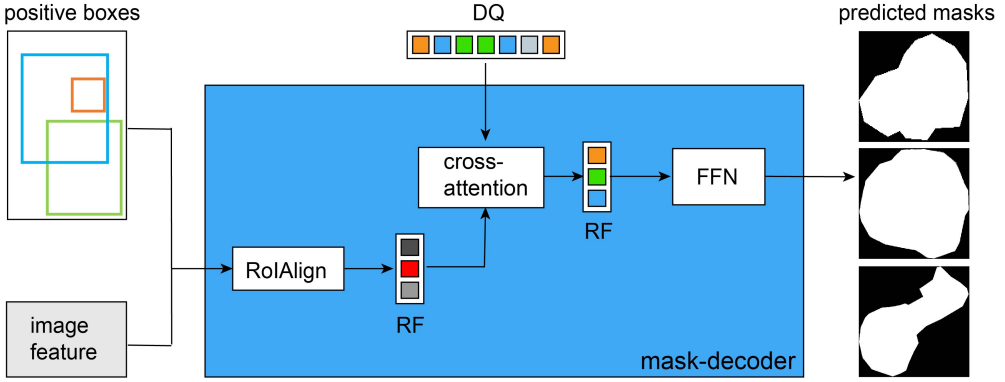


Figure 5.5: The mask-decoder takes multi-scale feature maps and a set of positive boxes as input and predicts instance masks. The dynamic queries interact with regional features via cross-attention and highlight only pixels within proposal boxes. Final instance masks are generated from the enhanced regional feature maps. Abbreviations include DQ: dynamic queries, FFN: feed-forward network, and RF: regional features.

5.3.3 Lesion classifier

The lesion classifier comprises multiple independent lesion prediction heads. Each head specializes in a specific anatomical structure (e.g., the glomerulus) and processes cropped images of that structure to predict probabilities for multi-label lesions. Figure 5.6 illustrates this design. The architecture of a single head consists of three convolutional blocks, followed by a max pooling operator. Each block contains a 3×3 convolution, batch normalization, and ReLU activation. We currently implement two prediction heads: one for glomeruli (predicting sclerosis) and one for tubuli (predicting atrophy).

Notably, the modular design of the instance segmentation component allows lesion prediction heads to operate as independent plug-ins. This modularity is crucial for seamlessly adapting the model to the expanding datasets, particularly in clinical scenarios. It allows for the straightforward integration or removal of prediction heads

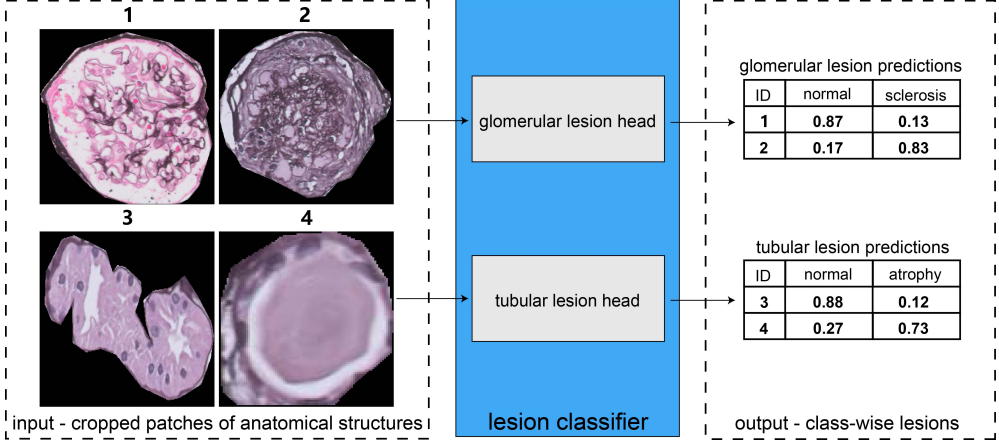


Figure 5.6: The lesion classifier consists of 2 prediction heads, one dedicated to glomeruli and the other to tubuli. It takes croppings of anatomical structures and outputs the probability of class-wise lesions.

with minimal impact on the overall model architecture. For example, if new annotations for arteries become available, a dedicated prediction head can be trained to classify arterial lesions and integrate them into the model without modifying existing components. Likewise, this design extends the model’s diagnostic scope to include additional lesions within a given structure (e.g., tubuli) by simply replacing an existing prediction head with a newly trained one on expanded data. This design promotes the model’s versatility and long-term adaptability within clinical scenarios where the scope of identifiable lesions may increase.

5.3.4 Implementation Details

The entire model employs distinct strategies for training and inference. During training, the DiffRegFormer learns to reconstruct ground-truth boxes through adaptation at state 0 from an arbitrary noisy state t (Figure 5.7(a)). Additionally, the sampling module provides class-balanced ground-truth boxes to facilitate the mask decoder’s training. Lastly, the lesion classifier operates on croppings generated from ground-truth instance masks. Each prediction head specializes in classifying lesions specific to the corresponding anatomical structure within these croppings. At the same time, during inference, the DiffRegFormer assembles multiple predictions to refine final results (Figure 5.7(b)). Moreover, the sampling module is deactivated, and all proposals are directly forwarded to the mask decoder for instance segmentation prediction. Likewise, the lesion classifier utilizes instance masks predicted by DiffRegFormer

5.3. Methods

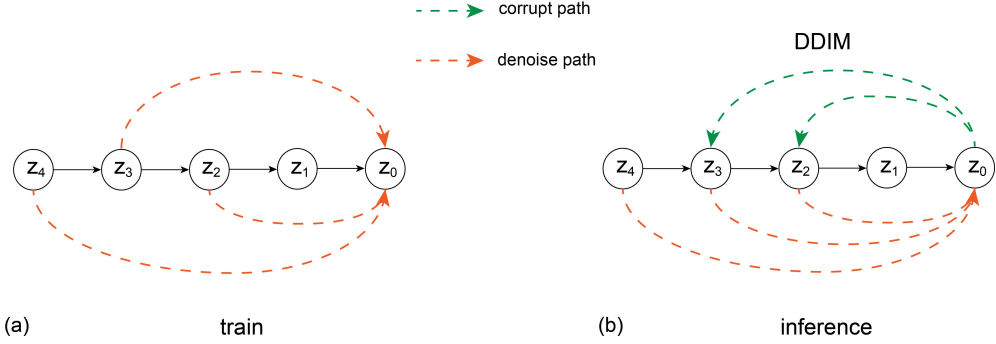


Figure 5.7: An illustration for a diffusion model at different stages depicts a simplified process with 5 states (skip step=1), where state 4 is pure Gaussian noise and state 0 is the ground-truth. (a) During training, a diffusion model randomly selects denoising paths to train its decoder; (b) In inference, the Denoising Diffusion Implicit Model (DDIM) flexibly alternates between denoising and corrupting paths, generating multiple predictions that are then ensembled for a refined final result.

to crop structures. As a final step, each prediction head outputs probabilities for class-specific lesions within the cropped regions.

Training

We start by constructing noisy boxes from the ground-truth. The DiffRegFormer is then trained to recover bounding boxes and iteratively generate instance masks from an arbitrary state t . The algorithm is detailed in Algorithm 1.

Diffusion forward. Ground-truth boxes are first concatenated with additional random boxes to align to a fixed length across images, accounting for the variability in instance counts. Gaussian noise is then introduced as the noise scale governed by $\bar{\alpha}$ as per Eq.(5.1). Following [492], the scale varies with state t using a monotonically decreasing *cosine* schedule. It should be noted that, for optimal performance, the signal-to-noise ratio requires a relatively high signal scaling value compared to image generation tasks [88, 260, 144, 90, 85].

Training Loss. The box decoder takes noisy boxes as input and predicts category labels and box coordinates. Given the one-to-many relationship between ground-truth and predicted boxes, first, a set-loss function is employed to match predictions to the ground-truth based on the lowest cost [53, 602, 870]. Furthermore, the mask decoder inputs sampled positive boxes and predicts instance masks. We use a binary cross-entropy loss function due to the one-to-one mapping. Finally, since multiple lesions can occur within a single crop, lesion classification uses a binary cross-entropy

loss function for multi-label prediction within each head.

Sampling. Our implementation uses 3 classes and a maximum of $N = 500$ objects per image. We set up to $n = \frac{500}{3} \simeq 166$ instances per class. More specifically, positive ground-truth boxes are sampled without replacement, ensuring they do not exceed the class-wise maximum. Moreover, we disregard classes that are absent in the input. For example, if there are 2 glomeruli, 0 vessels, and 200 tubuli ground-truth boxes within an ROI, then the sampling module would select 2 glomeruli, 0 vessels, and 166 random tubuli as positive samples without duplication.

Algorithm 1: DiffRegFormer Training

```
def train(images, gt_boxes, gt_masks):
    # images: [B,H,W,3], gt_boxes: [B,*,4], gt_masks: [B,H,W,G]
    # B: batch size, N: number of proposal boxes

    # generate multi-scale features via encoder
    feats = encoder(images)
    # generate noised boxes at state t, where t is a random integer
    # in diffusion forward; Pad boxes to N with t_boxes: [B,N,4]
    t_boxes, t = diffusion(gt_boxes, mean=0, std=1)
    # generate initial dynamic queries
    d_query = initialize_query(t_boxes, feats)
    # learn to reverse noised boxes at state t back to ground-truth
    # boxes at state 0 and return refined dynamic queries
    [pred_boxes, d_query] = box_decoder(t_boxes, feats, d_query, t)
    # obtain bbox loss via the set objective function
    loss_bbox = set_loss(pred_boxes, gt_boxes)
    # randomly select balanced positive boxes from ground-truth boxes
    pos_boxes = sampling(gt_boxes)
    # generate predicted instance masks
    pred_masks = mask_decoder(pos_boxes, feats, d_query)
    # obtain mask loss via the set objective function
    loss_mask = set_prediction_loss(pred_masks, gt_masks)
    return loss_bbox, loss_mask
```

Inference

During inference, the DiffRegFormer starts from a standard Gaussian distribution corresponding to the final state T as defined in Eq.(5.1). Then, a progressive denoising operation reverses the predictions to the initial state 0. The algorithm, as depicted in Algorithm 2, outlines this process.

Inference steps. Each inference iteration involves two primary operations. First, the box decoder processes noisy boxes generated from the previous state t_{now} and predicts both categories and box coordinates. Subsequently, a Denoising Diffusion

5.4. Methods

Implicit Model (DDIM) [480, 583] introduces noise into the previously predicted boxes. This step generates new noisy boxes for the subsequent state t_{next} , thereby facilitating progressive denoising across states. Notably, each iteration performs a single denoising operation, transitioning from state t_{now} to t_{next} . This characteristic allows for the assembly of intermediate predictions from multiple denoising steps (i.e., $t_i \rightarrow 0$ where $i \in [1, T]$) to refine the outcome. However, DiffRegFormer performs inference only once to improve efficiency, directly transitioning from pure Gaussian noise at state T to state 0 to obtain the final predictions.

Box replacement. Predicted boxes can be categorized as either *positive* (scores above a threshold, containing objects of interest) or *negative* (scores below the threshold, arbitrarily located). Applying negative boxes directly to DDIM would significantly degrade the quality of newly generated noisy boxes. This issue stems from negative boxes that originated from corrupted boxes during the training phase, which significantly deviate from the Gaussian distribution. To ensure consistency between the training and inference phases, we substitute negative boxes with random boxes sampled from a Gaussian distribution.

Algorithm 2: DiffRegFormer Inference

```
def infer(images, step, T):
    # images: [B,H,W,3], step: skip length of state transform
    # T: chain length, B: batch size, N: proposal box number

    # generate multi-scale features via encoder
    feats = encoder(images)
    # return Gaussian noise as noisy boxes at state T;
    # Pad boxes to N with T_boxes: [B,N,4]
    [t_boxes, _] = diffusion(mean=0, std=1)
    # generate state transform pairs skipping every step
    # [(T, T-step), (T-step, T-2*step), ..., (step, 0)]
    time_pairs = uniform(0, T, step)
    # generate initial dynamic queries
    d_query = initialize_query(t_boxes, feats)
    # iterate over stages
    for (t_now, t_next) in time_pairs:
        # predict boxes and dynamic query at state t_now
        [pred_boxes, d_query] = box_decoder(t_boxes, feats, d_query, t_now)
        # generate new noisy boxes from state t_now to t_next
        t_boxes = ddim(t_boxes, pred_boxes, t_now, t_next)
        # replace undesired boxes with random Gaussian noise
        t_boxes = box_repalce(t_boxes)
    # generate predicted instance masks
    pred_masks = mask_decoder(pred_boxes, feats, d_query)
    return pred_boxes, pred_masks
```

5.4 Results

This section begins with an overview of the kidney biopsy dataset used in this study. We then make a fair comparison between DiffRegFormer and established end-to-end instance segmentation models specifically designed to handle dense, multi-class, and multi-scale objects at the ROI level. In addition, to include retrospective reflection in our analysis, we have compared our model with a semantic segmentation based method [253]. The results of this comparison are presented in section 5.4.2 and the tables therein. All benchmark methods are reproduced utilizing the **mmdetection** package [75] to ensure consistency. Further, we evaluate the performance of our lesion classifier on cropped images of anatomical structures derived from the instance masks predicted by DiffRegFormer. Additionally, we have conducted a complexity analysis of DiffRegFormer to identify an optimal balance between its performance and computational burden. Finally, we present extensive ablation studies on DiffRegFormer that evaluate the impact of various training strategies.

Source Center	Scanner	WSIs	ROIs	Train Dataset	Eval Dataset
LUMC	Philips UFS	65	142	130	12
AUMC	Philips UFS	42	80	74	6
UMCU	Hamamatsu XR	25	49	45	4
Leuven	Philips UFS	16	32	-	32

Table 5.1: Origin of datasets digitized from various source centers with different scanners. LUMC: Leiden University Medical Center. AUMC: Amsterdam University Medical Center. UMCU: The University Medical Center Utrecht. Leuven: KU Leuven.

5.4.1 Datasets

The biopsy specimens were prepared according to the Pathology Laboratory Protocol for kidney biopsies. A standard procedure of operations collected the tissues: core needle biopsy, formalin fixed, rapidly processed by a routine tissue processor, embedded in paraffin, serially sectioned at $2-4\mu m$, mounted onto adhesive slides, and stained with Jones silver staining. The dataset comprises 148 patient biopsy samples (both transplant and native kidneys) collected from the multi-center archives of the Departments of Pathology at LUMC, AUMC, and UMCU in the Netherlands, or from the archive of the Department of Nephrology at Leuven in Belgium. Before analysis, all biopsies were anonymized by a pathology staff member. Annotation was performed using the **ASAP**

5.4. Results

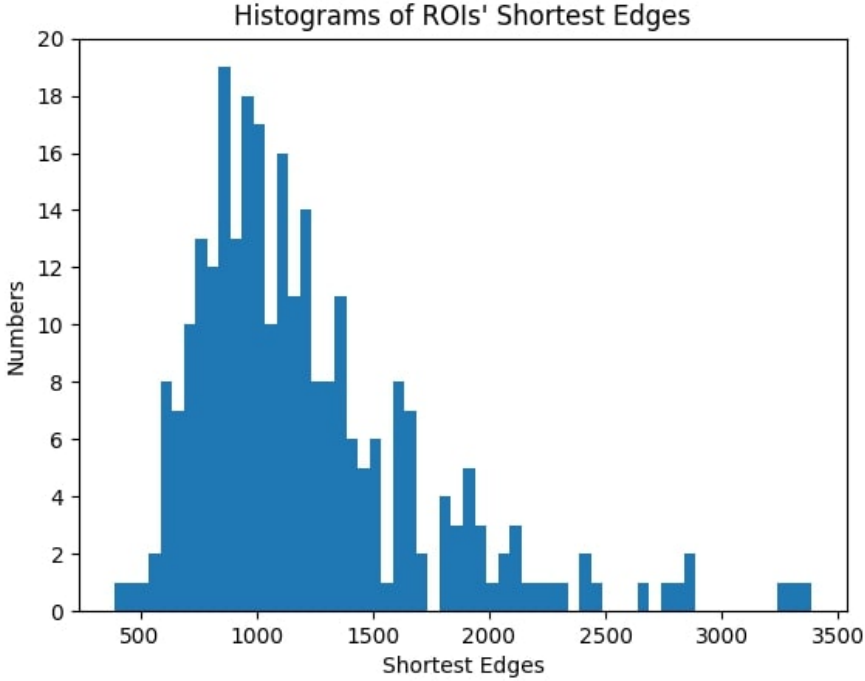


Figure 5.8: The histogram of the shortest edges of ROIs extracted from WSIs. In our dataset, the shortest ROI length is 387 pixels, and the largest is 4565 pixels.

version 1.9 for Windows⁴ and our custom software **Slidescape**⁵. The 148 Jones-stained renal WSIs were acquired from scanners at a resolution of $\text{PPM} = 2$ in BIG-TIFF format⁶. Openslide⁷ was used to extract 303 ROIs at level 0 of the BIG-TIFF images. Table 5.1 outlines the dataset composition, and Figure 5.8 illustrates the distribution of ROI short sizes. Furthermore, only 115 LUMC PAS-stained renal WSIs were used to test stain transfer. In particular, ROIs from the same WSIs were exclusively assigned to either the training or the validation set. To train the lesion classifier, 21,889 patches of individual objects were extracted from 303 ROIs, including 916 glomeruli patches with binary lesion labels (382 with sclerosis and 534 normal) and 20,973 tubuli patches with binary labels (10,647 with atrophy and 10,326 normal), as detailed in Table 5.4.

⁴<https://computationalpathologygroup.github.io/ASAP/>

⁵<https://github.com/amspath/slidescape>

⁶<https://www.awaresystems.be/imaging/tiff/bigtiff.html>

⁷<https://openslide.org/>

5.4.2 Experiments

In our experimental analysis, we conducted a fair comparison of our proposed DiffRegFormer against well-established end-to-end instance segmentation models: Mask R-CNN [245], Cascade Mask R-CNN [50], and QueryInst [177], specifically designed to simultaneously handle dense, multi-class, multi-scale objects at the ROI level. All the models under comparison employed a ResNet-50 [248] backbone pre-trained on the ImageNet [138] and were trained on 249 ROIs over 40000 iterations. We utilized the **mmde-tection** data augmentation pipeline: each ROI was first subjected to *RandomFlip* with a 0.5 probability, followed by a random selection between *RandomResize* and *RandomCrop*, also with a probability of 0.5. *RandomResize* adjusted the ROI to a size with the shortest edge between 480 and 1333 pixels, while *RandomCrop* extracted patches with the shortest edge between 384 and 600 pixels. Finally, pixel values were normalized to $(-1, 1)$. As **mmde-tection** supports training with variable image sizes, all models were allowed to train on objects at multiple scales.

For feature extraction, we used 256 channels for queries. During both training and inference, DiffRegFormer employed **500** dynamic queries, while QueryInst used 300 and 500 static queries, respectively. Cascade Mask R-CNN, QueryInst, and our method each used 6 cascading stages. Table 5.2 presents performance comparisons for object detection w.r.t. bounding boxes and instance segmentation w.r.t. masks, respectively. For all experiments, DiffRegFormer did not utilize refinement during inference (i.e., iteration set to 1). Notably, Mask R-CNN [245] and cascade Mask R-CNN [50] are two-stage methods that employ RPN networks rather than dynamic queries. QueryInst is a one-stage method with static queries. In contrast, DiffRegFormer is categorized as a one-stage method with dynamic queries. We adopted the Average Precision (AP) used in MSCOCO [415] as the metric to evaluate the models' performance in detection and segmentation.

Detection. DiffRegFormer achieves an Average Precision (AP) of 52.1% for detection with a ResNet-50 backbone, thereby surpassing competitors such as QueryInst, Mask R-CNN, and Cascade Mask R-CNN by significant margins. We observe that iterative refinement improves detection performance for medium- to large-sized objects while degrading performance for small objects. It suggests additional operations to mitigate information loss for small objects during the refinement process. Furthermore, static queries are less effective for handling dense objects, particularly when the number of queries significantly exceeds the number of feature channels (300 or 500 compared to

5.4. Results

		Bounding Boxes						Instances					
		AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
one-stage	QueryInst_300	37.8	59.3	47.4	13.5	21.5	47.4	38.9	59.4	43.5	8.8	19.0	44.1
	QueryInst_500	30.1	31.5	22.8	0.9	8.7	24.1	32.5	31.9	22.7	0.4	7.8	27.7
two-stage	Mask R-CNN	49.0	68.6	56.2	18.0	24.4	59.2	45.5	69.0	53.5	16.4	21.5	54.4
	CMask-RCNN	49.9	67.3	58.4	7.6	24.1	61.4	45.0	67.8	54.5	6.5	21.6	54.5
our model	DiffRegFormer	52.1	71.1	57.7	15.4	27.4	61.4	46.8	71.6	52.8	14.1	23.6	54.2

Table 5.2: Overall evaluation results. All models use ResNet-50 as a backbone network and have been trained for 40000 iterations on an NVIDIA GeForce RTX 3090 GPU with a batch size of 2. AP₅₀ and AP₇₅ denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively. AP_S, AP_M, and AP_L correspond to AP on small, medium, and large-sized objects, respectively. The best results are highlighted in bold. CMask-RCNN represents Cascade Mask R-CNN.

256), leading to a marked decline in detection performance.

Instance segmentation. DiffRegFormer achieves a 46.8% AP for instance segmentation and thus outperforms its counterparts. Moreover, applying iterative bounding box refinement significantly affects instance segmentation performance. This effect can be attributed to the binary segmentation process within the boxes, where the quality of the instance masks is inherently linked to the recall and precision rates of the bounding boxes.

Table 5.3 presents the performance per object class. Our model demonstrates an overall superiority in detecting all anatomical structures compared to other methods. Additionally, it performs better in glomeruli and arterial instance segmentation, while maintaining a slight advantage in tubular segmentation. These findings indicate that our model can process dense, multi-class objects with diverse scales at the ROI level. However, for each model, we observe a performance bottleneck in processing arteries, limiting clinical application due to significant size variation within the dataset. For instance, the largest arteries are two orders of magnitude larger than the smallest. The difficulty suggests that more sophisticated attention mechanisms are needed to capture global contextual information in the face of these extreme scale variations.

Figure 5.9 demonstrates DiffRegFormer’s capability to detect objects at various scales effectively. It accurately generates instance masks within bounding boxes, demonstrating precise localization of larger arteries (see the second row). In contrast, Cascade Mask R-CNN struggles to localize large objects within a single bounding box because it lacks an attention mechanism to capture long-range dependencies. Furthermore, Mask R-CNN tends to produce multiple detection results for a single object (see the third row) because it lacks cascade refinement. These observations suggest that iterative bounding box refinement during training significantly reduces multiple predictions per object. At the same time, the attention mechanism with

Chapter 5. Instance Segmentation

		Bounding Boxes - Glomeruli						Instances - Glomeruli					
		AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
one-stage	QueryInst_300	62.6	74.5	71.6	-	-	63.7	56.1	74.5	68.1	-	-	57.1
	QueryInst_500	36.7	47.5	44.0	-	-	37.6	35.9	47.6	43.6	-	-	37.3
two-stage	Mask R-CNN	69.4	85.8	78.6	-	-	70.4	67.0	85.8	78.8	-	-	68.1
	CMask-RCNN	70.7	84.0	81.7	-	-	72.1	65.1	84.0	79.8	-	-	66.4
our model	DiffRegFormer	80.9	94.2	89.8	-	-	80.9	74.3	94.2	80.9	-	-	74.3
		Bounding Boxes - Arteries						Instances - Arteries					
		AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
one-stage	QueryInst_300	18.8	37.0	19.1	5.2	20.2	20.6	17.5	36.3	14.8	4.5	17.9	19.4
	QueryInst_500	4.5	10.0	2.9	0.5	5.6	4.5	4.7	10.5	3.8	0.1	5.3	5.3
two-stage	Mask R-CNN	22.5	40.3	23.5	20.2	19.4	28.9	20.1	41.4	18.0	20.2	17.4	25.4
	CMask-RCNN	23.0	38.0	26.9	-	17.8	33.3	20.1	39.4	21.9	-	16.9	27.5
our model	DiffRegFormer	25.7	46.9	25.8	13.5	23.9	31.6	23.4	48.8	20.2	20.2	20.4	30.1
		Bounding Boxes - Tubuli						Instances - Tubuli					
		AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
one-stage	QueryInst_300	45.8	66.4	51.4	21.8	44.3	57.9	41.2	67.3	47.6	13.1	39.0	55.8
	QueryInst_500	21.2	37.1	21.7	1.3	20.3	30.1	20.6	37.5	20.7	0.7	18.1	40.5
two-stage	Mask R-CNN	57.5	79.6	66.4	15.8	53.8	78.2	51.4	79.9	62.3	12.7	47.3	69.8
	CMask-RCNN	58.1	79.9	66.5	15.1	54.4	78.8	51.8	80.0	61.7	13.0	47.9	69.6
our model	DiffRegFormer	58.5	80.2	67.9	19.9	51.0	75.4	55.8	80.6	62.3	16.6	45.6	64.5

Table 5.3: Per-class evaluation results. All models use ResNet-50 as the backbone network and have been trained for 40000 iterations on a single NVIDIA GeForce RTX 3090 GPU with a batch size of 2. AP₅₀ and AP₇₅ denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively. AP_S, AP_M, and AP_L correspond to mAP on small, medium, and large-sized objects, respectively. The best results are highlighted in bold. CMask-RCNN represents Cascade Mask R-CNN.

dynamic queries processes large-scale objects. Nevertheless, QueryInst, with its reliance on static queries, may struggle to effectively detect all instances in datasets with significant size and shape variations. QueryInst-300 tends to miss many structures, while QueryInst-500 can even degenerate into generating multiple false predictions for a single object. Furthermore, RCNN-based models share a limitation: a single pixel may be allocated to multiple instance masks due to pixel-wise predictions within each independent ROI. It inherently prevents enforcing the spatial exclusivity of pixels to instances. Additionally, Figure 5.10 shows DiffRegFormer’s potential for domain transfer, generating reasonable results on PAS-stained images despite being trained exclusively on Jones’ silver-stained images. Moreover, Figure 5.11 highlights that DiffRegFormer, combined with sliding window and stitching techniques, can be effectively applied to WSIs, further demonstrating its effectiveness.

Lesion classification. The data augmentation pipeline for training the lesion classifier is relatively straightforward. We set the batch size to 96 for the input croppings, with 32 patches per class. Each crop is first subjected to *RandomFlip* with probability 0.5, followed by resizing to 256×256 pixels. Finally, pixel values are normalized to the $(-1, 1)$ range. As shown in Table 5.4, the classifier achieves an overall precision of 89.2% and a recall of 64.6%. More specifically, per-class precision for atrophic tubuli

5.4. Results

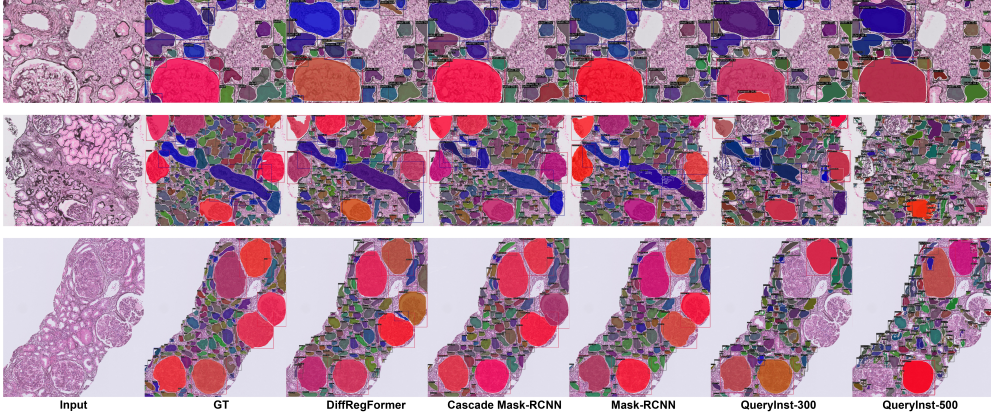


Figure 5.9: A visual comparison of different instance segment methods on Jones-stained kidney biopsies. It shows that our model, i.e., DiffRegFormer, outperforms in detecting objects of diverse sizes and shapes. Despite some instances overlapping, it generates precise instance masks over each region. Mask R-CNN based models struggle to localize a single large object within a single bounding box because they lack mechanisms to capture long-range dependencies. QueryInst with static queries fails to process dense objects since it fails to detect objects of each class within one ROI.

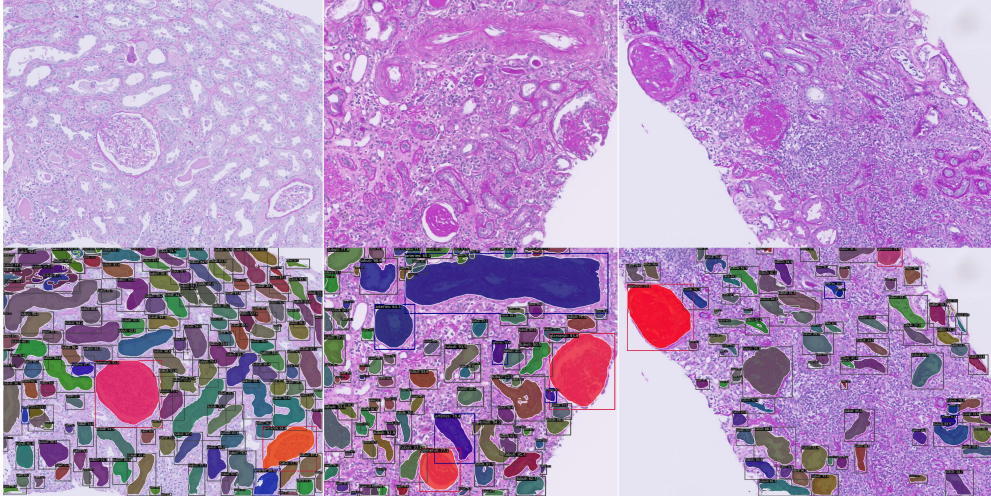


Figure 5.10: A visualization of the results of our model (DiffRegFormer) on the PAS-stained kidney biopsies with increasing degrees of fibrosis and inflammation (from left to right), along with variations in stain intensity.

is 85.1%, and 93.2% for sclerotic glomeruli. In addition, per-class recall is 40.5% for atrophic tubuli and 88.7% for sclerotic glomeruli. Since the lesion classifier operates on anatomical structure croppings generated by DiffRegFormer during inference, its recall

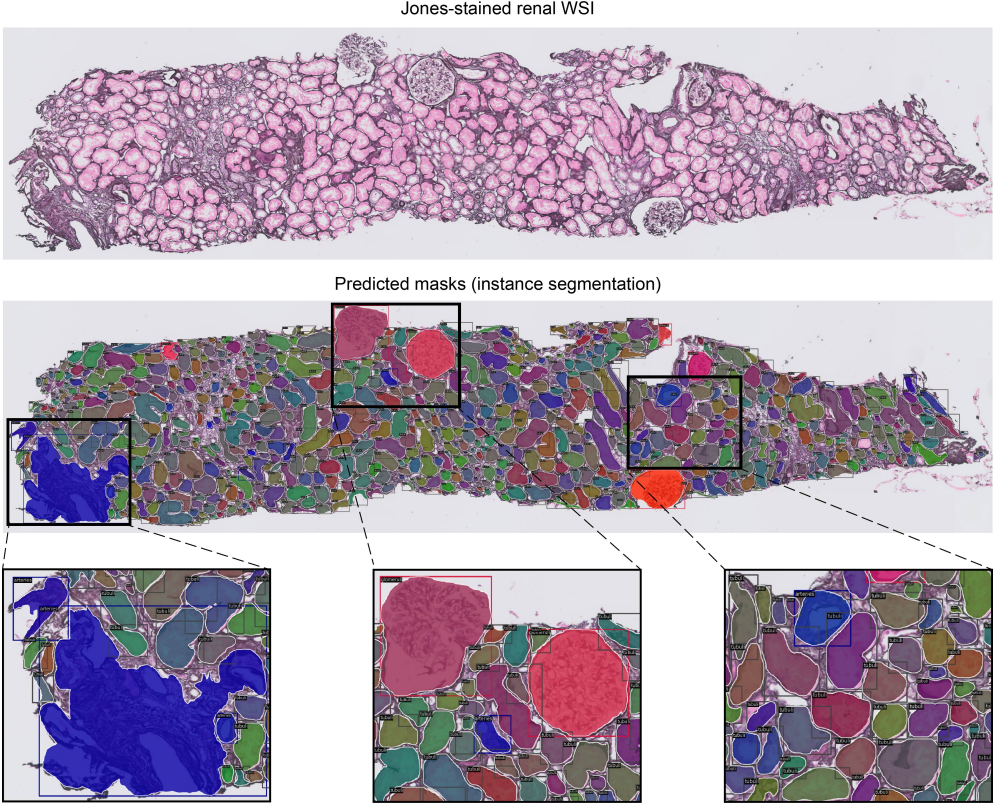


Figure 5.11: A representative visualization of results from DiffRegFormer stitches showing predicted dense instance objects from a WSI of a Jones-stained renal biopsy.

rate heavily depends on the performance of our dense instance segmentation model. However, the high precision indicates that lesion classification on correctly cropped structures is very accurate. In summary, improving the performance of DiffRegFormer is expected to further improve the lesion classifier’s recall. Figure 5.12 visualizes the mapping of lesion predictions onto the instance segmentation results. It reveals that our model can serve as a solid foundation for further clinical diagnosis.

Metrics for Instance Segmentation

We have computed a set of commonly used metrics to assess the performance of our method in a broader scope. These are DICE, MASD (mean average surface distance), FLOPS, inference speed, and the number of parameters. In Table 5.5, we observe that our method outperforms other approaches in segmentation performance, as indicated by the DICE and MASD values. Resource efficiency, however, is moderate. The

5.4. Results

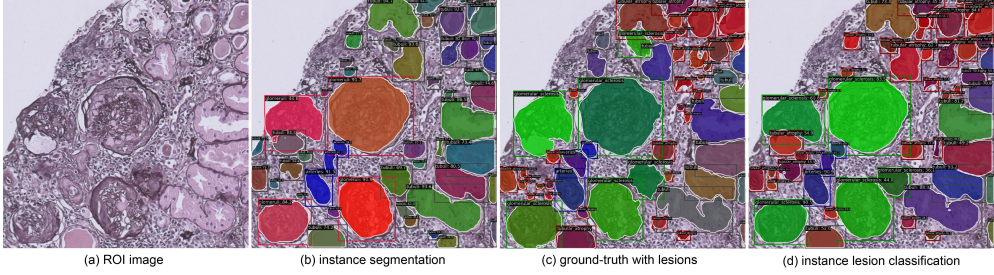


Figure 5.12: A visualization of the results of the lesion classifier predicting the probability of class-wise lesions on an ROI. Combined with dense instance segmentation, our model can generate per-instance lesion identification.

Experimental Settings	Tubuli Lesion Classification		Glomeruli Lesion Classification	
	Normal	Atrophy	Normal	Sclerosis
Train Dataset	10326	10647	534	382
Eval Dataset	906	568	51	22
Per-class Precision	85.1		93.2	
Per-class Recall	40.5		88.7	
Overall Precision	89.2			
Overall Recall	64.6			

Table 5.4: Lesion dataset configuration and classification Performance.

main cause of resource cost is the attention mechanism, as seen in QueryInst and DiffRegFormer. To that end, it is beneficial to adopt more efficient attention algorithms to accelerate inference speed and reduce resource burden, tailored to the practical clinical agenda. In addition, the varying arteries across multiple scales make it difficult to obtain high-quality instance masks. All methods have illustrated a large average surface distance between the predicted and ground-truth boundaries.

Comparison with Segmentation Based Method

To supplement our analysis in terms of completeness, we have conducted a retrospective analysis that reproduces a semantic segmentation based instance segmentation framework [253]. We denote this method as a semantic method and compare its results

		MASD	DICE	FLOPS (Trillion)	Inference Speed (Second)	# Parameters (Million)
One-Stage	QueryInst_300	7.49	42.6	0.41	1.425	173
	QueryInst_500	10.95	35.9	0.55	1.647	246
Two-Stage	Mask R-CNN	4.78	50.7	0.20	0.864	43.98
	CMask-RCNN	3.87	52.1	1.72	1.156	77.03
Our Model	DiffRegFormer	3.64	56.6	0.55	1.324	140

Table 5.5: Additional metrics for instance segmentation assessment.

	Bounding Boxes			Instance			Extra Metrics				
	AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	DICE	MASD	FLOPS (Trillion)	Inference Speed (Second)	# Parameter (Million)
DiffRegFormer	52.1	71.1	57.7	46.8	71.6	52.8	56.6	3.87	0.55	1.324	140
Semantic Method	44.6	64.8	46.7	40.3	65.3	44.6	45.4	7.86	1.24	0.760	32.8

Table 5.6: Overall evaluation result. The semantic method is an instance segmentation framework that combines semantic segmentation with a post-processing operation. We have reproduced it following the literature [253]. *AP* denotes average precision. *AP₅₀* and *AP₇₅* denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively.

with ours. In Table 5.6, an overall comparison is depicted. Our method outperforms the semantic method in prediction quality at the cost of increased inference time and computational resources. It should be noted that these results are specific to this dataset. In Table 5.7, the superiority of our method per class is depicted. Furthermore, Figure 5.13 illustrates that the semantic method is more vulnerable to the segmentation quality of the auxiliary boundary and interstitium class, which is insufficient in separating objects that are tightly touching. Such touching objects are typical in kidney tissue. Our method, on the contrary, can detect and segment each object.

	Bounding Boxes - glo			Instance - glo			Extra Metrics	
	AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	DICE	MASD
DiffRegFormer	80.9	94.2	89.8	74.3	94.2	80.9	78.8	1.21
semantic method	75.6	82.4	81.2	68.3	84.3	70.1	70.6	8.68
	Bounding Boxes - art			Instance - art			Extra Metrics	
	AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	DICE	MASD
DiffRegFormer	25.7	46.9	25.8	23.4	48.8	20.2	30.2	25.94
semantic method	10.3	25.8	13.6	11.5	24.9	10.3	16.3	49.14
	Bounding Boxes - tub			Instance - tub			Extra Metrics	
	AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅	DICE	MASD
DiffRegFormer	58.5	80.2	67.9	55.8	80.6	62.3	64.8	0.89
semantic method	43.6	73.5	54.6	46.9	75.2	52.4	50.5	2.16

Table 5.7: Evaluation results per object class. *AP* denotes average precision. *AP₅₀* and *AP₇₅* denote AP with fixed thresholds for IoU ratio 50% and 75%, respectively.

5.4.3 Complexity Assessment of Model Training

We start by analyzing the intrinsic class imbalance within our dataset. Of the total structures (21,889 objects), 88% are tubuli, followed by 7% arteries (1679 objects), and 5% glomeruli (1226 objects). This distribution is depicted in Figure 5.14 (a), highlighting the necessity of sampling to include more rare instances, such as arteries and glomeruli, among the dominant tubuli. A violin plot, shown in Figure 5.14 (b), details the distribution of structure counts at the ROI level. On average, ROIs contain

5.4. Results

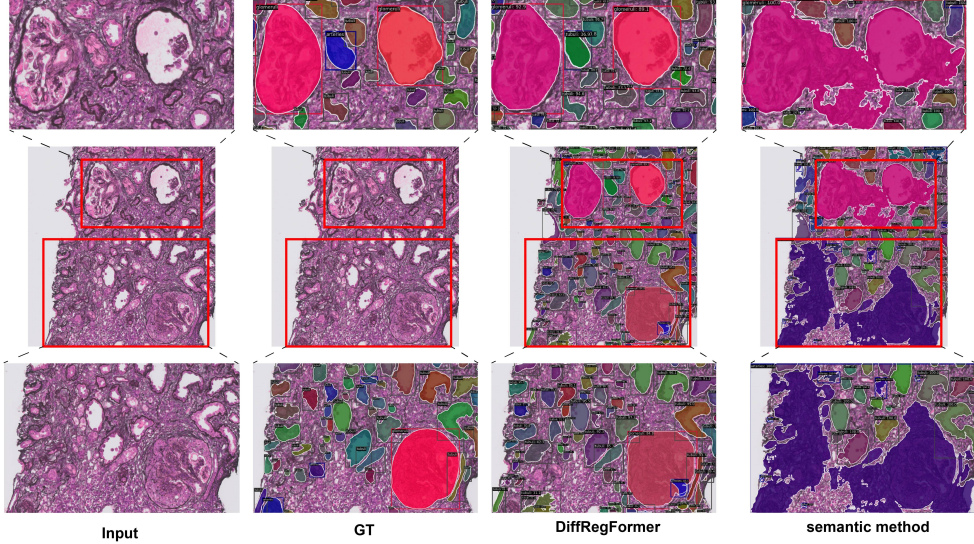


Figure 5.13: A visual comparison between DiffRegFormer and the semantic method. The upper row represents a zoom of the upper-box in the middle row. The lower row represents a zoom of the lower-box in the middle row. The quality of the result w.r.t. the ground truth labels can be convincingly assessed in favor of DiffRegFormer.

95 objects, with a maximum of 434. To accommodate all objects, DiffRegFormer is configured to handle more than 434 per image. Figure 5.14 (c) depicts the correlation between the increase in computational cost and performance improvement. As expected, we observe a turning point at 500 dynamic queries. This configuration yields a favorable trade-off between complexity and efficiency, achieving 52.1% AP with 20.2 GB of GPU memory. Specifically, ROIs contain an average of 5 glomeruli, 7 arteries, and 84 tubuli with corresponding maximums of 74, 84, and 408. Setting the maximum number per class to 166 is another feasible option for positive sample generation, as our sampling approach ensures comprehensive inclusion of rare instances while maintaining the rational proportion of the majority class. In conclusion, selecting a maximum of 500 objects per image and 166 per class can effectively balance performance and computational efficiency.

5.4.4 Ablation Study

The ablation study focuses on the effects of various training strategies and query types on model performance. By analyzing our results, we can gain insights into the specific characteristics of each strategy. This analysis highlights the significance of our

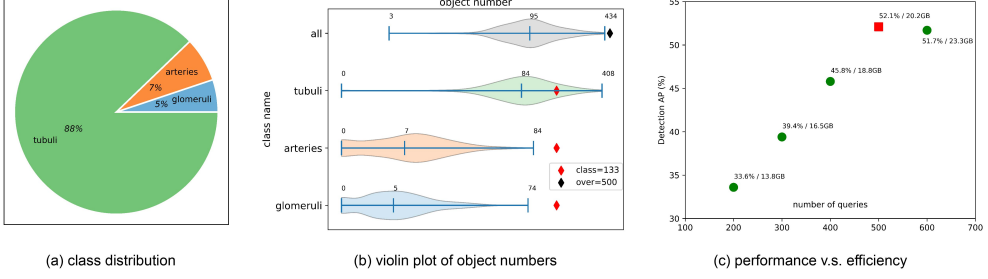


Figure 5.14: Complexity Analysis on Jones-stained renal biopsy dataset. Figure (a) is the sample occupation across our dataset. Figure (b) shows a violin plot for the sample number distribution per class. Figure (c) shows trade-offs between performance and efficiency for different experiment settings. The red square is the optimal setting, while the green circles are sub-optimal settings.

improvements as key contributions to adapting an end-to-end instance segmentation model to datasets with dense objects.

Queries

Static queries in transformer-based models struggle to represent dense objects in large-scale datasets due to their static nature. Adequate accommodation to extensive variations in object size and shape necessitates a proportional increase in query count. Alternatively, dynamic queries, generated on-the-fly for each object in a mini-batch, demonstrate superior adaptability to dataset complexity. Theoretically, assigning more static queries could achieve performance comparable to that of dynamic queries. However, our experiments indicate that increasing the number of static queries without a corresponding increase in feature channels does not yield improvements. Consequently, employing static queries for dense object processing in transformer-based models is significantly less cost-effective than dynamic queries.

Table 5.8 delineates our comparative analysis of varying query counts for both dynamic and static queries, with the number of feature channels fixed at 256. We observe that increasing the number of dynamic queries improves performance, whereas increasing the number of static queries degrades performance. Figure 5.15 further illustrates the inherent challenge of achieving a favorable trade-off between complexity and performance with static queries. Even with 500 queries, the model with static queries fails to consistently detect structures across all classes. This limitation is not attributable to class imbalance compared to the dynamic query strategy. In conclusion, the dynamic query is a more efficient approach for datasets with dense

5.4. Results

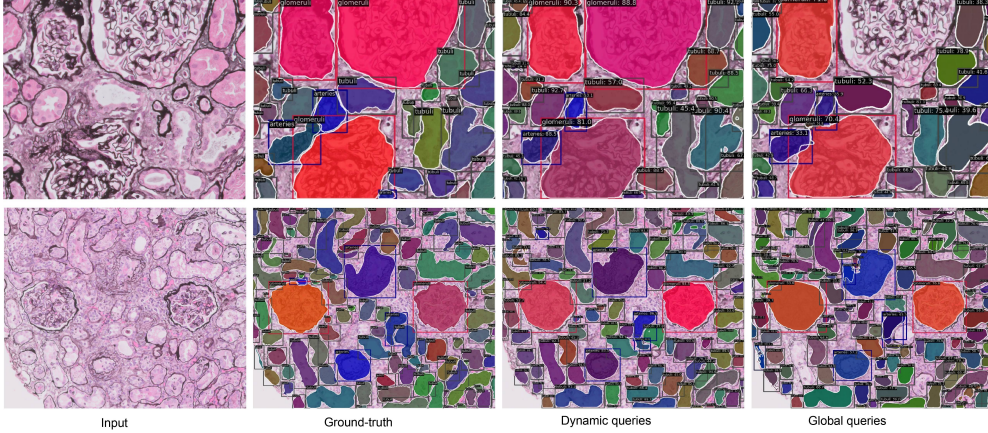


Figure 5.15: A visual comparison from an ablation study on different query types. Dynamic queries outperform static queries on detection. It should be noted that the static queries fail to detect objects of all classes within one ROI.

objects, highlighting their superiority over static queries.

Number mAP	# 200 Query	# 300 Query	# 400 Query	# 500 Query	# 600 Query
Dynamic Query	33.6	39.4	45.8	52.1	51.7
Static Query	23.8	38.9	37.4	34.6	31.6

Table 5.8: Evaluation of detection performances on different types of queries. The best result of each type is highlighted in bold.

Training strategies

feature	boxes	sample	glo-det	art-det	tub-det	all-det	glo-ins	art-ins	tub-ins	all-ins
✗	✗	✗	-	-	-	-	-	-	-	-
✗	✗	✓	-	-	-	-	-	-	-	-
✗	✓	✗	75.8	24.9	56.4	52.4	-	-	33.6	-
✗	✓	✓	76.1	26.2	57.6	53.3	7.3	4.8	35.2	15.8
✓	✗	✗	-	-	-	-	-	-	-	-
✓	✗	✓	8.5	4.3	13.5	8.8	6.8	4.1	12.5	7.8
✓	✓	✗	12.6	6.8	56.7	25.4	10.4	5.3	45.8	20.5
✓	✓	✓	80.9	25.7	58.5	55.1	74.3	23.4	55.8	51.2

Table 5.9: Evaluation on combinations of different strategies. The symbol ✓ denotes taking one strategy, while the symbol ✗ means taking the converse strategy. The best results are highlighted in bold. *glo*: glomeruli. *art*: arteries. *tub*: tubuli. *all*: overall. *det*: detection. *ins*: instance. "-": not working. *feature*: if use separate feature maps between bbox-decoder and mask-decoder. *boxes*: if use ground-truth boxes in the sampling process. *sample*: if using class balance in the sampling process.

Training an instance segmentation model on datasets with dense objects, such as

kidney biopsies, presents distinct challenges compared to common datasets with sparser object distributions, such as MSCOCO. Instance segmentation models typically employ three strategies: (1) shared features between the bbox-decoder and mask-decoder; (2) selection of positive samples from proposal bounding boxes; and (3) class-imbalanced sampling. However, for dense object segmentation, we propose three alternative strategies: (a) separate features for the bbox and mask decoders; (b) positive sample selection directly from ground-truth boxes; and (c) class-balanced sampling. While the conventional combination of shared features, proposal boxes, and biased sampling is adequate for standard instance segmentation (see Figure 5.16a), only the combination of separate features, ground-truth boxes, and unbiased sampling proves to be successful for dense instance segmentation. Table 5.9 evaluates the combinations above, with conventional strategies marked with ✗ and our proposed strategy with ✓. Notably, only five out of eight possible combinations can train DiffRegFormer.

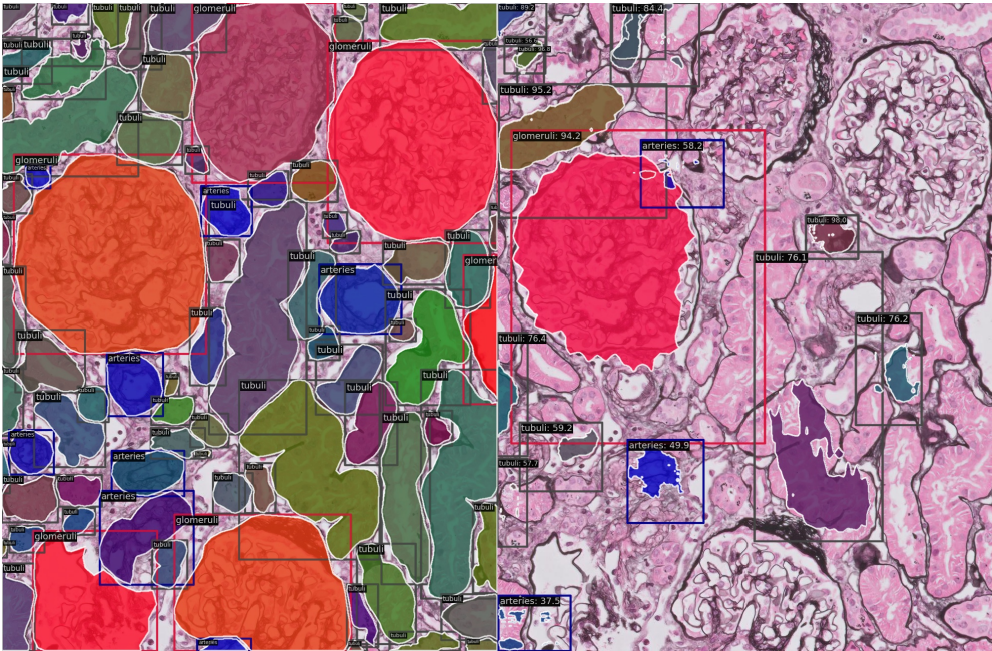
Figure 5.16b demonstrates that combining shared features, ground-truth boxes, and biased sampling enables object detection across all classes. However, it fails to produce accurate instance masks within these bounding boxes. The observation suggests that the bbox-decoder dominates the shared feature representation, thus hindering the mask decoder’s ability to extract pixel-wise information. Consequently, all pixels within detected boxes are classified as tubuli. Moreover, biased sampling impairs the mask decoder’s ability to learn features for rare instances (glomeruli and arteries), leading to no pixel-wise classification within those boxes.

Additionally, Figure 5.16c shows that shared features combined with ground-truth boxes and unbiased sampling can detect and segment objects across all classes, albeit with degraded mask quality. This degradation is likely due to the bbox-decoder’s dominance within the shared feature representation.

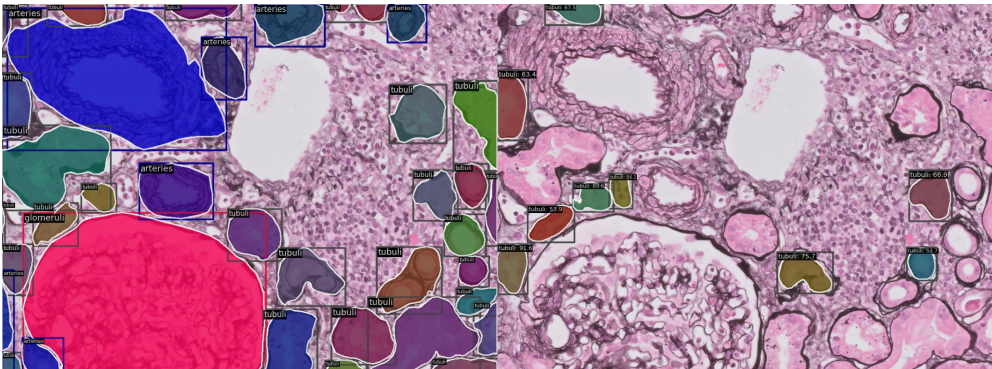
In contrast, Figure 5.16d reveals that separate features, proposal boxes, and unbiased sampling lead to inaccurate object detection, probably due to the accumulation of errors introduced by proposals during the early stages of diffusion model training.

Lastly, Figure 5.16e demonstrates that separate features, ground-truth boxes, and biased sampling result in correct detection and segmentation for only a small proportion of tubuli. This finding further highlights the challenges caused by biased sampling in learning features for rare instances.

In summary, shared feature maps lead the bbox-decoder to dominate feature representation learning, impairing the mask-decoder’s ability to extract pixel-wise information necessary for accurate binary instance masks. Biased sampling further hinders the mask-decoder from learning information to segment rare instances (glomeruli



(d) separate feature + proposals + unbiased sampling on a ROI



(e) separate feature + ground-truth + biased sampling on a ROI

Figure 5.16: Ablation study on combinations of different strategies. The images on the left are ground-truth. The images on the right are our predictions.

and arteries), causing it to be overwhelmed by the majority class (tubuli). Finally, using proposals for sampling rather than ground-truth boxes introduces cumulative errors that disrupt training for both decoders, preventing the learning of contextual information necessary for long-range correlations via attention mechanisms.

5.5 Conclusion and Future Work

Our research introduces a novel framework for processing large-scale renal WSI datasets that accounts for potential changes in lesion combinations. Our model performs better in predicting lesions within dense objects, effectively handling multi-class and multi-scale challenges at the ROI level. We present the first dense instance segmentation module, DiffRegFormer, that seamlessly integrates a diffusion model with a transformer in an RCNN-style architecture. This approach leverages the diffusion model to generate refined object boxes, upon which a regional transformer creates dynamic queries that are ultimately transformed into accurate instance masks within the boxes. Furthermore, our lesion classifier accurately predicts class-specific lesions for each cropped structure. The model demonstrates several advantages in handling dense object datasets: (1) its iterative noise-to-box approach eliminates the dependence on prior knowledge about object size and shape, facilitating rapid object localization; (2) dynamic queries, generated on-the-fly for each object in a mini-batch style, ensuring scalability to large datasets; (3) the modular design of the model, with independent plug-in prediction heads, allows for replacement and promotes adaptability. This flexibility enables the efficient reuse of trained components when upscaling to large-scale datasets. Experimental results on Jones' stained renal biopsies demonstrate that our model surpasses existing benchmarks. Comparing results between methods based on the data is tricky. Measures like DICE behave differently with different datasets. The comparison should be based on the same dataset (cf. 5.4.2). In that case, results can be well understood and objectively compared.

The current model needs further exploration to meet the rigorous requirements for clinical applications. Still, it establishes a robust framework for dense, multi-class, multi-scale object recognition at the ROI level, setting a solid foundation for future computational improvements. Potential areas for future research include: (1) further enhancing the compatibility of diffusion models with transformers like DETR to avoid NMS post-processing. (2) developing innovative regional feature extractors to generate more informative dynamic queries and mitigate the loss of information for small objects during iterative refinements. (3) integrating spatial constraints into regional mask inference to prevent erroneous assignment of single pixels to multiple instances. (4) designing lightweight attention mechanisms to lessen the computational burden. In conclusion, the approach presented paves the way for efficient and reliable automated assessment as a tool for the workbench of renal pathologists.

Chapter 6

Conclusions and Discussion

In this thesis, we have addressed six research questions. These questions comprise a survey (Chapter 2) and two application domains: the development of deep learning in the image processing community and a specific focus on classification problems in the computer vision domain. The themes investigated cover a general review of deep learning variants and narrow to two specific classification cases: image segmentation and object detection. In this chapter, we summarize the main findings of our research. Additionally, we discuss and identify the challenges posed by our methods and possible solutions to address them. Lastly, we suggest potential developments for future work.

6.1 Main Findings

In each chapter, we have proposed new views and approaches to address the corresponding research questions. Here, we summarize these answers and present the main findings inspired by experimental results.

Chapter 2 answers **RQ1: What is the current development in deep learning models and their applications in image processing?** **Chapter 2** provides a new perspective to track the development of deep learning models from CNN via Transformers to the hybrid models, based on the extension and improvement of the mathematical definitions of their core components: convolution and attention mechanisms. Next, we elaborate on the fundamental elements of the typical architectures and list other stand-alone methodologies that can help deep learning models in general. Subsequently, a broad range of active research topics in the image processing domain

6.1. Main Findings

has been highlighted, each with a revolutionized cornerstone application. Lastly, we are interested in the limitations and challenges of modern deep learning architectures and in providing direction for future research.

Chapter 3 takes the results from **Chapter 2**, and narrows down our research topic on one specific case of image classification: semantic segmentation. In this chapter, we offer a novel perspective to understand the cross-attention mechanism in the context of semantic segmentation. The research question **RQ2: How can we explain an attention mechanism as a probabilistic model that captures global cross-class correlations in semantic segmentation tasks?** is answered by our Variational Gaussian Mixture Model. The framework interprets cross-attention as building a multi-modal distribution via a Gaussian mixture and supports online updates (i.e., mini-batch style) via Bayesian update rules. Specifically, each semantic concept is a centroid of a Gaussian distribution. It measures its uncertainty using correlation matrices and gathers global class-specific information from the input feature maps in the dataset. However, images with complex interactions among semantic concepts yield ambiguous decision boundaries between classes. In general, a solution can be to integrate more geometrical or structural constraints. Adding constraints has led to the formulation of an additional research question **RQ3: How can we integrate geometric constraints into a probabilistic model and solve it efficiently?** To answer this, we have extended our Gaussian Mixture Model to support arbitrary additive geometric regularization. For simplicity without losing generalizability, we choose the popular Total-Variance constraints and solve them with a level-set algorithm for topological flexibility. Based on our experimental analysis, we conclude that our probabilistic model is resource-friendly and scalable as the number of semantic classes increases.

Chapter 4 further explores the cross-attention mechanism in semantic segmentation under different conditions and constraints. In this chapter, we focus on generating high-quality dense annotations without providing explicit pixel-wise supervision. The so-called weakly supervised semantic segmentation (WSSS) is a very important task, as annotation can be expensive, especially in expertise-intensive fields such as biology and medicine. Although cross-attention is a powerful tool for extracting long-range dependencies, WSSS suffers from an intrinsic deficit: incomplete information cannot infer the correct full class assignment. Therefore, we propose a conditional probabilistic model to separate the foreground segmentation from the missing background information. To answer **RQ4: How can we interpret weakly supervised semantic segmentation**

as a feature space decomposition problem using conditional probabilistic models? We adopt contrastive-learning algorithms to infer background information across the whole dataset and interpret it as out-of-distribution information. Then, foreground assignment only focuses on likely supplementary regions, subtracting the background. Furthermore, we have observed that loss functions in WSSS exhibit equipotential sets on their loss surfaces, leading to perplexity in distinguishing good from bad assignment configurations. This phenomenon has given rise to an additional research question **RQ5: How can we eliminate the equipotential of the loss surface in weakly supervised semantic segmentation using geometric regularization?** Our solution is to add geometric regularization, like Total-Variance constraints, to provide extra evidence for WSSS loss functions to reject the bad assignments. Our implementation of the conditional model uses a masked cross-attention with a level-set method. Through our experiments, we have achieved competitive results on a typical dataset and demonstrated an explainable prototype transformation in the latent feature space.

Chapter 5 focuses on another special case of image classification: instance segmentation/object detection on a biomedical dataset. We find that kidney pathology images are too complicated to get acceptable performance by simply adapting existing algorithms. The crux is that each image contains a large number (about 500) of anatomical structures, spanning a wide range of scales and tightly packed together. To exacerbate the situation, there are multiple anatomical types, but with a severe class imbalance. Therefore, instance segmentation/object detection for Kidney pathology images requires flexibility to handle varying scales and efficiency to accommodate dense objects. It is promising to integrate the diffusion model and attention mechanism into RCNN with three-fold advantages: (1) generative methods benefit from generating various scaled candidates; (2) attention can capture long-range correlation to accurately locate large objects; (3) RCNN can only focus on potential regions and ignore the irrelevant background. From this analysis, we formulated the last research question **RQ6: How can we equip an attention mechanism with an R-CNN in a diffusion manner in support of efficiency and generalization of dense instance segmentation tasks?** Our solution is to design a dynamic attention mechanism combining the diffusion models and RCNN. Specifically, each dynamic attention module takes as input noisy box candidates generated by diffusion models, computes global interactions among them, and outputs refined boxes for RCNN. In order to avoid sampling imbalance in the RCNN, we have also designed a class-balanced sampling algorithm to

6.2. Discussion

stabilize training. The experiments with our framework and the subsequent analysis demonstrate that it has domain transferability (i.e., the ability to ignore the effect of stains), thereby demonstrating flexibility, generalizability, and efficiency.

6.2 Discussion

The methodologies and frameworks that we have proposed in this thesis are addressed in six research questions and have achieved promising results. They, however, still have some challenges with respect to algorithmic and theoretical perspectives.

6.2.1 Algorithmic Perspective

(1) Amortized Geometrical Constraint: In Chapter 3, the Total Variation regularization is a computational bottleneck. Since it lacks a closed-form solution, we must resort to iterative optimization via level-set methods. That means it is necessary to evolve the class-wise decision boundaries from the initial surfaces for each input feature map. In general, this process requires more than 15 iterations to converge from scratch, resulting in unexpected computational time. For efficiency, we adopt a compromise solution that requires an additional sub-network to predict a better initial surface. The refined surface can accelerate the convergence of the level-set algorithm but increase the difficulty of the training process, requiring more sophisticated hyperparameter tuning, such as the learning rate.

(2) Dynamic regional candidate: In Chapter 5, we adopt RoI-align Pooling to generate the initial region candidates as the input dynamic queries. Although it is a classical operator used in RCNN models, it assumes even sampling within each candidate region, thereby ignoring differences in scale and geometric irregularity. This intrinsic drawback significantly degrades object representations, especially for extremely small or large objects. Even worse, RoI-align Pooling fails to account for spatial exclusivity among candidates because it runs each candidate in parallel. A possible solution to this is to design a box attention that captures non-local information within each region for completeness and interacts between candidates to ensure clear spatial affiliation.

6.2.2 Theoretical Perspective

Variational Bayesian Inference: In Chapter 3, Variational Bayesian Inference (VB) is a family of graphical models for approximating intractable distributions arising in Bayesian inference. VB is an extension of the expectation-maximization (EM) algorithm and approximates the true posterior distribution as a factorization of simple distributions. By optimizing the evidence lower bound (ELBO), VB alternatively approximates the posterior distribution and updates the parameters. Although VB is scalable for inferring complex models with a large number of parameters, it has an intrinsic drawback: finding a suitable factorization often requires many trials. Therefore, we cannot actually estimate our factorization outperforming the others. The deficiency means that performance comparison of different factorizations in the same training setting should be considered to provide experimental evidence of selection over factorizations.

6.3 Future Work

In previous chapters, we have presented many methods for addressing research questions in the image processing community related to deep learning models. Deep learning models need novel ideas to evolve further. Therefore, it is crucial to discuss potential advancements in future work. These specifically concern interpretability, data efficiency, multi-modal understanding, and scalability. In this section, we briefly indicate possible future research directions.

- (1) Deep learning models demand more effort to enhance their transparency. It is crucial to build trustworthy systems, especially in reliability-sensitive applications. Researchers are currently actively developing techniques in explainable AI (XAI). There are two promising directions for increasing human supervision of complex models. One direction is to develop more convenient visualization systems that can easily assess the responses/activations/graduations of arbitrary components to support decision-making. The other is the adaptation of the traditional algorithms or probabilistic models into differentiable modules, interpreting local components with solid mathematical formulations.
- (2) Deep learning models require vast amounts of high-quality annotated data. However, it is very expensive to generate trustworthy supervision information, especially for

6.3. Future Work

expertise-intensive fields. Therefore, advancements in areas such as transfer learning, few-shot learning, self-supervised learning, weakly supervised learning, and unsupervised learning can effectively address the data-hungry problem, enabling them to learn effectively with less data. Stand-alone techniques open up new possibilities for extending existing models to new applications where large labeled datasets are scarce.

(3) Deep learning models intend to solve increasingly complex problems that need to integrate information from multiple sources, such as text, images, audio, etc. Therefore, multiple-modal algorithms have the potential to facilitate more sophisticated reasoning on complex systems, such as autonomous systems, virtual assistants, and medical diagnostics. In addition, multiple-modal applications can revolutionize human-computer interaction and create more immersive experiences.

(4) The scalability of deep learning is advancing with better hardware, optimized algorithms, and distributed computing frameworks. This progress enables training on massive datasets and complex architectures, leading to the development of more sophisticated models with improved performance and generalization. Meanwhile, scaling deep learning models down to portable devices is enabled with cloud computing and high-performance processors. The advancement makes them more accessible, helping leverage AI to tackle real-world problems on mobile devices.

We have investigated a broad range of aspects of deep learning models in the context of image processing. We have been covering both technical applications and mathematical interpretability. In the rapidly evolving field of artificial intelligence, inevitably, our coverage is not comprehensive. Nevertheless, our contributions move the field forward and yield incremental insights and knowledge. We are pleased that our research adds to these increments and look forward to the further development of the field.

Bibliography

- [1] Nima Aghli and Eraldo Ribeiro. Combining Weight Pruning and Knowledge Distillation for CNN Compression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3191–3198, 2021.
- [2] Karim Ahmed and Lorenzo Torresani. Star-Caps: Capsule Networks with Straight-Through Attentive Routing. *Advances in Neural Information Processing Systems*, 32, 2019.
- [3] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly Supervised Learning of Instance Segmentation with Inter-Pixel Relations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019.
- [4] Namhyuk Ahn, Byungkoon Kang, and Kyung-Ah Sohn. Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. In *European Conference on Computer Vision*, pages 252–268, 2018.
- [5] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyang Wu. Spatio-Temporal Representation Factorization for Video-Based Person Re-Identification. In *IEEE International Conference on Computer Vision*, pages 152–162, 2021.
- [6] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. XCiT: Cross-Covariance Image Transformers. *Advances in Neural Information Processing Systems*, 34:20014–20027, 2021.
- [7] Zeyuan Allen-Zhu and Yuanzhi Li. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. *ArXiv Preprint ArXiv:2012.09816*, 2020.
- [8] Zeyuan Allen-Zhu and Yuanzhi Li. Forward Super-Resolution: How Can GANs Learn Hierarchical Generative Models for Real-World Distributions. *ArXiv Preprint ArXiv:2106.02619*, 2021.
- [9] Israa Alnazer, Pascal Bourdon, Thierry Urruty, Omar Falou, Mohamad Khalil, Ahmad Shahin, and Christine Fernandez-Maloigne. Recent Advances in Medical

Bibliography

- Image Processing for the Evaluation of Chronic Kidney Disease. *Medical Image Analysis*, 69:101960, 2021.
- [10] Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated Feedback Refinement Network for Dense Image Labeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3751–3759, 2017.
- [11] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. Structured Pruning of Deep Convolutional Neural Networks. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 13(3):1–18, 2017.
- [12] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale Combinatorial Grouping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 328–335, 2014.
- [13] Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble Knowledge Distillation for Learning Improved and Efficient Networks. In *ECAI 2020*, pages 953–960. IOS Press, 2020.
- [14] Marcella Astrid and Seung-Ik Lee. CP-Decomposition with Tensor Power Method for Convolutional Neural Networks Compression. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 115–118. IEEE, 2017.
- [15] AR Aswani, R Chithra, and Alex Pappachen James. Unstructured Weight Pruning in Variability-Aware Memristive Crossbar Neural Networks. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 3458–3462. IEEE, 2022.
- [16] Sara Atito, Muhammad Awais, and Josef Kittler. SIT: Self-Supervised Vision Transformer. *ArXiv Preprint ArXiv:2104.03602*, 2021.
- [17] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025.
- [18] Jimmy Ba and Rich Caruana. Do Deep Nets Really Need to Be Deep? *Advances in Neural Information Processing Systems*, 27, 2014.
- [19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *ArXiv Preprint ArXiv:1607.06450*, 2016.
- [20] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *ArXiv Preprint ArXiv:1511.00561*, 2015.

-
- [21] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.
 - [22] Jong-Hyeon Baek, Hee Kyung Lee, Hyon-Gon Choo, Soon-heung Jung, and Yeong Jun Koh. Center-Guided Transformer for Panoptic Segmentation. *Electronics*, 12(23), 2023.
 - [23] Alexei Baevski and Michael Auli. Adaptive Input Representations for Neural Language Modeling. *ArXiv Preprint ArXiv:1809.10853*, 2018.
 - [24] Neeraj Baghel, Shiv Ram Dubey, and Satish Kumar Singh. SRTransGAN: Image Super-Resolution Using Transformer Based Generative Adversarial Network. *ArXiv Preprint ArXiv:2312.01999*, 2023.
 - [25] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023.
 - [26] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. *ArXiv Preprint ArXiv:2106.08254*, 2021.
 - [27] Qiqi Bao, Yunmeng Liu, Bowen Gang, Wenming Yang, and Qingmin Liao. SCTANet: A Spatial Attention-Guided CNN-Transformer Aggregation Network for Deep Face Image Super-Resolution. *IEEE Transactions on Multimedia*, 25:8554–8565, 2023.
 - [28] Jonathan T Barron. Squareplus: A Softplus-Like Algebraic Rectifier. *ArXiv Preprint ArXiv:2112.11687*, 2021.
 - [29] Syed Muhammad Arsalan Bashir and Yi Wang. Small Object Detection in Remote Sensing Images with Residual Feature Aggregation-Based Super-Resolution and Object Detector Network. *Remote Sensing*, 13(9):1854, 2021.
 - [30] Josh Beal, Eric Kim, Eric Tzeng, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Toward Transformer-Based Object Detection. *ArXiv Preprint ArXiv:2012.09958*, 2020.
 - [31] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.
 - [32] Irwan Bello. LambdaNetworks: Modeling Long-Range Interactions Without Attention. *ArXiv Preprint ArXiv:2102.08602*, 2021.

Bibliography

- [33] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention Augmented Convolutional Networks. In *IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.
- [34] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *ArXiv Preprint ArXiv:2004.05150*, 2020.
- [35] Gedas Bertasius, Lorenzo Torresani, Stella X Yu, and Jianbo Shi. Convolutional Random Walk Networks for Semantic Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2017.
- [36] Xiao Bian, Ser Nam Lim, and Ning Zhou. Multiscale Fully Convolutional Network with Application to Industrial Inspection. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016.
- [37] Christopher M Bishop and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.
- [38] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv Preprint ArXiv:2004.10934*, 2020.
- [39] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-Time Instance Segmentation. In *IEEE International Conference on Computer Vision*, pages 9157–9166, 2019.
- [40] Y-Lan Boureau, Nicolas Le Roux, Francis Bach, Jean Ponce, and Yann LeCun. Ask the Locals: Multi-Way Local Pooling for Image Recognition. In *IEEE International Conference on Computer Vision*, pages 2651–2658. IEEE, 2011.
- [41] Y-Lan Boureau, Jean Ponce, and Yann LeCun. A Theoretical Analysis of Feature Pooling in Visual Recognition. In *27th International Conference on Machine Learning (ICML-10)*, pages 111–118, 2010.
- [42] Nassim Bouteldja, Barbara M Klinkhammer, Roman D Bülow, Patrick Droste, Simon W Otten, Saskia Freifrau von Stillfried, Julia Moellmann, Susan M Sheehan, Ron Korstanje, Sylvia Menzel, et al. Deep Learning–Based Segmentation and Quantification in Experimental Kidney Histopathology. *Journal of the American Society of Nephrology: JASN*, 32(1):52, 2021.
- [43] Soumeya Brachemi and Guillaume Bollée. Renal Biopsy Practice: What Is the Gold Standard? *World Journal of Nephrology*, 3(4):287, 2014.
- [44] Gianni Brauwers and Flavius Frasinicar. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [45] Antoni Buades, Bartomeu Coll, and J-M Morel. A Non-Local Algorithm for Image Denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 2, pages 60–65. Ieee, 2005.

-
- [46] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model Compression. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541, 2006.
 - [47] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction. In *IEEE International Conference on Computer Vision*, pages 17302–17313, 2023.
 - [48] Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly Kernel Inception Network for Remote Sensing Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 27706–27716, 2024.
 - [49] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into High Quality Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.
 - [50] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2019.
 - [51] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.
 - [52] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. *ArXiv Preprint ArXiv:1904.11492*, 2019.
 - [53] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
 - [54] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.
 - [55] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *IEEE International Conference on Computer Vision*, pages 9650–9660, 2021.
 - [56] Vicent Caselles, Antonin Chambolle, Matteo Novaga, et al. Total Variation in Imaging. *Handbook of Mathematical Methods in Imaging*, 1(2):3, 2015.

Bibliography

- [57] Kelvin CK Chan, Xiangyu Xu, Xintao Wang, Jinwei Gu, and Chen Change Loy. GLEAN: Generative Latent Bank for Image Super-Resolution and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3154–3168, 2022.
- [58] Siddhartha Chandra, Nicolas Usunier, and Iasonas Kokkinos. Dense and Low-Rank Gaussian CRFs Using Deep Embeddings. In *International Conference on Computer Vision*, volume 2, 2017.
- [59] S Chandrakala, K Deepak, and G Revathy. Anomaly Detection in Surveillance Videos: A Thematic Taxonomy of Deep Models, Review and Performance Analysis. *Artificial Intelligence Review*, 56(4):3319–3368, 2023.
- [60] Chia-Yuan Chang, Shuo-En Chang, Pei-Yung Hsiao, and Li-Chen Fu. EPSNet: Efficient Panoptic Segmentation Network with Cross-Layer Attention Fusion. In *Asian Conference on Computer Vision*, 2020.
- [61] Chih-Lin Chang, Shiou-Chi Li, and Jen-Wei Huang. Ensemble2N et: Learning from Ensemble Teacher Networks via Knowledge Transfer. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE, 2021.
- [62] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-Supervised Semantic Segmentation via Sub-Category Exploration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020.
- [63] Abhishek Chaurasia and Eugenio Culurciello. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.
- [64] Boyu Chen, Peixia Li, Chuming Li, Baopu Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, and Wanli Ouyang. GLiT: Neural Architecture Search for Global and Local Image Transformer. In *IEEE International Conference on Computer Vision*, pages 12–21, 2021.
- [65] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. RegionViT: Regional-to-Local Attention for Vision Transformers. *ArXiv Preprint ArXiv:2106.02689*, 2021.
- [66] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *IEEE International Conference on Computer Vision*, pages 357–366, 2021.
- [67] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-Trained Image Processing Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

-
- [68] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.
 - [69] Hao Chen, Chunhua Shen, and Zhi Tian. Unifying Instance and Panoptic Segmentation with Dynamic Rank-1 Convolutions. *ArXiv Preprint ArXiv:2011.09796*, 2020.
 - [70] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8573–8581, 2020.
 - [71] Honggang Chen, Xiaohai He, Linbo Qing, Yuanyuan Wu, Chao Ren, Ray E Sheriff, and Ce Zhu. Real-World Single Image Super-Resolution: A Brief Review. *Information Fusion*, 79:124–145, 2022.
 - [72] Jiaxin Chen, Dong Xing, Mohammad Shabaz, Yongpei Zhu, Yong Wang, and Xianxun Zhu. Dnlx: Image super-resolution with deformable non-local attention and multi-branch weighted feature fusion. *Image and Vision Computing*, 162:105721, 2025.
 - [73] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *ArXiv Preprint ArXiv:2102.04306*, 2021.
 - [74] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid Task Cascade for Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.
 - [75] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *ArXiv Preprint ArXiv:1906.07155*, 2019.
 - [76] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018.
 - [77] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *ArXiv Preprint ArXiv:1412.7062*, 2014.
 - [78] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional

Bibliography

- Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.
- [79] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv Preprint ArXiv:1706.05587*, 2017.
- [80] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018.
- [81] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative Pretraining from Pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020.
- [82] Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. AutoFormer: Searching Transformers for Visual Recognition. In *IEEE International Conference on Computer Vision*, pages 12270–12280, 2021.
- [83] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-Supervised Image-Specific Prototype Exploration for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4288–4298, 2022.
- [84] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.
- [85] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. DiffusionDet: Diffusion Model for Object Detection. In *IEEE International Conference on Computer Vision*, pages 19830–19843, 2023.
- [86] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating Learning via Knowledge Transfer. *ArXiv Preprint ArXiv:1511.05641*, 2015.
- [87] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- [88] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J. Fleet. A Generalist Framework for Panoptic Segmentation of Images and Videos, 2023.
- [89] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2Seq: A Language Modeling Framework for Object Detection. In *International Conference on Learning Representations*, 2021.

-
- [90] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog Bits: Generating Discrete Data Using Diffusion Models with Self-Conditioning. *ArXiv Preprint ArXiv:2208.04202*, 2022.
 - [91] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing Neural Networks with the Hashing Trick. In *International Conference on Machine Learning*, pages 2285–2294. PMLR, 2015.
 - [92] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating More Pixels in Image Super-Resolution Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023.
 - [93] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *ArXiv Preprint ArXiv:2003.04297*, 2020.
 - [94] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. TensorMask: A Foundation for Dense Object Segmentation. In *IEEE International Conference on Computer Vision*, pages 2061–2069, 2019.
 - [95] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *IEEE International Conference on Computer Vision*, pages 9640–9649, 2021.
 - [96] Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, and Han Hu. RepPoints V2: Verification Meets Regression for Object Detection. In *Advances in Neural Information Processing Systems*, 2020.
 - [97] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic Convolution: Attention over Convolution Kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020.
 - [98] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A²-Nets: Double Attention Networks. *Advances in Neural Information Processing Systems*, 31, 2018.
 - [99] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual Path Networks. *Advances in Neural Information Processing Systems*, 30, 2017.
 - [100] Zhaozheng Chen and Qianru Sun. Extracting Class Activation Maps from Non-Discriminative Features as Well. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3135–3144, 2023.
 - [101] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *The Eleventh International Conference on Learning Representations*, 2023.

Bibliography

- [102] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The Vision-Friendly Transformer. In *IEEE International Conference on Computer Vision*, pages 589–598, 2021.
- [103] Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. DPT: Deformable Patch-Based Transformer for Visual Recognition. In *29th ACM International Conference on Multimedia*, pages 2899–2907, 2021.
- [104] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020.
- [105] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.
- [106] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-Pixel Classification Is Not All You Need for Semantic Segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.
- [107] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-Free Oriented Proposal Generator for Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.
- [108] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebin Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [109] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse Instance Activation for Real-Time Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4433–4442, 2022.
- [110] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers. *ArXiv Preprint ArXiv:1904.10509*, 2019.
- [111] Jae-Seok Choi and Munchurl Kim. A Deep Convolutional Neural Network with Selection Units for Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 154–160, 2017.
- [112] François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

-
- [113] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking Attention with Performers. *ArXiv Preprint ArXiv:2009.14794*, 2020.
 - [114] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.
 - [115] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional Positional Encodings for Vision Transformers. *ArXiv Preprint ArXiv:2102.10882*, 2021.
 - [116] Fan Chung and Linyuan Lu. The Average Distances in Random Graphs with Given Expected Degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.
 - [117] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.
 - [118] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). ArXiv 2015. *ArXiv Preprint ArXiv:1511.07289*, 2020.
 - [119] Taco S Cohen and Max Welling. Steerable CNNs. *ArXiv Preprint ArXiv:1612.08498*, 2016.
 - [120] Shuang Cong and Yang Zhou. A Review of Convolutional Neural Network Architectures and Their Optimizations. *Artificial Intelligence Review*, 56(3):1905–1969, 2023.
 - [121] MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. <https://github.com/open-mmlab/mms Segmentation>, 2020.
 - [122] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the Relationship Between Self-Attention and Convolutional Layers. *ArXiv Preprint ArXiv:1911.03584*, 2019.
 - [123] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.
 - [124] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

Bibliography

- [125] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-Sensitive Fully Convolutional Networks. In *European Conference on Computer Vision*, pages 534–549. Springer, 2016.
- [126] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object Detection via Region-Based Fully Convolutional Networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.
- [127] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision*, pages 764–773, 2017.
- [128] Linhui Dai, Hong Liu, Hao Tang, Zhiwei Wu, and Pinhao Song. AO2-DETR: Arbitrary-Oriented Object Detection Transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2342–2356, 2023.
- [129] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-Order Attention Network for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.
- [130] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic Head: Unifying Object Detection Heads with Attentions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021.
- [131] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. CoAtNet: Marrying Convolution and Attention for All Data Sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.
- [132] Lanxue Dang, Peidong Pang, and Jay Lee. Depth-Wise Separable Convolution Neural Network with Residual Connection for Hyperspectral Image Classification. *Remote Sensing*, 12(20):3408, 2020.
- [133] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024.
- [134] Brent De Weerd, Yonina C. Eldar, and Nikos Deligiannis. Designing transformer networks for sparse recovery of sequential data using deep unfolding. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.
- [135] Sukanya Deka, Kuntal Deka, Nhan Thanh Nguyen, Sanjeev Sharma, Vimal Bhatia, and Nandana Rajatheva. Comprehensive review of deep unfolding techniques for next-generation wireless communication systems, 2025.
- [136] Federico Del Pup and Manfredo Atzori. Applications of Self-Supervised Learning to Biomedical Signals: A Survey. *IEEE Access*, 2023.

-
- [137] Chunfang Deng, Mengmeng Wang, Liang Liu, Yong Liu, and Yunliang Jiang. Extended Feature Pyramid Network for Small Object Detection. *IEEE Transactions on Multimedia*, 24:1968–1979, 2021.
 - [138] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.
 - [139] Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jun Long, Zuhayr Asad, R Michael Womick, Zheyu Zhu, Agnes B Fogo, Shilin Zhao, et al. Omni-Seg: A Scale-Aware Dynamic Network for Renal Pathological Image Segmentation. *IEEE Transactions on Biomedical Engineering*, 2023.
 - [140] Ruining Deng, Quan Liu, Can Cui, Tianyuan Yao, Jialin Yue, Juming Xiong, Lining Yu, Yifei Wu, Mengmeng Yin, Yu Wang, et al. PrPSeg: Universal Proposition Learning for Panoramic Renal Pathology Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11736–11746, 2024.
 - [141] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr2: Image Style Transfer with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11326–11336, 2022.
 - [142] Misha Denil, Babak Shakibi, Laurent Dinh, Marc’Aurelio Ranzato, and Nando De Freitas. Predicting Parameters in Deep Learning. *Advances in Neural Information Processing Systems*, 26, 2013.
 - [143] Emily L. Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. *Advances in Neural Information Processing Systems*, 27, 2014.
 - [144] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
 - [145] Haisong Ding, Kai Chen, and Qiang Huo. Compressing CNN-DBLSTM Models for OCR with Teacher-Student Learning and Tucker Decomposition. *Pattern Recognition*, 96:106957, 2019.
 - [146] Ruizhou Ding, Ting-Wu Chin, Zeye Liu, and Diana Marculescu. Regularizing Activation Distribution for Training Binarized Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11408–11417, 2019.
 - [147] Xiaohan Ding, Tianxiang Hao, Jianchao Tan, Ji Liu, Jungong Han, Yuchen Guo, and Guiguang Ding. ResRep: Lossless CNN Pruning via Decoupling Remembering and Forgetting. In *IEEE International Conference on Computer Vision*, pages 4510–4520, 2021.

Bibliography

- [148] Zixiang Ding, Yaran Chen, Nannan Li, Dongbin Zhao, Zhiquan Sun, and CL Philip Chen. BNAS: Efficient Neural Architecture Search Using Broad Scalable Architecture. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):5004–5018, 2021.
- [149] Yasser Abdelaziz Dahou Djilali, Tarun Krishna, Kevin McGuinness, and Noel E O’Connor. Rethinking 360° Image Visual Attention Modelling with Unsupervised Learning. In *IEEE International Conference on Computer Vision*, pages 15414–15424, 2021.
- [150] Carl Doersch. Tutorial on Variational Autoencoders. *ArXiv Preprint ArXiv:1606.05908*, 2016.
- [151] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2018.
- [152] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a Deep Convolutional Network for Image Super-Resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.
- [153] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the Super-Resolution Convolutional Neural Network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016.
- [154] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.
- [155] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Baining Guo. PeCo: Perceptual Codebook for BERT Pre-Training of Vision Transformers. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 552–560, 2023.
- [156] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv Preprint ArXiv:2010.11929*, 2020.
- [157] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35*:

- Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- [158] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022.
 - [159] Chengzhen Duan, Zhiwei Wei, Chi Zhang, Siying Qu, and Hongpeng Wang. Coarse-Grained Density Map Guided Object Detection in Aerial Images. In *IEEE International Conference on Computer Vision*, pages 2789–2798, 2021.
 - [160] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. CenterNet: Keypoint Triplets for Object Detection. In *IEEE International Conference on Computer Vision*, pages 6569–6578, 2019.
 - [161] Stéphane d’Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.
 - [162] Mohamed Elasri, Omar Elharrouss, Somaya Al-Maadeed, and Hamid Tairi. Image Generation: A Review. *Neural Processing Letters*, 54(5):4609–4646, 2022.
 - [163] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural Architecture Search: A Survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019.
 - [164] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable Object Detection Using Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2154, 2014.
 - [165] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in Style. In *IEEE International Conference on Computer Vision*, pages 2251–2261, 2023.
 - [166] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.
 - [167] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
 - [168] Angela Fan, Edouard Grave, and Armand Joulin. Reducing Transformer Depth on Demand with Structured Dropout. *ArXiv Preprint ArXiv:1909.11556*, 2019.
 - [169] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 605–613, 2017.

Bibliography

- [170] Jiansong Fan, Tianxu Lv, Pei Wang, Xiaoyan Hong, Yuan Liu, Chunjuan Jiang, Jianming Ni, Lihua Li, and Xiang Pan. DCDiff: Dual-Granularity Cooperative Diffusion Models for Pathology Image Analysis. *IEEE Transactions on Medical Imaging*, 2024.
- [171] Yijian Fan, Zhenbang Yang, Rui Liu, Mingjie Li, and Xiaojun Chang. Medical Report Generation Is a Multi-Label Classification Problem. *ArXiv Preprint ArXiv:2409.00250*, 2024.
- [172] Yuchen Fan, Honghui Shi, Jiahui Yu, Ding Liu, Wei Han, Haichao Yu, Zhangyang Wang, Xinchao Wang, and Thomas S Huang. Balanced Two-Stage Residual Networks for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 161–168, 2017.
- [173] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data Augmentation for Object Detection via Controllable Diffusion Models. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1257–1266, 2024.
- [174] Jiemin Fang, Lingxi Xie, Xinggang Wang, Xiaopeng Zhang, Wenyu Liu, and Qi Tian. MSG-Transformer: Exchanging Local Spatial Information by Manipulating Messenger Tokens. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12063–12072, 2022.
- [175] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear Attention Networks for Person Retrieval. In *IEEE International Conference on Computer Vision*, pages 8030–8039, 2019.
- [176] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You Only Look at One Sequence: Rethinking Transformer in Vision Through Object Detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021.
- [177] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as Queries, 2021.
- [178] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. QueryInst: Parallely Supervised Mask Query for Instance Segmentation. *ArXiv Preprint ArXiv:2105.01928*, 3(10), 2021.
- [179] Chunyue Feng, Kokhaur Ong, David M Young, Bingxian Chen, Longjie Li, Xinmi Huo, Haoda Lu, Weizhong Gu, Fei Liu, Hongfeng Tang, et al. Artificial Intelligence-Assisted Quantification and Assessment of Whole Slide Images for Pediatric Kidney Disease Diagnosis. *Bioinformatics*, 40(1):btad740, 2024.
- [180] Michael Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially Adaptive Computation Time for Residual Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2017.

-
- [181] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. DSSD: Deconvolutional Single Shot Detector. *ArXiv Preprint ArXiv:1701.06659*, 2017.
 - [182] Jiamei Fu, Xian Sun, Zhirui Wang, and Kun Fu. An Anchor-Free Method Based on Feature Balancing and Refinement Network for Multiscale Ship Detection in SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(2):1331–1344, 2020.
 - [183] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018.
 - [184] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual Attention Network for Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.
 - [185] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient Knowledge Distillation from an Ensemble of Teachers. In *Interspeech*, pages 3697–3701, 2017.
 - [186] Kunihiro Fukushima. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36(4):193–202, 1980.
 - [187] Michael Gadermayr, Laxmi Gupta, Vitus Appel, Peter Boor, Barbara M Klinkhammer, and Dorit Merhof. Generative Adversarial Networks for Facilitating Stain-Independent Supervised and Unsupervised Segmentation: A Study on Kidney Histology. *IEEE Transactions on Medical Imaging*, 38(10):2293–2302, 2019.
 - [188] Hongyang Gao, Hao Yuan, Zhengyang Wang, and Shuiwang Ji. Pixel Transposed Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1218–1227, 2019.
 - [189] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast Convergence of DETR with Spatially Modulated Co-Attention. In *IEEE International Conference on Computer Vision*, pages 3621–3630, 2021.
 - [190] Penglei Gao, Xi Yang, Rui Zhang, John Y Goulermas, Yujie Geng, Yuyao Yan, and Kaizhu Huang. Generalized Image Outpainting with U-Transformer. *Neural Networks*, 162:1–10, 2023.
 - [191] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked Diffusion Transformer Is a Strong Image Synthesizer. In *IEEE International Conference on Computer Vision*, pages 23164–23173, 2023.
 - [192] Lihui Ge, Guanqun Wang, Tong Zhang, Yin Zhuang, He Chen, Hao Dong, and Liang Chen. Regression-Guided Refocusing Learning with Feature Alignment for Remote Sensing Tiny Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Bibliography

- [193] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is Attention Better Than Matrix Decomposition? *ArXiv Preprint ArXiv:2109.04553*, 2021.
- [194] Pascal Getreuer. Rudin-Osher-Fatemi Total Variation Denoising Using Split Bregman. *Image Processing On Line*, 2:74–95, 2012.
- [195] Zoubin Ghahramani and Michael Jordan. Factorial Hidden Markov Models. *Advances in Neural Information Processing Systems*, 8, 1995.
- [196] R Girshick, J Donahue, T Darrell, and J Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *ArXiv Preprint ArXiv:1311.2524*, 2014.
- [197] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.
- [198] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. Compressing Deep Convolutional Networks Using Vector Quantization. *ArXiv Preprint ArXiv:1412.6115*, 2014.
- [199] Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, and Zhenjun Han. Effective Fusion Factor in FPN for Tiny Object Detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1160–1168, 2021.
- [200] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., USA, 2006.
- [201] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [202] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [203] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. LeViT: A Vision Transformer in ConvNet’s Clothing for Faster Inference. In *IEEE International Conference on Computer Vision*, pages 12259–12269, 2021.
- [204] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9224–9232, 2018.
- [205] Alex Graves. Long Short-Term Memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012.

-
- [206] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhao-han Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.
- [207] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.
- [208] Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, et al. DataSeg: Taming a Universal Multi-Dataset Multi-Task Segmentation Model. *Advances in Neural Information Processing Systems*, 36:67329–67354, 2023.
- [209] Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, Jun Lan, Changhua Meng, and Weiqiang Wang. DiffusionInst: Diffusion Model for Instance Segmentation. *ArXiv Preprint ArXiv:2212.02773*, 2022.
- [210] Yigit Gunduc. ViT-GAN: Image-to-Image Translation with Vision Transformers and Conditional GANS. *Authorea Preprints*, 2023.
- [211] J Guo, Y Li, W Lin, Y Chen, and J Li. Network Decoupling: From Regular to Depthwise Separable Convolutions. *ArXiv Preprint ArXiv:1808.05517*, 2018.
- [212] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling Object Detectors via Decoupled Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021.
- [213] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. CMT: Convolutional Neural Networks Meet Vision Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022.
- [214] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media*, 8(3):331–368, 2022.
- [215] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online Knowledge Distillation via Collaborative Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020.
- [216] Yong Guo, Yin Zheng, Mingkui Tan, Qi Chen, Jian Chen, Peilin Zhao, and Junzhou Huang. NAT: Neural Architecture Transformer for Accurate and Compact Architectures. *Advances in Neural Information Processing Systems*, 32, 2019.
- [217] Yuanhao Guo, Zhan Xiong, and Fons J. Verbeek. An Efficient and Robust Hybrid Method for Segmentation of Zebrafish Objects from Bright-Field Microscope Images. *Machine Vision and Applications*, 29(8):1211–1225, 2018.

Bibliography

- [218] Zonghao Guo, Chang Liu, Xiaosong Zhang, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Beyond Bounding-Box: Convex-Hull Feature Adaptation for Oriented and Densely Packed Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8792–8801, 2021.
- [219] Gousia Habib, Tausifa Jan Saleem, and Brejesh Lall. Knowledge Distillation in Vision Transformers: A Critical Review. *ArXiv Preprint ArXiv:2302.02108*, 2023.
- [220] Taeyoung Hahn, Myeongjang Pyeon, and Gunhee Kim. Self-Routing Capsule Networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- [221] Cheng Han, James Chenhao Liang, Qifan Wang, MAJID RABBANI, Sohail Dianat, Raghuveer Rao, Ying Nian Wu, and Dongfang Liu. Image Translation as Diffusion Visual Programmers. In *The Twelfth International Conference on Learning Representations*, 2024.
- [222] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align Deep Features for Oriented Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [223] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. ReDet: A Rotation-Equivariant Detector for Aerial Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021.
- [224] Jian Han, Yaping Liao, Junyou Zhang, Shufeng Wang, and Sixian Li. Target Fusion Detection of LiDAR and Camera Based on the Improved YOLO Algorithm. *Mathematics*, 6(10):213, 2018.
- [225] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A Survey on Visual Transformer. *ArXiv Preprint ArXiv:2012.12556*, 2020.
- [226] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in Transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.
- [227] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. Image Super-Resolution via Dual-State Recurrent Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1654–1663, 2018.
- [228] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic Contours from Inverse Detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011.
- [229] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous Detection and Segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

-
- [230] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep Back-Projection Networks for Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018.
- [231] Kyle Harlow, Hyesu Jang, Timothy D Barfoot, Ayoung Kim, and Christof-fer Heckman. A New Wave in Robotics: Survey on Recent mmWave Radar Applications in Robotics. *IEEE Transactions on Robotics*, 2024.
- [232] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the Big Data Paradigm with Compact Transformers. *ArXiv Preprint ArXiv:2104.05704*, 2021.
- [233] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. DiffiT: Diffusion Vision Transformers for Image Generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2025.
- [234] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22046–22055, 2023.
- [235] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *NeurIPS*, 2023.
- [236] Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Youwei Pang, Longxiang Tang, Chengyu Fang, Yulun Zhang, Linghe Kong, Xiu Li, and Sina Farsiu. Segment concealed objects with incomplete supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7832–7851, 2025.
- [237] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4630–4651, 2025.
- [238] Chunming He, Fengyang Xiao, Rihan Zhang, Chengyu Fang, Deng-Ping Fan, and Sina Farsiu. Reversible unfolding network for concealed visual perception with generative refinement, 2025.
- [239] Chunming He, Rihan Zhang, Longxiang Tang, Ziyun Yang, Kai Li, Deng-Ping Fan, and Sina Farsiu. Scaler: Sam-enhanced collaborative learning for label-deficient concealed object segmentation, 2025.
- [240] Chunming He, Rihan Zhang, Dingming Zhang, Fengyang Xiao, Deng-Ping Fan, and Sina Farsiu. Nested unfolding network for real-world concealed object segmentation, 2025.

Bibliography

- [241] Junling He, Pieter A. Valkema, Jingmin Long, Jia Li, Sandrine Florquin, Maarten Naesens, Priyanka Koshy, Tri Q. Nguyen, Soufian Meziyeh, Aiko P. J. de Vries, Onno J. De Boer, Fons J. Verbeek, Zhan Xiong, and Jesper Kers. Transformer-Based Multiclass Segmentation Pipeline for Basic Kidney Histology. *Scientific Reports*, 15(1):38957, 2025.
- [242] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [243] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. *ArXiv Preprint ArXiv:1911.05722*, 2019.
- [244] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.
- [245] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.
- [246] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [247] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.
- [248] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [249] Yang He and Lingao Xiao. Structured Pruning for Deep Convolutional Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [250] Yihui He, Xiangyu Zhang, and Jian Sun. Channel Pruning for Accelerating Very Deep Neural Networks. In *IEEE International Conference on Computer Vision*, pages 1389–1397, 2017.
- [251] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *ArXiv Preprint ArXiv:1606.08415*, 2016.
- [252] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking Spatial Dimensions of Vision Transformers. In *IEEE International Conference on Computer Vision*, pages 11936–11945, 2021.

-
- [253] Meyke Hermesen, Thomas de Bel, Marjolijn Den Boer, Eric J Steenbergen, Jesper Kers, Sandrine Florquin, Joris JTH Roelofs, Mark D Stegall, Mariam P Alexander, Byron H Smith, et al. Deep Learning–Based Histopathologic Assessment of Kidney Tissue. *Journal of the American Society of Nephrology: JASN*, 30(10):1968, 2019.
- [254] Geoffrey Hinton. Distilling the Knowledge in a Neural Network. *ArXiv Preprint ArXiv:1503.02531*, 2015.
- [255] Geoffrey E Hinton, Zoubin Ghahramani, and Yee Whye Teh. Learning to Parse Images. *Advances in Neural Information Processing Systems*, 12, 1999.
- [256] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix Capsules with EM Routing. In *International Conference on Learning Representations*, 2018.
- [257] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, 2006.
- [258] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *ArXiv Preprint ArXiv:1207.0580*, 2012.
- [259] Geoffrey E Hinton and Richard Zemel. Autoencoders, Minimum Description Length and Helmholtz Free Energy. *Advances in Neural Information Processing Systems*, 6, 1993.
- [260] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [261] Mingbo Hong, Shuiwang Li, Yuchao Yang, Feiyu Zhu, Qijun Zhao, and Li Lu. SSPNet: Scale Selection Pyramid Network for Tiny Person Detection from UAV Images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.
- [262] Weixiang Hong, Qingpei Guo, Wei Zhang, Jingdong Chen, and Wei Chu. LPSNet: A Lightweight Solution for Fast Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16746–16754, 2021.
- [263] Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. DynaBERT: Dynamic BERT with Adaptive Width and Depth. *Advances in Neural Information Processing Systems*, 33:9782–9793, 2020.
- [264] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for MobileNetV3. In *IEEE International Conference on Computer Vision*, pages 1314–1324, 2019.
- [265] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017). *ArXiv Preprint ArXiv:1704.04861*, 126, 2017.

Bibliography

- [266] H Hu. Network Trimming: A Data-Driven Neuron Pruning Approach Towards Efficient Deep Architectures. *ArXiv Preprint ArXiv:1607.03250*, 2016.
- [267] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local Relation Networks for Image Recognition. In *IEEE International Conference on Computer Vision*, pages 3464–3473, 2019.
- [268] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.
- [269] Kai Hu, Junlan Jin, Fei Zheng, Liguang Weng, and Yiwu Ding. Overview of Behavior Recognition Based on Deep Learning. *Artificial Intelligence Review*, 56(3):1833–1865, 2023.
- [270] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep Level Sets for Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2300–2309, 2017.
- [271] Qinghao Hu, Peisong Wang, and Jian Cheng. From Hashing to CNNs: Training Binary Weight Networks via Hashing. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [272] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert Knowledge-Aware Image Difference Graph Representation Learning for Difference-Aware Medical Visual Question Answering. In *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165, 2023.
- [273] Yanting Hu, Xinbo Gao, Jie Li, Yuanfei Huang, and Hanzi Wang. Single Image Super-Resolution via Cascaded Multi-Scale Cross Network. *ArXiv Preprint ArXiv:1802.08808*, 2018.
- [274] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.
- [275] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024.
- [276] Gao Huang, Shichen Liu, Laurens Van der Maaten, and Kilian Q Weinberger. CondenseNet: An Efficient DenseNet Using Learned Group Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2761, 2018.
- [277] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.

-
- [278] Qiao Huang and Jean-Claude Zambrini. Stochastic Geometric Mechanics in Nonequilibrium Thermodynamics: Schrödinger Meets Onsager. *Journal of Physics A: Mathematical and Theoretical*, 56(13):134003, 2023.
- [279] Rachel Huang, Jonathan Pedoeem, and Cuixian Chen. YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers. In *2018 IEEE International Conference on Big Data*, pages 2503–2510. IEEE, 2018.
- [280] Xiaofei Huang and Hongfang Gong. A Dual-Attention Learning Network with Word and Sentence Embedding for Medical Visual Question Answering. *IEEE Transactions on Medical Imaging*, 2023.
- [281] Xinyang Huang, Guanghui Zhang, Dongchen Zhu, Yunpeng Sun, Wenjun Shi, Gang Ye, Yang Xiao, Lei Wang, Xiaolin Zhang, Bo Li, and Jiamao Li. BEE-Net: Bridging Semantic and Instance with Gated Encoding and Edge Constraint for Efficient Panoptic Segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4281–4287, 2024.
- [282] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal Unsupervised Image-to-Image Translation. In *European Conference on Computer Vision*, pages 172–189, 2018.
- [283] Zhanchao Huang, Wei Li, Xiang-Gen Xia, Xin Wu, Zhaoquan Cai, and Ran Tao. A Novel Nonlocal-Aware Pyramid and Multiscale Multitask Refinement Detector for Object Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2021.
- [284] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask Scoring R-CNN. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019.
- [285] Zhongzhan Huang, Senwei Liang, Mingfu Liang, and Haizhao Yang. Dianet: Dense-and-Implicit Attention Network. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 4206–4214, 2020.
- [286] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer. *ArXiv Preprint ArXiv:2106.03650*, 2021.
- [287] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-Cross Attention for Semantic Segmentation. In *IEEE International Conference on Computer Vision*, pages 603–612, 2019.
- [288] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv Preprint ArXiv:1502.03167*, 2015.

Bibliography

- [289] Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G Derpanis, and Neil DB Bruce. Position, Padding and Predictions: A Deeper Look at Position Information in CNNs. *International Journal of Computer Vision*, pages 1–22, 2024.
- [290] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.
- [291] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding Up Convolutional Neural Networks with Low Rank Expansions. *ArXiv Preprint ArXiv:1405.3866*, 2014.
- [292] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General Perception with Iterative Attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021.
- [293] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.
- [294] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. SeMask: Semantically Masked Transformers for Semantic Segmentation. In *IEEE International Conference on Computer Vision*, pages 752–761, 2023.
- [295] Yonten Jamtsho, Panomkhawn Riyamongkol, and Rattapoom Waranusast. Real-Time License Plate Detection for Non-Helmeted Motorcyclist Using YOLO. *ICT Express*, 7(1):104–109, 2021.
- [296] Dong-Won Jang and Rae-Hong Park. Densenet with Deep Residual Channel-Attention Blocks for Single Image Super Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [297] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’23/IAAI’23/EAAI’23. AAAI Press, 2023.
- [298] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2010.
- [299] Simon Jégou, Michal Drozdal, David Vazquez, Adriana Romero, and Yoshua Bengio. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for

- Semantic Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183. IEEE, 2017.
- [300] Yunho Jeon and Junmo Kim. Active Convolution: Learning the Shape of Convolution for Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4201–4209, 2017.
- [301] Aadarsh Jha, Haichun Yang, Ruining Deng, Meghan E Kapp, Agnes B Fogo, and Yuankai Huo. Instance Segmentation for Whole Slide Imaging: End-to-End or Detect-Then-Segment. *Journal of Medical Imaging*, 8(1):014001–014001, 2021.
- [302] Mingi Ji, Seungjae Shin, Seunghyun Hwang, Gibeom Park, and Il-Chul Moon. Refine Myself by Teaching Myself: Feature Refinement via Self-Knowledge Distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10664–10673, 2021.
- [303] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.
- [304] D Jia, K Han, Y Wang, Y Tang, J Guo, C Zhang, and D Tao. Efficient Vision Transformers via Fine-Grained Manifold Distillation, 2021.
- [305] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic Filter Networks. *Advances in Neural Information Processing Systems*, 29, 2016.
- [306] Yangqing Jia, Chang Huang, and Trevor Darrell. Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3370–3377. IEEE, 2012.
- [307] Zhen Jia, Zhang Zhang, Liang Wang, and Tieniu Tan. Human Image Generation: A Comprehensive Survey. *ACM Computing Surveys*, 56(11):1–39, 2024.
- [308] Lei Jiang, Wenkai Chen, Bao Dong, Ke Mei, Chuang Zhu, Jun Liu, Meishun Cai, Yu Yan, Gongwei Wang, Li Zuo, et al. A Deep Learning-Based Approach for Glomeruli Instance Segmentation from Multistained Renal Biopsy Pathologic Images. *The American Journal of Pathology*, 191(8):1431–1441, 2021.
- [309] Peng Jiang, Fanglin Gu, Yunhai Wang, Changhe Tu, and Baoquan Chen. DifNet: Semantic Segmentation by Diffusion Networks. In *Advances in Neural Information Processing Systems*, pages 1637–1646, 2018.
- [310] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021.

Bibliography

- [311] Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. ConvBERT: Improving BERT with Span-Based Dynamic Convolution. *Advances in Neural Information Processing Systems*, 33:12837–12848, 2020.
- [312] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Andy J Ma, Yaowei Wang, and Wei-Shi Zheng. DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition. *IEEE Transactions on Multimedia*, 25:8906–8919, 2023.
- [313] Baoyu Jing, Pengtao Xie, and Eric Xing. On the Automatic Generation of Medical Imaging Reports. *ArXiv Preprint ArXiv:1711.08195*, 2017.
- [314] Tao Jing, Qing-Hao Meng, and Hui-Rang Hou. SmokeSegger: A Transformer-CNN Coupled Model for Urban Scene Smoke Segmentation. *IEEE Transactions on Industrial Informatics*, 20(2):1385–1396, 2023.
- [315] Sanghyun Jo, In-Jae Yu, and Kyungsu Kim. MARS: Model-Agnostic Biased Object Removal Without Additional Supervision for Weakly-Supervised Semantic Segmentation. *ArXiv Preprint ArXiv:2304.09913*, 2023.
- [316] Hongtao Kang, Die Luo, Weihua Feng, Shaoqun Zeng, Tingwei Quan, Junbo Hu, and Xiuli Liu. StainNet: A Fast and Robust Stain Normalization Network. *Frontiers in Medicine*, 8:746307, 2021.
- [317] Wei-Cheng Kao, Hong-Xia Xie, Chih-Yang Lin, and Wen-Huang Cheng. Specific Expert Learning: Enriching Ensemble Diversity via Knowledge Distillation. *IEEE Transactions on Cybernetics*, 53(4):2494–2505, 2021.
- [318] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.
- [319] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask Transfuser for High-Quality Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4412–4421, 2022.
- [320] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.
- [321] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.
- [322] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-Recursive Convolutional Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016.

-
- [323] Taeksoo Kim, Moon-su Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017.
 - [324] Wonjik Kim, Asako Kanezaki, and Masayuki Tanaka. Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering. *IEEE Transactions on Image Processing*, 29:8055–8068, 2020.
 - [325] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications. *ArXiv Preprint ArXiv:1511.06530*, 2015.
 - [326] Diederik P Kingma. Auto-Encoding Variational Bayes. *ArXiv Preprint ArXiv:1312.6114*, 2013.
 - [327] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.
 - [328] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.
 - [329] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.
 - [330] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
 - [331] Alexander Kirillov, Yuxin Wu, Kaiming He, and Piotr Dollár. PointRend: Image Segmentation as Rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9808, 2020.
 - [332] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer. *ArXiv Preprint ArXiv:2001.04451*, 2020.
 - [333] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen. RON: Reverse Connection with Objectness Prior Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5936–5944, 2017.
 - [334] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 845–853, 2016.

Bibliography

- [335] Vasiliki Kouni, Georgios Paraskevopoulos, Holger Rauhut, and George C. Alexandropoulos. Admm-dad net: A deep unfolding network for analysis compressed sensing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1506–1510, 2022.
- [336] Onur Can Koyun, Reyhan Kevser Keser, Ibrahim Batuhan Akkaya, and Behçet Uğur Töreyn. Focus-and-Detect: A Small Object Detection Framework for Aerial Images. *Signal Processing: Image Communication*, 104:116675, 2022.
- [337] Nikolay Kozyrskiy and Anh-Huy Phan. CNN Acceleration by Low-Rank Approximation with Quantized Factors. *ArXiv Preprint ArXiv:2006.08878*, 2020.
- [338] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *Advances in Neural Information Processing Systems*, pages 109–117, 2011.
- [339] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, Dec 2013.
- [340] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [341] Uday Kulkarni, Sitanshu S Hallad, Abhishek Patil, Tanvi Bhujannavar, Satwik Kulkarni, and SM Meena. A Survey on Filter Pruning Techniques for Optimization of Deep Neural Networks. In *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 610–617. IEEE, 2022.
- [342] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the Potential of Ordinary Classifier: Class-Specific Adversarial Erasing Framework for Weakly Supervised Semantic Segmentation. In *IEEE International Conference on Computer Vision*, pages 6994–7003, 2021.
- [343] Hyeokjun Kweon, Sung-Hoon Yoon, and Kuk-Jin Yoon. Weakly Supervised Semantic Segmentation via Adversarial Learning of Classifier and Reconstructor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11329–11339, 2023.
- [344] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, 2018.

- [345] Varsha S Lalapura, Joseph Amudha, and Hariramn Selvamuruga Satheesh. Recurrent neural networks for edge intelligence: a survey. *ACM Computing Surveys (CSUR)*, 54(4):1–38, 2021.
- [346] Hei Law and Jia Deng. CornerNet: Detecting Objects as Paired Keypoints. In *European Conference on Computer Vision*, pages 734–750, 2018.
- [347] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. Speeding-Up Convolutional Neural Networks Using Fine-Tuned CP-Decomposition. *ArXiv Preprint ArXiv:1412.6553*, 2014.
- [348] Vadim Lebedev and Victor Lempitsky. Fast ConvNets Using Group-Wise Brain Damage. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2554–2564, 2016.
- [349] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015.
- [350] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 1989.
- [351] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [352] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.
- [353] Donghyun Lee, Dingheng Wang, Yukuan Yang, Lei Deng, Guangshe Zhao, and Guoqi Li. QTTNet: Quantized Tensor Train Neural Networks for 3D Object and Video Recognition. *Neural Networks*, 141:420–432, 2021.
- [354] Heng Lee, Yi-Heng Wu, Yu-Sheng Lin, and Shao-Yi Chien. Convolutional Neural Network Accelerator with Vector Quantization. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019.
- [355] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing Information Bottleneck for Weakly Supervised Semantic Segmentation. *Advances in Neural Information Processing Systems*, 34:27408–27421, 2021.
- [356] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021.

Bibliography

- [357] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly Supervised Semantic Segmentation Using Out-of-Distribution Data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022.
- [358] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. ViTGAN: Training GANs with Vision Transformers. *ArXiv Preprint ArXiv:2107.04589*, 2021.
- [359] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. MPViT: Multi-Path Vision Transformer for Dense Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2022.
- [360] Youngwan Lee and Jongyoul Park. CenterMask: Real-Time Anchor-Free Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2020.
- [361] Dawa Chyophel Lepcha, Bhawna Goyal, Ayush Dogra, and Vishal Goyal. Image Super-Resolution: A Comprehensive Review, Recent Trends, Challenges and Applications. *Information Fusion*, 91:230–260, 2023.
- [362] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023.
- [363] Changlin Li, Tao Tang, Guangrun Wang, Jiefeng Peng, Bing Wang, Xiaodan Liang, and Xiaojun Chang. BossNAS: Exploring Hybrid CNN-Transformers with Block-Wisely Self-Supervised Neural Architecture Search. In *IEEE International Conference on Computer Vision*, pages 12281–12291, 2021.
- [364] Chao Li, Aojun Zhou, and Anbang Yao. Omni-Dimensional Dynamic Convolution. *ArXiv Preprint ArXiv:2209.07947*, 2022.
- [365] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. Distance Regularized Level Set Evolution and Its Application to Image Segmentation. *IEEE Transactions on Image Processing*, 19(12):3243–3254, 2010.
- [366] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient Self-Supervised Vision Transformers for Representation Learning. *ArXiv Preprint ArXiv:2106.09785*, 2021.
- [367] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. DN-DETR: Accelerate DETR Training by Introducing Query Denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.
- [368] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask DINO: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.

-
- [369] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023.
 - [370] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models. *Neurocomputing*, 479:47–59, 2022.
 - [371] Hongyang Li, Yu Liu, Wanli Ouyang, and Xiaogang Wang. Zoom Out-and-In Network with Map Attention Decision for Region Proposal and Object Detection. *International Journal of Computer Vision*, pages 1–14, 2017.
 - [372] Huantong Li, Xiangmiao Wu, Fanbing Lv, Daihai Liao, Thomas H Li, Yonggang Zhang, Bo Han, and Mingkui Tan. Hard Sample Matters a Lot in Zero-Shot Quantization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 24417–24426, 2023.
 - [373] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios. *ArXiv Preprint ArXiv:2207.05501*, 2022.
 - [374] Jingxiong Li, Sunyi Zheng, Chenglu Zhu, Yuxuan Sun, Pingyi Chen, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. PathUp: Patch-Wise Timestep Tracking for Multi-Class Large Pathology Image Synthesising Diffusion Model. In *32nd ACM International Conference on Multimedia*, pages 3984–3993, 2024.
 - [375] Juan Li, Wenkai Xu, Limiao Deng, Ying Xiao, Zhongzhi Han, and Haiyong Zheng. Deep Learning for Visual Recognition and Detection of Aquatic Animals: A Review. *Reviews in Aquaculture*, 15(2):409–433, 2023.
 - [376] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022.
 - [377] Ruibo Li, Chi Zhang, Guosheng Lin, Zhe Wang, and Chunhua Shen. RigidFlow: Self-Supervised Scene Flow Learning on Point Clouds by Local Rigidity Prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16959–16968, 2022.
 - [378] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-Buddy GANs for Highly Detailed Image Super-Resolution. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 1412–1420, 2022.

Bibliography

- [379] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-Maximization Attention Networks for Semantic Segmentation. In *IEEE International Conference on Computer Vision*, pages 9167–9176, 2019.
- [380] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM Improves Controllable Text Generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.
- [381] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective Kernel Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019.
- [382] Xiangtai Li, Shilin Xu, Yibo Yang Cheng, Yunhai Tong, Dacheng Tao, et al. Panoptic-PartFormer: Learning a Unified Model for Panoptic Part Segmentation. *ECCV*, 2022.
- [383] Xiangtai Li, Shilin Xu, Yibo Yang, Haobo Yuan, Guangliang Cheng, Yunhai Tong, Zhouchen Lin, Ming-Hsuan Yang, and Dacheng Tao. PanopticPartFormer++: A Unified and Decoupled View for Panoptic Part Segmentation, 2024.
- [384] Xiaopeng Li, Yuyun Xiang, and Shuqin Li. Combining Convolutional and Vision Transformer Structures for Sheep Face Recognition. *Computers and Electronics in Agriculture*, 205:107651, 2023.
- [385] Xingjian Li, Haoyi Xiong, Zeyu Chen, Jun Huan, Cheng-Zhong Xu, and Dejing Dou. “In-Network Ensemble”: Deep Ensemble Learning with Diversified Knowledge Distillation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–19, 2021.
- [386] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [387] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring Plain Vision Transformer Backbones for Object Detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.
- [388] Yangyang Li, Qin Huang, Xuan Pei, Yanqiao Chen, Licheng Jiao, and Ronghua Shang. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2148–2161, 2020.
- [389] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully Convolutional Networks for Panoptic Segmentation with Point-Based Supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4552–4568, 2022.

-
- [390] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully Convolutional Networks for Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2021.
 - [391] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. EfficientFormer: Vision Transformers at MobileNet Speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022.
 - [392] Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group Sparsity: The Hinge Between Filter Pruning and Decomposition for Network Compression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2020.
 - [393] Yawei Li, Kai Zhang, Jiezhong Cao, Radu Timofte, and Luc Van Gool. LocalViT: Bringing Locality to Vision Transformers. *ArXiv Preprint ArXiv:2104.05707*, 2021.
 - [394] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty Estimation via Response Scaling for Pseudo-Mask Noise Mitigation in Weakly-Supervised Semantic Segmentation. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 1447–1455, 2022.
 - [395] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. *Advances in Neural Information Processing Systems*, 31, 2018.
 - [396] Yue Li, Wenrui Ding, Chunlei Liu, Baochang Zhang, and Guodong Guo. TRQ: Ternary Neural Networks with Residual Quantization. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 8538–8546, 2021.
 - [397] Yuxuan Li, Xiang Li, Yimain Dai, Qibin Hou, Li Liu, Yongxiang Liu, Ming-Ming Cheng, and Jian Yang. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *International Journal of Computer Vision*, 133(3):1410–1431, 2025.
 - [398] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-Head R-CNN: In Defense of Two-Stage Object Detector. *ArXiv Preprint ArXiv:1711.07264*, 2017.
 - [399] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. MST: Masked Self-Supervised Transformer for Visual Representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.
 - [400] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1280–1289, 2022.

Bibliography

- [401] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. LViT: Language Meets Vision Transformer in Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 2023.
- [402] Zuoxin Li and Fuqiang Zhou. FSSD: Feature Fusion Single Shot Multibox Detector. *ArXiv Preprint ArXiv:1712.00960*, 2017.
- [403] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image Restoration Using Swin Transformer. In *IEEE International Conference on Computer Vision*, pages 1833–1844, 2021.
- [404] Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meir, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid transformer-mamba language model, 2024.
- [405] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017.
- [406] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.
- [407] Guoyu Lin, Zhentai Zhang, Kaixing Long, Yiwen Zhang, Yanmeng Lu, Jian Geng, Zhitao Zhou, Qianjin Feng, Lijun Lu, and Lei Cao. GCLR: A Self-Supervised Representation Learning Pretext Task for Glomerular Filtration Barrier Segmentation in TEM Images. *Artificial Intelligence in Medicine*, 146:102720, 2023.
- [408] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. CAT: Cross Attention in Vision Transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.
- [409] Jia-Ping Lin and Min-Te Sun. A YOLO-Based Traffic Counting System. In *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 82–85. IEEE, 2018.
- [410] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. *CoRR*, abs/1312.4400, 2013.
- [411] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. HRank: Filter Pruning Using High-Rank Feature Map. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2020.

-
- [412] Rui Lin, Ching-Yun Ko, Zhuolun He, Cong Chen, Yuan Cheng, Hao Yu, Graziano Chesi, and Ngai Wong. Hotcake: Higher Order Tucker Articulated Kernels for Deeper CNN Compression. In *2020 IEEE 15th International Conference on Solid-State & Integrated Circuit Technology (ICSICT)*, pages 1–4. IEEE, 2020.
- [413] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge Distillation via the Target-Aware Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10915–10924, 2022.
- [414] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.
- [415] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [416] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312, 2014.
- [417] Wenxiang Lin, Yan Ding, Hua-Liang Wei, Xinglin Pan, and Yutong Zhang. LD-SConv: Learned Depthwise Separable Convolutions by Group Pruning. *Sensors*, 20(15):4349, 2020.
- [418] W Johnson J Lindenstrauss and J Johnson. Extensions of Lipschitz Maps into a Hilbert Space. *Contemp. Math*, 26(189-206):2, 1984.
- [419] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind Image Super-Resolution: A Survey and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5461–5480, 2022.
- [420] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pinsky. Sparse Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–814, 2015.
- [421] Chenguang Liu, Guangshuai Gao, Ziyue Huang, Zhenghui Hu, Qingjie Liu, and Yunhong Wang. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images. *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [422] Chenxi Liu, Barret Zoph, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan L. Yuille, Jonathan Huang, and Kevin Murphy. Progressive Neural Architecture Search. *CoRR*, abs/1712.00559, 2017.

Bibliography

- [423] Dongnan Liu, Donghao Zhang, Yang Song, Heng Huang, and Weidong Cai. Panoptic Feature Fusion Net: A Novel Instance Segmentation Paradigm for Biomedical and Biological Images. *IEEE Transactions on Image Processing*, 30:2045–2059, 2021.
- [424] Dongnan Liu, Donghao Zhang, Yang Song, Chaoyi Zhang, Fan Zhang, Lauren O’Donnell, and Weidong Cai. Nuclei Segmentation via a Deep Panoptic Model with Semantic Feature Fusion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 861–868, 2019.
- [425] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation. *Advances in Neural Information Processing Systems*, 34:16266–16279, 2021.
- [426] Gang Liu, Yanxin Hu, Zhiyu Chen, Jianwei Guo, and Peng Ni. Lightweight Object Detection Algorithm for Robots with Improved YOLOv5. *Engineering Applications of Artificial Intelligence*, 123:106217, 2023.
- [427] Hou-I Liu, Yu-Wen Tseng, Kai-Cheng Chang, Pin-Jyun Wang, Hong-Han Shuai, and Wen-Huang Cheng. A DeNoising FPN with Transformer R-CNN for Tiny Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [428] Jiayu Liu, He Li, and Weihang Kong. Multi-Level Learning Counting via Pyramid Vision Transformer and CNN. *Engineering Applications of Artificial Intelligence*, 123:106184, 2023.
- [429] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, 128:261–318, 2020.
- [430] Liyang Liu, Shilong Zhang, Zhanghui Kuang, Aojun Zhou, Jing-Hao Xue, Xinjiang Wang, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Group Fisher Pruning for Practical Network Compression. In *International Conference on Machine Learning*, pages 7021–7032. PMLR, 2021.
- [431] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised Image-to-Image Translation Networks. *Advances in Neural Information Processing Systems*, 30, 2017.
- [432] Ruiping Liu, Kailun Yang, Alina Roitberg, Jiaming Zhang, Kunyu Peng, Huayao Liu, and Rainer Stiefelhagen. TransKD: Transformer Knowledge Distillation for Efficient Semantic Segmentation. *ArXiv Preprint ArXiv:2202.13393*, 2022.
- [433] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path Aggregation Network for Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.
- [434] Songtao Liu, Di Huang, et al. Receptive Field Block Net for Accurate and Fast Object Detection. In *European Conference on Computer Vision*, pages 385–400, 2018.

-
- [435] W Liu. ParseNet: Looking Wider to See Better. *ArXiv Preprint ArXiv:1506.04579*, 2015.
 - [436] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot Multibox Detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
 - [437] Xueyu Liu, Yongfei Wu, Yilin Chen, Dongna Hui, Jianan Zhang, Fang Hao, Yuanyue Lu, Hangbei Cheng, Yue Zeng, Weixia Han, et al. Diagnosis of Diabetic Kidney Disease in Whole Slide Images via AI-Driven Quantification of Pathological Indicators. *Computers in Biology and Medicine*, 166:107470, 2023.
 - [438] Yichen Liu, Huajian Zhang, and Daqing Gao. DiffYOLO: Object Detection for Anti-Noise via YOLO and Diffusion Models. *ArXiv Preprint ArXiv:2401.01659*, 2024.
 - [439] Yuchen Liu, Natasha Ong, Kaiyan Peng, Bo Xiong, Qifan Wang, Rui Hou, Madian Khabsa, Kaiyue Yang, David Liu, Donald S Williamson, et al. MMViT: Multiscale Multiview Vision Transformers. *ArXiv Preprint ArXiv:2305.00104*, 2023.
 - [440] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022.
 - [441] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *IEEE International Conference on Computer Vision*, pages 10012–10022, 2021.
 - [442] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-Training Quantization for Vision Transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021.
 - [443] Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal Likelihood Computation for Model Selection and Hypothesis Testing: An Extensive Review. *SIAM Review*, 65(1):3–58, 2023.
 - [444] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
 - [445] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *ArXiv Preprint ArXiv:1711.05101*, 2017.
 - [446] Junyan Lu, Chi Ma, Li Li, Xiaoyan Xing, Yong Zhang, Zhigang Wang, and Jiuwei Xu. A Vehicle Detection Method for Aerial Image Based on YOLO. *Journal of Computer and Communications*, 6(11):98–107, 2018.

Bibliography

- [447] Wanjie Lu, Chaozhen Lan, Chaoyang Niu, Wei Liu, Liang Lyu, Qunshan Shi, and Shiju Wang. A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:1211–1231, 2023.
- [448] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See More, Know More: Unsupervised Video Object Segmentation with Co-Attention Siamese Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3623–3632, 2019.
- [449] Xiaocong Lu, Jian Ji, Zhiqi Xing, and Qiguang Miao. Attention and Feature Fusion SSD for Remote Sensing Object Detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–9, 2021.
- [450] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tiejiong Zeng. Transformer for Single Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 457–466, 2022.
- [451] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 23033–23044. PMLR, 2023.
- [452] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression. In *IEEE International Conference on Computer Vision*, pages 5058–5066, 2017.
- [453] Mingnan Luo, Guihua Wen, Yang Hu, Dan Dai, and Yingxue Xu. Stochastic Region Pooling: Make Attention More Expressive. *Neurocomputing*, 409:119–130, 2020.
- [454] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, and Xiaoou Tang. Face Model Compression by Distilling Knowledge from Neurons. In *AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [455] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 4898–4906, 2016.
- [456] Jun Lyu, Guangyuan Li, Chengyan Wang, Chen Qin, Shuo Wang, Qi Dou, and Jing Qin. Region-Focused Multi-View Transformer-Based Generative Adversarial Network for Cardiac Cine MRI Reconstruction. *Medical Image Analysis*, 85:102760, 2023.

- [457] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *European Conference on Computer Vision*, pages 116–131, 2018.
- [458] Pingli Ma, Chen Li, Md Mamunur Rahaman, Yudong Yao, Jiawei Zhang, Shuojia Zou, Xin Zhao, and Marcin Grzegorzek. A State-of-the-Art Survey of Object Detection Techniques in Microorganism Image Analysis: From Classical Methods to Deep Learning Approaches. *Artificial Intelligence Review*, 56(2):1627–1698, 2023.
- [459] Ruijun Ma, Shuyi Li, Bob Zhang, Leyuan Fang, and Zhengming Li. Flexible and generalized real photograph denoising exploiting dual meta attention. *IEEE Transactions on Cybernetics*, 53(10):6395–6407, 2023.
- [460] Ruijun Ma, Shuyi Li, Bob Zhang, and Zhengming Li. Generative adaptive convolutions for real-world noisy image denoising. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1935–1943, Jun. 2022.
- [461] Ruijun Ma, Yaoxuan Zhang, Bob Zhang, Leyuan Fang, Dong Huang, and Long Qi. Learning attention in the frequency domain for flexible real photograph denoising. *IEEE Transactions on Image Processing*, 33:3707–3721, 2024.
- [462] Teli Ma, Mingyuan Mao, Honghui Zheng, Peng Gao, Xiaodi Wang, Shumin Han, Errui Ding, Baochang Zhang, and David Doermann. Oriented Object Detection with Transformer. *ArXiv Preprint ArXiv:2106.03146*, 2021.
- [463] Xiaolong Ma, Fu-Ming Guo, Wei Niu, Xue Lin, Jian Tang, Kaisheng Ma, Bin Ren, and Yanzhi Wang. PConv: The Missing but Desirable Sparsity in DNN Weight Pruning for Real-Time Execution on Mobile Devices. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5117–5124, 2020.
- [464] Xiaolong Ma, Sheng Lin, Shaokai Ye, Zhezhi He, Linfeng Zhang, Geng Yuan, Sia Huat Tan, Zhengang Li, Deliang Fan, Xuehai Qian, et al. Non-Structured DNN Weight Pruning—Is It Beneficial in Any Platform? *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4930–4944, 2021.
- [465] Wei Mao, Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. Panoptic Lintention Network: Towards Efficient Navigational Perception for the Visually Impaired. In *2021 IEEE International Conference on Real-Time Computing and Robotics (RCAR)*, pages 857–862. IEEE, 2021.
- [466] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. *Advances in Neural Information Processing Systems*, 29, 2016.
- [467] Geoffrey J McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.

Bibliography

- [468] Lotfi Abdelkrim Mecharbat, Hadjer Benmeziane, Hamza Ouarnoughi, and Smail Niar. HyT-NAS: Hybrid Transformers Neural Architecture Search for Edge Devices. In *2023 Workshop on Compilers, Deployment, and Tooling for Edge AI*, pages 41–45, 2023.
- [469] Sachin Mehta and Mohammad Rastegari. MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer. *ArXiv Preprint ArXiv:2110.02178*, 2021.
- [470] Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural Architecture Search Without Training. In *International Conference on Machine Learning*, pages 7588–7598. PMLR, 2021.
- [471] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *ArXiv Preprint ArXiv:2108.01073*, 2021.
- [472] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for Fast Training Convergence. In *IEEE International Conference on Computer Vision*, pages 3651–3660, 2021.
- [473] Fanxu Meng, Hao Cheng, Ke Li, Huixiang Luo, Xiaowei Guo, Guangming Lu, and Xing Sun. Pruning Filter in Filter. *Advances in Neural Information Processing Systems*, 33:17629–17640, 2020.
- [474] Pablo Meseguer, Rocío Del Amor, and Valery Naranjo. MICIL: Multiple-Instance Class-Incremental Learning for Skin Cancer Whole Slide Images. *Artificial Intelligence in Medicine*, 152:102870, 2024.
- [475] Paul Michel, Omer Levy, and Graham Neubig. Are Sixteen Heads Really Better Than One? *Advances in Neural Information Processing Systems*, 32, 2019.
- [476] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth International Conference on 3D vision (3DV)*, pages 565–571, 2016.
- [477] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved Knowledge Distillation via Teacher Assistant. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020.
- [478] Diganta Misra. Mish: A Self Regularized Non-Monotonic Activation Function. *ArXiv Preprint ArXiv:1908.08681*, 2019.
- [479] Shentong Mo, Zhun Sun, and Chao Li. Multi-Level Contrastive Learning for Self-Supervised Vision Transformers. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2778–2787, 2023.
- [480] Shakir Mohamed and Balaji Lakshminarayanan. Learning in Implicit Generative Models. *ArXiv Preprint ArXiv:1610.03483*, 2016.

-
- [481] Rohit Mohan and Abhinav Valada. EfficientPS: Efficient Panoptic Segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021.
 - [482] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.
 - [483] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion Models, Image Super-Resolution and Everything: A Survey. *ArXiv Preprint ArXiv:2401.00736*, 2024.
 - [484] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.
 - [485] Subhabrata Mukherjee and Ahmed Awadallah. XtremeDistil: Multi-Stage Distillation for Massive Multilingual Models. *ArXiv Preprint ArXiv:2004.05686*, 2020.
 - [486] Abdul Muqeet, Md Tauhid Bin Iqbal, and Sung-Ho Bae. Hybrid Residual Attention Network for Single Image Super Resolution. *ArXiv Preprint ArXiv:1907.05514*, 2019.
 - [487] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.
 - [488] Mahyar Najibi, Mohammad Rastegari, and Larry S Davis. G-CNN: An Iterative Grid Based Object Detector. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2369–2377, 2016.
 - [489] Usman Naseem, Matloob Khushi, and Jinman Kim. Vision-Language Transformer for Interpretable Pathology Visual Question Answering. *IEEE Journal of Biomedical and Health Informatics*, 27(4):1681–1690, 2022.
 - [490] S Naveen, MS S Ram Kiran, M Indupriya, TV Manikanta, and PV Sudeep. Transformer Models for Enhancing AttnGAN Based Text to Image Generation. *Image and Vision Computing*, 115:104284, 2021.
 - [491] Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated Generation of Accurate & Fluent Medical X-Ray Reports. *ArXiv Preprint ArXiv:2108.12126*, 2021.
 - [492] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

Bibliography

- [493] Ishan Nigam, Chen Huang, and Deva Ramanan. Ensemble Knowledge Transfer for Semantic Segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1499–1508. IEEE, 2018.
- [494] Ruigang Niu, Xian Sun, Yu Tian, Wenhui Diao, Kaiqiang Chen, and Kun Fu. Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.
- [495] Wei Niu, Xiaolong Ma, Sheng Lin, Shihao Wang, Xuehai Qian, Xue Lin, Yanzhi Wang, and Bin Ren. PatDNN: Achieving Real-Time DNN Execution on Mobile Devices with Pattern-Based Weight Pruning. In *Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 907–922, 2020.
- [496] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing*, 452:48–62, 2021.
- [497] Alexander Novikov, Dmitrii Podoprikin, Anton Osokin, and Dmitry P Vetrov. Tensorizing Neural Networks. *Advances in Neural Information Processing Systems*, 28, 2015.
- [498] Ini Oguntola, Subby Olubeko, and Christopher Sweeney. SlimNets: An Exploration of Deep Model Compression and Acceleration. In *2018 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.
- [499] Hyun-Jic Oh and Won-Ki Jeong. DiffMix: Diffusion Model-Based Data Synthesis for Nuclei Segmentation and Classification in Imbalanced Pathology Image Datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–345. Springer, 2023.
- [500] Naoki Okamoto, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Deep Ensemble Learning by Diverse Knowledge Distillation for Fine-Grained Object Classification. In *European Conference on Computer Vision*, pages 502–518. Springer, 2022.
- [501] Naoki Okamoto, Soma Minami, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Deep Ensemble Collaborative Learning by Using Knowledge-Transfer Graph for Fine-Grained Object Classification. *ArXiv Preprint ArXiv:2103.14845*, 2021.
- [502] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

-
- [503] Wanli Ouyang, Kun Wang, Xin Zhu, and Xiaogang Wang. Chained Cascade Network for Object Detection. In *IEEE International Conference on Computer Vision*, pages 1938–1946, 2017.
- [504] Haiwei Pan, Shuning He, Kejia Zhang, Bo Qu, Chunling Chen, and Kun Shi. AMAM: An Attention-Based Multimodal Alignment Model for Medical Visual Question Answering. *Knowledge-Based Systems*, 255:109763, 2022.
- [505] Xingjia Pan, Fan Tang, Weiming Dong, Yang Gu, Zhichao Song, Yiping Meng, Pengfei Xu, Oliver Deussen, and Changsheng Xu. Self-Supervised Feature Augmentation for Large Image Object Detection. *IEEE Transactions on Image Processing*, 29:6745–6758, 2020.
- [506] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the Integration of Self-Attention and Convolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–825, 2022.
- [507] Yanwei Pang, Manli Sun, Xiaoheng Jiang, and Xuelong Li. Convolution in Convolution for Network in Network. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1587–1597, 2017.
- [508] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: Bottleneck Attention Module. *ArXiv Preprint ArXiv:1807.06514*, 2018.
- [509] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [510] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *ArXiv Preprint ArXiv:1606.02147*, 2016.
- [511] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
- [512] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *IEEE International Conference on Computer Vision*, pages 4195–4205, 2023.
- [513] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep Snake for Real-Time Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8533–8542, 2020.
- [514] Tengyi Peng, Changqing Shen, Shilong Sun, and Dong Wang. Fault Feature Extractor Based on Bootstrap Your Own Latent and Data Augmentation Algorithm for Unlabeled Vibration Signals. *IEEE Transactions on Industrial Electronics*, 69(9):9547–9555, 2021.

Bibliography

- [515] Zhiliang Peng, Zonghao Guo, Wei Huang, Yaowei Wang, Lingxi Xie, Jianbin Jiao, Qi Tian, and Qixiang Ye. Conformer: Local Features Coupling Global Representations for Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8), 2023.
- [516] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local Features Coupling Global Representations for Visual Recognition. In *IEEE International Conference on Computer Vision*, pages 367–376, 2021.
- [517] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to Segment Object Candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.
- [518] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech, 2021.
- [519] Lorenzo Porzi, Samuel Rota Buló, Aleksander Colovic, and Peter Kotschieder. Seamless Scene Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8277–8286, 2019.
- [520] Lorenzo Porzi, Samuel Rota Buló, and Peter Kotschieder. Improving Panoptic Segmentation at All Scales. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7302–7311, 2021.
- [521] Tim Prangemeier, Christoph Reich, and Heinz Koeppl. Attention-Based Transformers for Instance Segmentation of Cells in Microstructures. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 700–707. IEEE, 2020.
- [522] Sai Prasanna, Anna Rogers, and Anna Rumshisky. When BERT Plays the Lottery, All Tickets Are Winning. *ArXiv Preprint ArXiv:2005.00561*, 2020.
- [523] Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. Fully Quantized Transformer for Machine Translation. *ArXiv Preprint ArXiv:1910.10485*, 2019.
- [524] Yifan Pu, Yiru Wang, Zhuofan Xia, Yizeng Han, Yulin Wang, Weihao Gan, Zidong Wang, Shiji Song, and Gao Huang. Adaptive Rotated Convolution for Rotated Object Detection. In *IEEE International Conference on Computer Vision*, pages 6589–6600, 2023.
- [525] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.
- [526] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

- [527] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10213–10224, 2021.
- [528] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency Channel Attention Networks. In *IEEE International Conference on Computer Vision*, pages 783–792, 2021.
- [529] Tran Minh Quan, David GC Hildebrand, and Won-Ki Jeong. FusionNet: A Deep Fully Residual Convolutional Neural Network for Image Segmentation in Connectomics. *ArXiv Preprint ArXiv:1612.05360*, 2016.
- [530] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
- [531] Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, and Piotr Dollár. On Network Design Spaces for Visual Recognition. In *IEEE International Conference on Computer Vision*, pages 1882–1890, 2019.
- [532] Aman Raj, Daniel Maturana, and Sebastian Scherer. Multi-Scale Convolutional Architecture for Semantic Segmentation. *Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RITR-15-21*, 2015.
- [533] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-Alone Self-Attention in Vision Models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [534] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.
- [535] Qiong Ran, Qing Wang, Boya Zhao, Yuanfeng Wu, Shengliang Pu, and Zijin Li. Lightweight Oriented Object Detection Using Multiscale Context and Enhanced Channel Attention in Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5786–5795, 2021.
- [536] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. TGFuse: An Infrared and Visible Image Fusion Approach Based on Transformer and Generative Adversarial Network. *IEEE Transactions on Image Processing*, 2023.

Bibliography

- [537] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- [538] Ryan Razani, Grégoire Morin, Eyyub Sari, and Vahid Partovi Nia. Adaptive Binary-Ternary Quantization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4613–4618, 2021.
- [539] Arthur J Redfern, Lijun Zhu, and Molly K Newquist. BCNN: A Binary CNN with All Matrix Ops Quantized to 1 Bit Precision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2021.
- [540] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [541] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. *ArXiv Preprint*, 2017.
- [542] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *ArXiv Preprint ArXiv:1804.02767*, 2018.
- [543] Haotian Ren, Wei Lu, Yun Xiao, Xiaojun Chang, Xuanhong Wang, Zhiqiang Dong, and Dingyi Fang. Graph convolutional networks in language and vision: A survey. *Knowledge-Based Systems*, 251:109250, 2022.
- [544] Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, Qing Du, Xiaodan Liang, and Xiaojun Chang. Beyond Fixation: Dynamic Window Visual Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11987–11997, 2022.
- [545] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.
- [546] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.
- [547] Pau Rodriguez, Diego Velazquez, Guillem Cucurull, Josep M Gonfaus, F Xavier Roca, and Jordi Gonzalez. Pay Attention to the Activations: A Modular Attention Mechanism for Fine-Grained Image Recognition. *IEEE Transactions on Multimedia*, 22(2):502–514, 2019.
- [548] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

-
- [549] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. *ArXiv Preprint ArXiv:1412.6550*, 2014.
- [550] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19574–19584, 2023.
- [551] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
- [552] Frank Rosenblatt. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6):386, 1958.
- [553] Frank Rosenblatt. Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo, NY, 1961.
- [554] T-YLPG Ross and GKHP Dollár. Focal Loss for Dense Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017.
- [555] Anirban Roy and Sinisa Todorovic. A Multi-Scale CNN for Affordance Segmentation in RGB Images. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016.
- [556] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token Contrast for Weakly-Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023.
- [557] Yossi Rubner and Carlo Tomasi. Earth mover’s distances: theory and applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000.
- [558] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning Representations by Back-Propagating Errors. *Nature*, 323(6088):533, 1986.
- [559] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic Routing Between Capsules. *Advances in Neural Information Processing Systems*, 30, 2017.
- [560] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-Image Diffusion Models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

Bibliography

- [561] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.
- [562] Massimo Salvi, Alessandro Mogetta, Alessandro Gambella, Luca Molinaro, Antonella Barreca, Mauro Papotti, and Filippo Molinari. Automated Assessment of Glomerulosclerosis and Tubular Atrophy Using Deep Learning. *Computerized Medical Imaging and Graphics*, 90:101930, 2021.
- [563] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [564] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017.
- [565] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks. *ArXiv Preprint ArXiv:1312.6229*, 2013.
- [566] Sajjad Shahabodini, Mobina Mansoori, Farnoush Bayatmakou, Jamshid Abouei, Konstantinos N. Plataniotis, and Arash Mohammadi. The Missing Point in Vision Transformers for Universal Image Segmentation, 2025.
- [567] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. SwiftFormer: Efficient Additive Attention for Transformer-Based Real-Time Mobile Vision Applications. In *IEEE International Conference on Computer Vision*, pages 17425–17436, 2023.
- [568] Ying Shan, Harpreet S Sawhney, and Rakesh Kumar. Unsupervised Learning of Discriminative Edge Measures for Vehicle Matching Between Nonoverlapping Cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):700–711, 2008.
- [569] Manish Sharma, Panos P Markopoulos, Eli Saber, M Salman Asif, and Ashley Prater-Bennette. Convolutional Auto-Encoder with Tensor-Train Factorization. In *IEEE International Conference on Computer Vision*, pages 198–206, 2021.
- [570] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. *ArXiv Preprint ArXiv:1803.02155*, 2018.
- [571] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient Attention: Attention with Linear Complexities. In *IEEE Winter Conference on Applications of Computer Vision*, pages 3531–3539, 2021.

-
- [572] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.
- [573] Benjamin Shickel, Nicholas Lucarelli, Adish Rao, Donghwan Yun, Kyung Chul Moon, Han Seung Seok, and Pinaki Sarder. Spatially Aware Transformer Networks for Contextual Prediction of Diabetic Nephropathy Progression from Whole Slide Images. In *Medical Imaging 2023: Digital and Computational Pathology*, volume 12471, pages 129–140. SPIE, 2023.
- [574] Nir Shlezinger, Santiago Segarra, Yi Zhang, Dvir Avrahami, Zohar Davidov, Tirza Routtenberg, and Yonina C. Eldar. Deep unfolding: Recent developments, theory, and design guidelines, 2025.
- [575] Kumar Shridhar, Harshil Jain, Akshat Agarwal, and Denis Kleyko. End to End Binarized Neural Networks for Text Classification. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 29–34, 2020.
- [576] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.
- [577] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.
- [578] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, 2021.
- [579] David So, Quoc Le, and Chen Liang. The Evolved Transformer. In *International Conference on Machine Learning*, pages 5877–5886. PMLR, 2019.
- [580] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. AdaptIS: Adaptive Instance Selection Network. In *IEEE International Conference on Computer Vision*, pages 7355–7363, 2019.
- [581] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Bibliography

- [582] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder Variational Autoencoders. *Advances in Neural Information Processing Systems*, 29, 2016.
- [583] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. *ArXiv Preprint ArXiv:2010.02502*, 2020.
- [584] Kaiyou Song, Jin Xie, Shan Zhang, and Zimeng Luo. Multi-Mode Online Knowledge Distillation for Self-Supervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11848–11857, 2023.
- [585] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32, 2019.
- [586] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling Through Stochastic Differential Equations. *ArXiv Preprint ArXiv:2011.13456*, 2020.
- [587] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer Tracking with Cyclic Shifting Window Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8791–8800, 2022.
- [588] Ziyang Song, Lin Liu, Feiyang Jia, Yadan Luo, Caiyan Jia, Guoxin Zhang, Lei Yang, and Li Wang. Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook. *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [589] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012.
- [590] Daniel A Spielman and Shang-Hua Teng. Spectral Sparsification of Graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.
- [591] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep Neural Network Models for Computational Histopathology: A Survey. *Medical Image Analysis*, 67:101813, 2021.
- [592] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck Transformers for Visual Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021.
- [593] Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. Hierarchical X-Ray Report Generation via Pathology Tags and Multi Head Attention. In *Asian Conference on Computer Vision*, 2020.

-
- [594] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [595] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. In *IEEE International Conference on Computer Vision*, pages 7262–7272, 2021.
- [596] Xiu Su, Shan You, Jiyang Xie, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. ViTAS: Vision Transformer Architecture Search. In *European Conference on Computer Vision*, pages 139–157. Springer, 2022.
- [597] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context Decoupling Augmentation for Weakly Supervised Semantic Segmentation. In *IEEE International Conference on Computer Vision*, pages 7004–7014, 2021.
- [598] Zhuo Su, Linpu Fang, Wenxiong Kang, Dewen Hu, Matti Pietikäinen, and Li Liu. Dynamic Group Convolution for Accelerating Convolutional Neural Networks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 138–155. Springer, 2020.
- [599] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *IEEE International Conference on Computer Vision*, pages 843–852, 2017.
- [600] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. Ecs-Net: Improving Weakly Supervised Semantic Segmentation by Using Connections Between Class Activation Maps. In *IEEE International Conference on Computer Vision*, pages 7283–7292, 2021.
- [601] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-Attention Multi-Class Constraint for Fine-Grained Image Recognition. In *European Conference on Computer Vision*, pages 805–821, 2018.
- [602] Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, and Ping Luo. What Makes for End-to-End Object Detection? In *International Conference on Machine Learning*, pages 9934–9944. PMLR, 2021.
- [603] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. SparseR-CNN: End-to-End Object Detection with Learnable Proposals. *ArXiv Preprint ArXiv:2011.12450*, 2020.
- [604] Shuyang Sun, Xiaoyu Yue, Hengshuang Zhao, Philip HS Torr, and Song Bai. Patch-Based Separable Transformer for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9241–9247, 2022.

Bibliography

- [605] Wei Sun, AoJun Zhou, Sander Stuijk, Rob Wijnhoven, Andrew O Nelson, Henk Corporaal, et al. DominoSearch: Find Layer-Wise Fine-Grained N: M Sparse Schemes from Dense Neural Networks. *Advances in Neural Information Processing Systems*, 34:20721–20732, 2021.
- [606] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking Transformer-Based Set Prediction for Object Detection. In *IEEE International Conference on Computer Vision*, pages 3611–3620, 2021.
- [607] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-V4, Inception-ResNet and the Impact of Residual Connections on Learning. *CoRR*, abs/1602.07261, 2016.
- [608] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *CoRR*, abs/1409.4842, 2014.
- [609] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep Neural Networks for Object Detection. In *Advances in Neural Information Processing Systems*, pages 2553–2561, 2013.
- [610] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CoRR*, abs/1512.00567, 2015.
- [611] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.
- [612] Mingxing Tan and Quoc V Le. MixConv: Mixed Depthwise Convolutional Kernels. *ArXiv Preprint ArXiv:1907.09595*, 2019.
- [613] Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet: Scalable and Efficient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020.
- [614] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch Slimming for Efficient Vision Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022.
- [615] Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhijun Tu, Kai Han, Hailin Hu, and Dacheng Tao. A Survey on Transformer Compression. *ArXiv Preprint ArXiv:2402.05964*, 2024.
- [616] Yehui Tang, Yunhe Wang, Yixing Xu, Yiping Deng, Chao Xu, Dacheng Tao, and Chang Xu. Manifold Regularized Dynamic Network Pruning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5028, 2021.

-
- [617] Ajay K Tanwani, Joelle Barral, and Daniel Freedman. RepsNet: Combining Vision with Language for Automated Medical Reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 714–724. Springer, 2022.
- [618] Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse Sinkhorn Attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR, 2020.
- [619] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. XrayGPT: Chest Radiographs Summarization Using Medical Vision-Language Models. *ArXiv Preprint ArXiv:2306.07971*, 2023.
- [620] Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster Gaze Prediction with Dense Networks and Fisher Pruning. *ArXiv Preprint ArXiv:1801.05787*, 2018.
- [621] Long Tian, Zhigang Tu, Dejun Zhang, Jun Liu, Baoxin Li, and Junsong Yuan. Unsupervised Learning of Optical Flow with CNN-Based Non-Local Filtering. *IEEE Transactions on Image Processing*, 29:8429–8442, 2020.
- [622] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional Convolutions for Instance Segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020.
- [623] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1922–1933, 2020.
- [624] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image Super-Resolution Using Dense Skip Connections. In *IEEE International Conference on Computer Vision*, pages 4799–4807, 2017.
- [625] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. UVCAN: UNet Vision Transformer Cycle-Consistent GAN for Unpaired Image-to-Image Translation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 702–712, 2023.
- [626] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training Data-Efficient Image Transformers & Distillation Through Attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.
- [627] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going Deeper with Image Transformers. In *IEEE International Conference on Computer Vision*, pages 32–42, 2021.

Bibliography

- [628] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.
- [629] Minh-Ngoc Tran, Trong-Nghia Nguyen, and Viet-Hung Dao. A Practical Tutorial on Variational Bayes. *ArXiv Preprint ArXiv:2103.01327*, 2021.
- [630] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11017, 2023.
- [631] Jingzhi Tu, Gang Mei, Zhengjing Ma, and Francesco Piccialli. SWCGAN: Generative Adversarial Network Combining Swin Transformer and CNN for Remote Sensing Image Super-Resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5662–5673, 2022.
- [632] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-Axis Vision Transformer. In *European Conference on Computer Vision*, pages 459–479. Springer, 2022.
- [633] Waseem Ullah, Tanveer Hussain, Fath U Min Ullah, Mi Young Lee, and Sung Wook Baik. TransCNN: Hybrid CNN and Transformer Mechanism for Surveillance Anomaly Detection. *Engineering Applications of Artificial Intelligence*, 123:106173, 2023.
- [634] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *ArXiv Preprint ArXiv:1607.08022*, 2016.
- [635] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 36–46. Springer, 2021.
- [636] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022.
- [637] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional Image Generation with PixelCNN Decoders. *Advances in Neural Information Processing Systems*, 29, 2016.
- [638] Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data Using T-SNE. *Journal of Machine Learning Research*, 9(11), 2008.
- [639] Vincent Vanhoucke, Andrew Senior, Mark Z Mao, et al. Improving the Speed of Neural Networks on CPUs. In *Proceedings of the Deep Learning and Unsupervised Feature Learning NIPS Workshop*, volume 1, page 4, 2011.

-
- [640] Ali Varamesh, Ali Diba, Tinne Tuytelaars, and Luc Van Gool. Self-Supervised Ranking for Representation Learning. *ArXiv Preprint ArXiv:2010.07258*, 2020.
- [641] Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Towards Histopathological Stain Invariance by Unsupervised Domain Augmentation Using Generative Adversarial Networks. *Neurocomputing*, 460:277–291, 2021.
- [642] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [643] Thomas Verelst and Tinne Tuytelaars. Dynamic Convolutions: Exploiting Spatial Sparsity for Faster Inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2320–2329, 2020.
- [644] Shourya Verma. Self-Trained Panoptic Segmentation. *ArXiv Preprint ArXiv:2311.10648*, 2023.
- [645] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *25th International Conference on Machine Learning*, pages 1096–1103, 2008.
- [646] Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. A Question-Centric Model for Visual Question Answering in Medical Imaging. *IEEE Transactions on Medical Imaging*, 39(9):2856–2868, 2020.
- [647] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. YOLOv10: Real-Time End-to-End Object Detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.
- [648] Chaofeng Wang, Zheng Li, and Jun Shi. Lightweight Image Super-Resolution with Adaptive Weighted Learning Network. *ArXiv Preprint ArXiv:1904.02358*, 2019.
- [649] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *ArXiv Preprint ArXiv:2402.13616*, 2024.
- [650] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
- [651] Dingheng Wang, Guangshe Zhao, Hengnu Chen, Zhexian Liu, Lei Deng, and Guoqi Li. Nonlinear Tensor Train Format for Deep Neural Network Compression. *Neural Networks*, 144:320–333, 2021.

Bibliography

- [652] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual Attention Network for Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.
- [653] Hefeng Wang, Jiale Cao, Rao Muhammad Anwer, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. DFormer: Diffusion-Guided Transformer for Universal Image Segmentation. *ArXiv Preprint ArXiv:2306.03437*, 2023.
- [654] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021.
- [655] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020.
- [656] Jianyong Wang, Mingliang Gao, Wenzhe Zhai, Imad Rida, Xianxun Zhu, and Qilei Li. Knowledge generation and distillation for road segmentation in intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13, 2025.
- [657] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.
- [658] Li Wang, Jiale Jia, and Hualin Dai. OrientedDiffDet: Diffusion Model for Oriented Object Detection in Aerial Images. *Applied Sciences*, 14(5):2000, 2024.
- [659] Liya Wang and Alex Tien. Aerial Image Object Detection with Vision Transformer Detector (ViTDet). In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 6450–6453. IEEE, 2023.
- [660] Naigang Wang, Chi-Chun Charlie Liu, Swagath Venkataramani, Sanchari Sen, Chia-Yu Chen, Kaoutar El Maghraoui, Vijayalakshmi Viji Srinivasan, and Leland Chang. Deep Compression of Pre-Trained Transformer Models. *Advances in Neural Information Processing Systems*, 35:14140–14154, 2022.
- [661] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding Convolution for Semantic Segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460. IEEE, 2018.
- [662] Peisong Wang and Jian Cheng. Accelerating Convolutional Neural Networks for Mobile Applications. In *24th ACM International Conference on Multimedia*, pages 541–545, 2016.

-
- [663] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. KVT: K-NN Attention for Boosting Vision Transformers. In *European Conference on Computer Vision*, pages 285–302. Springer, 2022.
- [664] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning Deep Transformer Models for Machine Translation. *ArXiv Preprint ArXiv:1906.01787*, 2019.
- [665] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020.
- [666] Rui Wang, Chaopeng Guo, Mohammad Shabaz, Imad Rida, Erik Cambria, and Xianxun Zhu. Cime: Contextual interaction-based multimodal emotion analysis with enhanced semantic information. *IEEE Transactions on Computational Social Systems*, pages 1–11, 2025.
- [667] Rui Wang, Yaoyang Wang, Erik Cambria, Xuhui Fan, Xiaohan Yu, Yao Huang, Xiang E, and Xianxun Zhu. Contrastive-based removal of negative information in multimodal emotion analysis. *Cogn. Comput.*, 17(3):107, 2025.
- [668] Rui Wang, Duyun Xu, Lucia Cascone, Yaoyang Wang, Hui Chen, Jianbo Zheng, and Xianxun Zhu. Raft: Robust adversarial fusion transformer for multimodal sentiment analysis. *Array*, 27:100445, 2025.
- [669] Rui Wang, Jiawei Zhu, Shoujin Wang, Tao Wang, Jingze Huang, and Xianxun Zhu. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. *Int. J. Multimed. Inf. Retr.*, 13(4):39, 2024.
- [670] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-Attention with Linear Complexity. *ArXiv Preprint ArXiv:2006.04768*, 2020.
- [671] Wei Wang, Chen Tang, Xin Wang, and Bin Zheng. A ViT-Based Multiscale Feature Fusion Approach for Remote Sensing Image Segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.
- [672] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023.
- [673] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In *IEEE International Conference on Computer Vision*, pages 568–578, 2021.

Bibliography

- [674] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT V2: Improved Baselines with Pyramid Vision Transformer. *Computational Visual Media*, 8(3):415–424, 2022.
- [675] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.
- [676] Wenxiao Wang, Wei Chen, Qibo Qiu, Long Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wei Liu. CrossFormer++: A Versatile Vision Transformer Hinging on Cross-Scale Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [677] Xiao Wang, Haijiang Zhu, and Zhiqing Li. DRIPNet: Decoupled Region of Interest Pooling Feature Network Based on Diffusion Model for UAV Small Object Detection. *ResearchSquare*, 2024.
- [678] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-Local Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [679] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting Objects by Locations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 649–665. Springer, 2020.
- [680] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and Fast Instance Segmentation. *Advances in Neural Information Processing Systems*, 33:17721–17732, 2020.
- [681] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *European Conference on Computer Vision Workshops*, pages 0–0, 2018.
- [682] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor DETR: Query Design for Transformer-Based Detector. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 2567–2575, 2022.
- [683] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.
- [684] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024.

-
- [685] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A General U-Shaped Transformer for Image Restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.
- [686] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep Learning for Image Super-Resolution: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2020.
- [687] Zhizhong Wang, Lei Zhao, and Wei Xing. StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models. In *IEEE International Conference on Computer Vision*, pages 7677–7689, 2023.
- [688] Ziwei Wang, Han Xiao, Jiwen Lu, and Jie Zhou. Generalizable Mixed-Precision Quantization via Attribution Rank Preservation. In *IEEE International Conference on Computer Vision*, pages 5291–5300, 2021.
- [689] Maurice Weiler and Gabriele Cesa. General E (2)-Equivariant Steerable CNNs. *Advances in Neural Information Processing Systems*, 32, 2019.
- [690] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. *Advances in Neural Information Processing Systems*, 31, 2018.
- [691] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *European Conference on Computer Vision*, pages 3–19, 2018.
- [692] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual Transformers: Token-Based Image Representation and Processing for Computer Vision. *ArXiv Preprint ArXiv:2006.03677*, 2020.
- [693] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *IEEE International Conference on Computer Vision*, pages 22–31, 2021.
- [694] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fast-FCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation. *ArXiv Preprint ArXiv:1903.11816*, 2019.
- [695] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized Convolutional Neural Networks for Mobile Devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016.
- [696] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and Improving Relative Position Encoding for Vision Transformer. In *IEEE International Conference on Computer Vision*, pages 10033–10041, 2021.

Bibliography

- [697] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast Pretraining Distillation for Small Vision Transformers. In *European Conference on Computer Vision*, pages 68–85. Springer, 2022.
- [698] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded Discriminative Attention Mechanism for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, 2021.
- [699] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. StyleFormer: Real-Time Arbitrary Style Transfer via Parametric Style Composition. In *IEEE International Conference on Computer Vision*, pages 14618–14627, 2021.
- [700] Yifan Wu, Shichao Kan, Min Zeng, and Min Li. SingularFormer: Learning to Decompose Self-Attention to Linearize the Complexity of Transformer. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4433–4441, 2023.
- [701] Yuxin Wu and Kaiming He. Group Normalization. In *European Conference on Computer Vision*, pages 3–19, 2018.
- [702] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [703] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite Transformer with Long-Short Range Attention. *ArXiv Preprint ArXiv:2004.11886*, 2020.
- [704] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. DiffI2I: Efficient Diffusion Model for Image-to-Image Translation. *ArXiv Preprint ArXiv:2308.13767*, 2023.
- [705] Mengfei Xia, Yu Zhou, Ran Yi, Yong-Jin Liu, and Wenping Wang. A Diffusion Model Translator for Efficient Image-to-Image Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [706] Zhengyu Xia and Joohee Kim. Mixed Spatial Pyramid Pooling for Semantic Segmentation. *Applied Soft Computing*, 91:106209, 2020.
- [707] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision Transformer with Deformable Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022.
- [708] Peng-Fei Xian, Lai-Man Po, Jing-Jing Xiong, Yu-Zhi Zhao, Wing-Yin Yu, and Kwok-Wai Cheung. Mask-Pyramid Network: A Novel Panoptic Segmentation Method. *Sensors*, 24(5), 2024.
- [709] Junyi Xiang, Xianxun Zhu, and Erik Cambria. Integrating audio-visual text generation with contrastive learning for enhanced multimodal emotion analysis. *Information Fusion*, 127:103809, 2026.

-
- [710] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *European Conference on Computer Vision*, pages 418–434, 2018.
- [711] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early Convolutions Help Transformers See Better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021.
- [712] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.
- [713] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross Language Image Matching for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4483–4492, 2022.
- [714] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. Contrastive Learning of Class-Agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation. *ArXiv Preprint ArXiv:2203.13505*, 2022.
- [715] Jiyang Xie, Zhanyu Ma, Dongliang Chang, Guoqiang Zhang, and Jun Guo. Gpca: A Probabilistic Framework for Gaussian Process Embedded Channel Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8230–8248, 2021.
- [716] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5987–5995. IEEE, 2017.
- [717] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.
- [718] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for Object Detection. In *IEEE International Conference on Computer Vision*, pages 3520–3529, 2021.
- [719] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-Supervised Learning with Swin Transformers. *ArXiv Preprint ArXiv:2105.04553*, 2021.
- [720] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

Bibliography

- [721] Zheng Xie, Zhiquan Wen, Jing Liu, Zhiqiang Liu, Xixian Wu, and Mingkui Tan. Deep Transferring Quantization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 625–642. Springer, 2020.
- [722] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 14138–14148, 2021.
- [723] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, et al. Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5652–5661, 2024.
- [724] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. UPSNet: A Unified Panoptic Segmentation Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019.
- [725] Zhan Xiong, Junling He, Pieter Valkema, Tri Q. Nguyen, Maarten Naesens, Jesper Kers, and Fons J. Verbeek. Advances in Kidney Biopsy Lesion Assessment Through Dense Instance Segmentation. *Artificial Intelligence in Medicine*, 164:103111, 2025.
- [726] Zhan Xiong, Xiaoqin Tang, and Fons J. Verbeek. A Survey of Deep Learning in Image Processing: Developments, Applications and Challenges. *Artificial Intelligence Review (Under Review)*, 2026.
- [727] Zhan Xiong, Xiaoqin Tang, and Fons J. Verbeek. VGMEA: Interpretable Cross-Attention Mechanisms in Semantic Segmentation via Variational Gaussian Model Mixture. *ArXiv Preprint*, 2026.
- [728] Zhan Xiong and Fons J. Verbeek. Segmentation of Zebrafish Larva Inhomogeneous 3D Images Using the Level-Set Method. In David L. Kao, Thomas Wischgoll, and Song Zhang, editors, *Visualization and Data Analysis 2016, San Francisco, California, USA, February 14–18, 2016*, pages 1–9. Society for Imaging Science and Technology, 2016.
- [729] Zhan Xiong and Fons J. Verbeek. Weakly Supervised Semantic Segmentation via Contextual Cross-Attention. In *9th International Conference on Frontiers of Signal Processing, ICFSP 2024, Paris, France, September 12–14, 2024*, pages 83–92. IEEE, 2024.
- [730] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical Evaluation of Rectified Activations in Convolutional Network. *ArXiv Preprint ArXiv:1505.00853*, 2015.

-
- [731] Jianjin Xu, Zhaoxiang Zhang, and Xiaolin Hu. Extracting Semantic Knowledge from GANs with Unsupervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9654–9668, 2023.
- [732] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.
- [733] Jingzhao Xu, Mengke Yuan, Dong-Ming Yan, and Tieru Wu. Deep unfolding multi-scale regularizer network for image denoising. *Computational Visual Media*, 9(2):335–350, 2023.
- [734] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-Class Token Transformer for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022.
- [735] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, and Baohua Lai. PP-YOLOE: An Evolved Version of YOLO, 2022.
- [736] Sheng Xu, Anran Huang, Lei Chen, and Baochang Zhang. Convolutional Neural Network Pruning: A Survey. In *2020 39th Chinese Control Conference (CCC)*, pages 7458–7463. IEEE, 2020.
- [737] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-Scale Conv-Attentional Image Transformers. In *IEEE International Conference on Computer Vision*, pages 9981–9990, 2021.
- [738] Yan Xu, Tao Mo, Qiwei Feng, Peilin Zhong, Maode Lai, I Eric, and Chao Chang. Deep Learning of Feature Representation with Multiple Instance Learning for Medical Image Analysis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1626–1630. IEEE, 2014.
- [739] Yixing Xu, Yunhe Wang, Kai Han, Yehui Tang, Shangling Jui, Chunjing Xu, and Chang Xu. ReNAS: Relativistic Evaluation of Neural Architecture Search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4411–4420, 2021.
- [740] Zhefan Xu, Xiaoyang Zhan, Yumeng Xiu, Christopher Suzuki, and Kenji Shimada. Onboard Dynamic-Object Detection and Tracking for Autonomous Robot Navigation with RGB-D Camera. *IEEE Robotics and Automation Letters*, 9(1):651–658, 2023.
- [741] Kohei Yamamoto. Learnable Companding Quantization for Accurate Low-Bit Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5029–5038, 2021.

Bibliography

- [742] Zhicheng Yan, Xiaoliang Dai, Peizhao Zhang, Yuandong Tian, Bichen Wu, and Matt Feiszli. FP-NAS: Fast Probabilistic Neural Architecture Search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15139–15148, 2021.
- [743] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. *Advances in Neural Information Processing Systems*, 32, 2019.
- [744] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2022.
- [745] Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. Clustered Object Detection in Aerial Images. In *IEEE International Conference on Computer Vision*, pages 8311–8320, 2019.
- [746] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning Texture Transformer Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.
- [747] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal Self-Attention for Local-Global Interactions in Vision Transformers. *ArXiv Preprint ArXiv:2107.00641*, 2021.
- [748] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19141–19151, 2022.
- [749] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3289–3299, 2020.
- [750] Shaopeng Yang, Weiyu Guo, and Yuheng Ren. CrowdFormer: An Overlap Patching Vision Transformer for Top-Down Crowd Counting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, volume 1, page 2, 2022.
- [751] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.
- [752] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. DeeperLab: Single-Shot Image Parser. *ArXiv Preprint ArXiv:1902.05093*, 2019.

-
- [753] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 3163–3171, 2021.
- [754] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In *International Conference on Machine Learning*, pages 11830–11841. PMLR, 2021.
- [755] Xue Yang, Gefan Zhang, Xiaojiang Yang, Yue Zhou, Wentao Wang, Jin Tang, Tao He, and Junchi Yan. Detecting Rotated Objects as Gaussian Distributions and Its 3-D Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4335–4354, 2022.
- [756] Xue Yang, Yue Zhou, Gefan Zhang, Jirui Yang, Wentao Wang, Junchi Yan, XIAOPENG ZHANG, and Qi Tian. The KFIOU Loss for Rotated Object Detection. In *The Eleventh International Conference on Learning Representations*, 2023.
- [757] Yadong Yang, Xiaofeng Wang, Quan Zhao, and Tingting Sui. Two-Level Attentions and Grouping Attention Convolutional Network for Fine-Grained Image Classification. *Applied Sciences*, 9(9):1939, 2019.
- [758] Yibo Yang, Zhisheng Zhong, Tiancheng Shen, and Zhouchen Lin. Convolutional Neural Networks with Alternately Updated Clique. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2413–2422, 2018.
- [759] Zhaohui Yang, Yunhe Wang, Kai Han, Chunjing Xu, Chao Xu, Dacheng Tao, and Chang Xu. Searching for Low-Bit Weights in Quantized Neural Networks. *Advances in Neural Information Processing Systems*, 33:4091–4102, 2020.
- [760] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, W. Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *ArXiv*, abs/2209.09407, 2022.
- [761] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- [762] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, et al. Radar-Camera Fusion for Object Detection and Semantic Segmentation in Autonomous Driving: A Comprehensive Review. *IEEE Transactions on Intelligent Vehicles*, 2023.

Bibliography

- [763] Ting Yao, Yehao Li, Yingwei Pan, Yu Wang, Xiao-Ping Zhang, and Tao Mei. Dual Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10870–10882, 2023.
- [764] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient DETR: Improving End-to-End Object Detector with Dense Prior. *ArXiv Preprint ArXiv:2104.01318*, 2021.
- [765] Jinmian Ye, Guangxi Li, Di Chen, Haiqin Yang, Shandian Zhe, and Zenglin Xu. Block-Term Tensor Neural Networks. *Neural Networks*, 130:11–21, 2020.
- [766] Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 728–737. IEEE, 2019.
- [767] Donggeun Yoo, Sunggyun Park, Joon-Young Lee, Anthony S Paek, and In So Kweon. AttentionNet: Aggregating Weak Directions for Accurate Object Detection. In *IEEE International Conference on Computer Vision*, pages 2659–2667, 2015.
- [768] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial Erasing Framework via Triplet with Gated Pyramid Pooling Layer for Weakly Supervised Semantic Segmentation. In *European Conference on Computer Vision*, pages 326–344. Springer, 2022.
- [769] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *European Conference on Computer Vision*, pages 325–341, 2018.
- [770] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *ArXiv Preprint ArXiv:1511.07122*, 2015.
- [771] Hongyuan Yu, Cheng Wan, Mengchen Liu, Dongdong Chen, Bin Xiao, and Xiyang Dai. Real-Time Image Segmentation via Hybrid Convolutional-Transformer Architecture Search. *ArXiv Preprint ArXiv:2403.10413*, 2024.
- [772] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. CMT-DeepLab: Clustering Mask Transformers for Panoptic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2560–2570, 2022.
- [773] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. K-Means Mask Transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022.
- [774] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means Mask Transformer. In *ECCV*, 2022.

-
- [775] Ruichi Yu, Ang Li, Chun-Fu Chen, Jui-Hsin Lai, Vlad I Morariu, Xintong Han, Mingfei Gao, Ching-Yung Lin, and Larry S Davis. NISP: Pruning Networks Using Neuron Importance Score Propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9194–9203, 2018.
- [776] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale Match for Tiny Person Detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1257–1265, 2020.
- [777] Zhihong Yu, Feifei Lee, and Qiu Chen. HCT-Net: Hybrid CNN-Transformer Model Based on a Neural Architecture Search Network for Medical Image Segmentation. *Applied Intelligence*, 53(17):19990–20006, 2023.
- [778] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic Radiology Report Generation Based on Multi-View Image Fusion and Medical Concept Enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019.
- [779] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating Convolution Designs into Visual Transformers. In *IEEE International Conference on Computer Vision*, pages 579–588, 2021.
- [780] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In *IEEE International Conference on Computer Vision*, pages 558–567, 2021.
- [781] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central Similarity Quantization for Efficient Image and Video Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2020.
- [782] Mingze Yuan, Yingda Xia, Hexin Dong, Zifan Chen, Jiawen Yao, Mingyan Qiu, Ke Yan, Xiaoli Yin, Yu Shi, Xin Chen, et al. Devil Is in the Queries: Advancing Mask Transformers for Real-World Medical Image Segmentation and Out-of-Distribution Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 23879–23889, 2023.
- [783] X Yuan, Z Zheng, Y Li, X Liu, L Liu, X Li, Q Hou, and MM Cheng. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. ArXiv 2025. *ArXiv Preprint ArXiv:2501.03775*, 2025.
- [784] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-Contextual Representations for Semantic Segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.

Bibliography

- [785] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. OCNet: Object Context Network for Scene Parsing. *ArXiv Preprint ArXiv:1809.00916*, 2018.
- [786] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. RAMM: Retrieval-Augmented Biomedical Visual Question Answering with Multi-Modal Pre-Training. In *31st ACM International Conference on Multimedia*, pages 547–556, 2023.
- [787] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient Diffusion Model for Image Super-Resolution by Residual Shifting. *Advances in Neural Information Processing Systems*, 36, 2024.
- [788] Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O (N) Connections Are Expressive Enough: Universal Approximability of Sparse Transformers. *Advances in Neural Information Processing Systems*, 33:13783–13794, 2020.
- [789] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-Level Representation Learning for Self-Supervised Vision Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8354–8363, 2022.
- [790] Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. A Multipath Network for Object Detection. *ArXiv Preprint ArXiv:1604.02135*, 2016.
- [791] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.
- [792] Samira Zare and Hien Van Nguyen. Evaluating and Improving Domain Invariance in Contrastive Self-Supervised Learning by Extrapolating the Loss Function. *IEEE Access*, 2023.
- [793] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.
- [794] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional Networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2528–2535. IEEE, 2010.
- [795] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive Deconvolutional Networks for Mid and High Level Feature Learning. In *IEEE International Conference on Computer Vision*, pages 2018–2025. IEEE, 2011.
- [796] Haijin Zeng, Xiangming Wang, Yongyong Chen, Jingyong Su, and Jie Liu. Vision-language gradient descent-driven all-in-one deep unfolding networks, 2025.

-
- [797] Chenlu Zhan, Yufei Zhang, Yu Lin, Gaoang Wang, and Hongwei Wang. UniDCP: Unifying Multiple Medical Vision-Language Tasks via Dynamic Cross-Modal Learnable Prompts. *IEEE Transactions on Multimedia*, 2024.
- [798] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal Image Synthesis and Editing: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [799] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. StyleSwin: Transformer-Based GAN for High-Resolution Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11304–11314, 2022.
- [800] Bowen Zhang, Liyang Liu, Minh Hieu Phan, Zhi Tian, Chunhua Shen, and Yifan Liu. SegViT V2: Exploring Efficient and Continual Semantic Segmentation with Plain Vision Transformers. *International Journal of Computer Vision*, 132(4):1126–1147, 2024.
- [801] Chengming Zhang, Geng Yuan, Wei Niu, Jiannan Tian, Sian Jin, Donglin Zhuang, Zhe Jiang, Yanzhi Wang, Bin Ren, Shuaiwen Leon Song, et al. ClickTrain: Efficient and Accurate End-to-End Deep Learning Training via Fine-Grained Architecture-Preserving Pruning. In *ACM International Conference on Supercomputing*, pages 266–278, 2021.
- [802] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal Intervention for Weakly-Supervised Semantic Segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.
- [803] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature Pyramid Transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 323–339. Springer, 2020.
- [804] Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. RefineMask: Towards High-Quality Instance Segmentation with Fine-Grained Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6861–6869, 2021.
- [805] Guanqun Zhang, Shuai Xu, Jing Li, and Alan JX Guo. Group-Based Network Pruning via Nonlinear Relationship Between Convolution Filters. *Applied Intelligence*, 52(8):9274–9288, 2022.
- [806] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Amrith Tyagi, and Amit Agrawal. Context Encoding for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Bibliography

- [807] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. ResNeSt: Split-Attention Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2736–2746, 2022.
- [808] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with Improved Denoising Anchor Boxes for End-to-End Object Detection. *ArXiv Preprint ArXiv:2203.03605*, 2022.
- [809] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. In *European Conference on Computer Vision*, 2020.
- [810] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.
- [811] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3214–3223, 2020.
- [812] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a Single Convolutional Super-Resolution Network for Multiple Degradations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018.
- [813] Li Zhang, Jiachen Lu, Sixiao Zheng, Xinxuan Zhao, Xiatian Zhu, Yanwei Fu, Tao Xiang, Jianfeng Feng, and Philip HS Torr. Vision Transformers: From Semantic Segmentation to Dense Prediction. *International Journal of Computer Vision*, 132(12):6142–6162, 2024.
- [814] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. DiffusionEngine: Diffusion Model Is Scalable Data Engine for Object Detection. *ArXiv Preprint ArXiv:2309.03893*, 2023.
- [815] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023.
- [816] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. In *IEEE International Conference on Computer Vision*, pages 2998–3008, 2021.
- [817] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. VSA: Learning Varied-Size Window Attention in Vision Transformers. In *European Conference on Computer Vision*, pages 466–483. Springer, 2022.

-
- [818] Qinglong Zhang and Yu-Bin Yang. REST: An Efficient Transformer for Visual Recognition. *Advances in Neural Information Processing Systems*, 34:15475–15485, 2021.
- [819] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. *ArXiv Preprint ArXiv:2303.00915*, 2023.
- [820] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-Shot Refinement Neural Network for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4203–4212, 2018.
- [821] Wei Zhang and Zhiming Wang. FPFS: Filter-Level Pruning via Distance Weight Measuring Filter Similarity. *Neurocomputing*, 512:40–51, 2022.
- [822] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards Unified Image Segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.
- [823] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.
- [824] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1943–1955, 2015.
- [825] Xiaofeng Zhang, Shuli Cheng, Liejun Wang, and Haojin Li. Asymmetric Cross-Attention Hierarchical Network Based on CNN and Transformer for Bitemporal Remote Sensing Images Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
- [826] Xinbang Zhang, Zehao Huang, Naiyan Wang, Shiming Xiang, and Chunhong Pan. You Only Search Once: Single Shot Neural Architecture Search via Direct Sparse Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):2891–2904, 2020.
- [827] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A Pursuit of Structural Diversity in Very Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3900–3908. IEEE, 2017.
- [828] Yang Zhang, Hui Chen, Imad Rida, and Xianxun Zhu. A generative random modality dropout framework for robust multimodal emotion recognition. *IEEE Intelligent Systems*, 40(5):62–69, 2025.
- [829] Yikang Zhang, Jian Zhang, Qiang Wang, and Zhao Zhong. DyNet: Dynamic Convolution for Accelerating Convolutional Neural Networks. *ArXiv Preprint ArXiv:2004.10694*, 2020.

Bibliography

- [830] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L Yuille, Yaoyao Liu, and Zongwei Zhou. Continual Learning for Abdominal Multi-Organ and Tumor Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2023.
- [831] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When Radiology Report Generation Meets Knowledge Graph. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917, 2020.
- [832] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *European Conference on Computer Vision*, pages 286–301, 2018.
- [833] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual Dense Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.
- [834] Zhaoyang Zhang, Jingyu Li, Wenqi Shao, Zhanglin Peng, Ruimao Zhang, Xiaogang Wang, and Ping Luo. Differentiable Learning-to-Group Channels via Groupable Convolutional Neural Networks. In *IEEE International Conference on Computer Vision*, pages 3542–3551, 2019.
- [835] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating Nested Transformers. *ArXiv Preprint ArXiv:2105.12723*, 2(3):5, 2021.
- [836] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring Self-Attention for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10076–10085, 2020.
- [837] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.
- [838] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-Wise Spatial Attention Network for Scene Parsing. In *European Conference on Computer Vision*, pages 267–283, 2018.
- [839] Pengbo Zhao, Zhenshen Qu, Yingjia Bu, Wenming Tan, and Qiuyu Guan. PolarDet: A Fast, More Precise Detector for Rotated Target in Aerial Images. *International Journal of Remote Sensing*, 42(15):5831–5861, 2021.
- [840] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic Diffusion Transformer. *ArXiv Preprint ArXiv:2410.03456*, 2024.
- [841] Xiang Zhao, Tiejun Yang, Bingjie Li, and Xin Zhang. SwinGAN: A Dual-Domain Swin Transformer-Based Generative Adversarial Network for MRI Reconstruction. *Computers in Biology and Medicine*, 153:106513, 2023.

-
- [842] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep Learning-Based Human Pose Estimation: A Survey. *ACM Computing Surveys*, 56(1):1–37, 2023.
- [843] Changyan Zheng, Liguu Xu, Xiaohu Fan, Jibin Yang, Junyi Fan, and Xian Huang. Dual-Path Transformer-Based Network with Equalization-Generation Components Prediction for Flexible Vibrational Sensor Speech Enhancement in the Time Domain. *The Journal of the Acoustical Society of America*, 151(5):2814–2825, 2022.
- [844] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-End Object Detection with Adaptive Clustering Transformer. *ArXiv Preprint ArXiv:2011.09315*, 2020.
- [845] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional Random Fields as Recurrent Neural Networks. In *IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.
- [846] Siming Zheng and Xin Yuan. Unfolding framework with prior of convolution-transformer mixture and uncertainty estimation for video snapshot compressive imaging, 2023.
- [847] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.
- [848] Wanfeng Zheng, Qiang Li, Guoxin Zhang, Pengfei Wan, and Zhongyuan Wang. ITTR: Unpaired Image-to-Image Translation with Transformers. *ArXiv Preprint ArXiv:2203.16015*, 2022.
- [849] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes Through the ADE20K Dataset. *ArXiv Preprint ArXiv:1608.05442*, 2016.
- [850] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing Through ADE20K Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [851] Daquan Zhou, Yujun Shi, Bingyi Kang, Weihao Yu, Zihang Jiang, Yuan Li, Xiaojie Jin, Qibin Hou, and Jiashi Feng. Refiner: Refining Self-Attention for Vision Transformers. *ArXiv Preprint ArXiv:2106.03714*, 2021.
- [852] Mingyi Zhou, Yipeng Liu, Zhen Long, Longxi Chen, and Ce Zhu. Tensor Rank Learning in CP Decomposition via Convolutional Neural Network. *Signal Processing: Image Communication*, 73:12–21, 2019.

Bibliography

- [853] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. Scale-Transferrable Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 528–537, 2018.
- [854] Sheng Zhou, Yucheng Wang, Defang Chen, Jiawei Chen, Xin Wang, Can Wang, and Jiajun Bu. Distilling Holistic Knowledge with Graph Neural Networks. In *IEEE International Conference on Computer Vision*, pages 10387–10396, 2021.
- [855] Tianfei Zhou and Wenguan Wang. Prototype-Based Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6858–6872, 2024.
- [856] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional Semantic Contrast and Aggregation for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022.
- [857] Xin Zhou, Aixin Sun, Yong Liu, Jie Zhang, and Chunyan Miao. SelfCF: A Simple Framework for Self-Supervised Collaborative Filtering. *ACM Transactions on Recommender Systems*, 1(2):1–25, 2023.
- [858] Yi-Tong Zhou and Rama Chellappa. Computation of Optical Flow Using a Neural Network. In *IEEE International Conference on Neural Networks*, volume 1998, pages 71–78, 1988.
- [859] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary Relation Contrastive Distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9260–9269, 2021.
- [860] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.
- [861] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. BiFormer: Vision Transformer with Bi-Level Routing Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10323–10333, 2023.
- [862] Lianghui Zhu, Yingyue Li, Jieming Fang, Yan Liu, Hao Xin, Wenyu Liu, and Xinggang Wang. WeakTr: Exploring Plain Vision Transformer for Weakly-Supervised Semantic Segmentation. *ArXiv Preprint ArXiv:2304.01184*, 2023.
- [863] Mingjian Zhu, Yehui Tang, and Kai Han. Vision Transformer Pruning. *ArXiv Preprint ArXiv:2104.08500*, 2021.
- [864] Xianxun Zhu, Heyang Feng, Erik Cambria, Yao Huang, Ming Ju, Haochen Yuan, and Rui Wang. Emvas: End-to-end multimodal emotion visualization analysis system. *Complex Intell. Syst.*, 11(8):1–15, 2025.

-
- [865] Xianxun Zhu, Chaopeng Guo, Heyang Feng, Yao Huang, Yichen Feng, Xiangyang Wang, and Rui Wang. A review of key technologies for emotion analysis using multimodal information. *Cognitive Computation*, 16(4):1504–1530, 2024.
- [866] Xianxun Zhu, Yao Huang, Xing Wang, and Rui Wang. Emotion recognition based on brain-like multimodal hierarchical perception. *Mult. Tools Appl.*, 83:56039–56057, 2024.
- [867] Xianxun Zhu, Zhaozhao Liu, Erik Cambria, Xiaohan Yu, Xuhui Fan, Hui Chen, and Rui Wang. A client–server based recognition system: Non-contact single/-multiple emotional and behavioral state assessment methods. *Computer Methods and Programs in Biomedicine*, 260:108564, 2025.
- [868] Xianxun Zhu, Yaoyang Wang, Erik Cambria, Imad Rida, José Santamaría López, Lin Cui, and Rui Wang. Rmer-dt: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. *Information Fusion*, 123:103268, 2025.
- [869] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable ConvNets V2: More Deformable, Better Results. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.
- [870] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *ArXiv Preprint ArXiv:2010.04159*, 2020.
- [871] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–557, 2018.
- [872] Zijian Zhu, Yichi Zhang, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Zhong, and Shibao Zheng. Understanding the Robustness of 3D Object Detection with Bird’s-Eye-View Representations in Autonomous Driving. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 21600–21610, 2023.
- [873] Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. *CoRR*, abs/1611.01578, 2016.
- [874] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. *CoRR*, abs/1707.07012, 2017.
- [875] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning Transferable Architectures for Scalable Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, 2018.

Bibliography

- [876] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once. *Advances in Neural Information Processing Systems*, 36:19769–19782, 2023.
- [877] Lojze Žust and Matej Kristan. PanSR: An Object-Centric Mask Transformer for Panoptic Segmentation, 2024.

Supplementary Materials in Chapter 5

As an extra, we have added two zoomed-in samples to highlight the details of our method and other approaches. The first row of Figure A.1 shows that our method can correctly divide two glomeruli without overlap. Cascade Mask R-CNN and Mask R-CNN, however, generate multiple responses at the border of two objects. QueryInst with static queries cannot detect or segment any glomerulus. The last row of Figure A.1 demonstrates four large arteries in a line. It is complicated because two large arteries pass over most of the input image regions. Although it had multiple responses at the artery border, our method can capture long-range spatial dependencies for large objects using attention mechanisms with dynamic queries. In contrast, Cascade Mask RCNN and Mask RCNN cannot model large objects without attention modules. QueryInst uses a static attention mechanism and cannot handle complex combinations of shapes and appearances with a fixed number of queries.

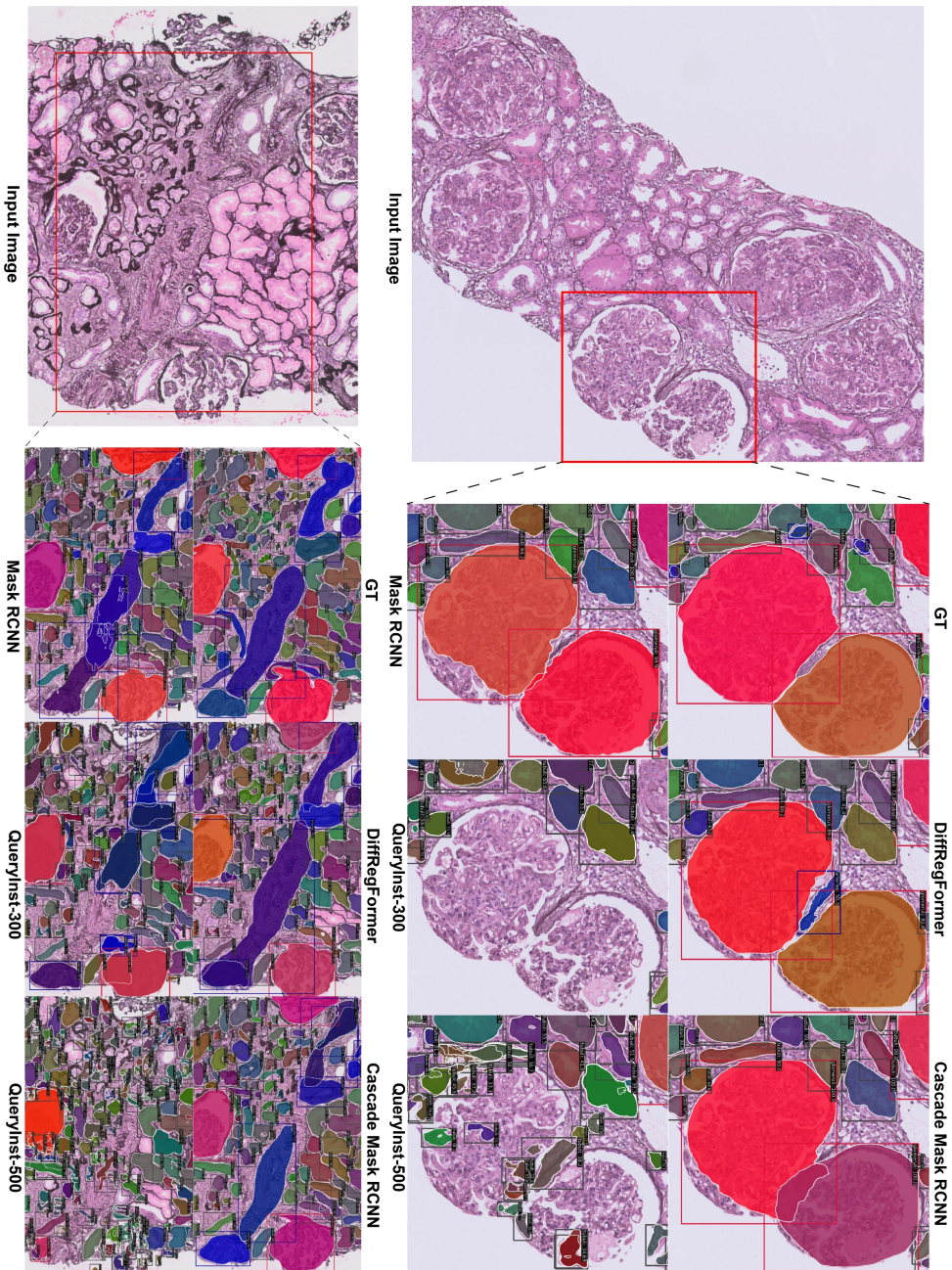


Figure A.1: Detailed zoom-in samples comparing DiffRegFormer with other methods.

Summary

Deep learning has revolutionized image processing, enabling remarkable advances in fields from autonomous driving to medical diagnostics. Yet the very success of these models, fueled by increasing complexity, has raised urgent questions about how they work, why they sometimes fail, and whether they can be trusted in critical applications. This thesis addresses these questions by offering both a coherent map of the rapidly evolving landscape of deep learning in computer vision and concrete contributions that advance model interpretability and reliability. The thesis is structured as a logical progression from a broad, systematic overview to focused investigations that together build a foundation for more transparent and robust vision systems.

Chapter 1 discusses the concepts relevant to this thesis. We then introduce the research questions that form the basis for the content of **Chapters 2 to 5**.

Chapter 2 provides a comprehensive review of the deep learning models that have come to dominate image processing. Starting from the mathematical principles shared by convolutional neural networks and vision transformers, it traces their parallel evolution and eventual convergence. Beyond classical architectures, the chapter introduces recent large-scale foundation models: pure vision, vision–language, and multi-modal models, highlighting their potential to serve as universal solutions for vision tasks. By identifying open challenges and future directions, this chapter establishes the conceptual toolkit and the research gaps that motivate the subsequent chapters.

Building on this foundation, **Chapter 3** addresses interpretability in a core vision task: semantic segmentation. Focusing on the cross-attention mechanism, we reformulate it as a probabilistic model and reveal that its operation can be interpreted as k-means clustering in the feature representation space. This novel perspective not only demystifies the widely used cross-attention component but also provides a principled basis for understanding its behavior under full supervision.

Chapter 4 extends this line of inquiry to the more challenging weakly supervised setting, where pixel-wise annotations are unavailable. We analyze how the absence of detailed labels shapes the learned feature space and propose a probabilistic framework that recovers missing background information through contrastive learning. Together, **Chapters 3 - 4** demonstrate that a unified probabilistic view can explain and improve attention-based models across different supervision regimes, thereby enhancing their interpretability.

Chapter 5 addresses reliability in a demanding clinical application: the analysis of microscope images of kidney biopsies. Our technical contributions result in a reliable and robust approach. In transplant medicine, monitoring a transplanted kidney is crucial. The status of the kidney is studied based on the micro-anatomy in a biopsy. The anatomical structures are densely packed, vary widely in size and shape, and often touch. We introduce a generative instance segmentation model that combines diffusion models with vision transformers. The architecture is explicitly designed to mitigate class imbalance, resolve objects at multiple scales, and disambiguate touching instances by emphasizing boundary information. Extensive experiments confirm its robustness, highlighting its potential for clinical deployment.

Finally, in **Chapter 6**, we summarize the results and conclusions of our research and answer the research questions. Based on these results, we outline a perspective for future developments in the field.

Collectively, this thesis reflects on how a macroscopic understanding of model evolution (**Chapter 2**) informs targeted investigations into interpretability (**Chapters 3–4**) and reliability (**Chapter 5**), demonstrating how these contributions together advance the goal of building vision systems that are not only powerful but also understandable and trustworthy. By bridging theoretical foundations with practical solutions, this work aims to pave the way for deep learning models that can be deployed with confidence in real-world settings, where transparency and dependability are as important as accuracy.

Nederlandse samenvatting

Beeldbewerking heeft met de komst van Deep Learning een revolutie doorgemaakt. Dit heeft, in het bijzonder, opmerkelijke voortgang gegeven in het onderzoeksveld variërend van autonome besturing van voertuigen tot medische diagnostiek. Ondanks het succes van de rekenmodellen onderliggend aan Deep Learning, heeft een steeds meer toenemende complexiteit de urgente vragen opgeworpen over hoe deze modellen werken, waarom ze soms niet goed werken, en of deze modellen vertrouwd kunnen worden in kritische toepassingen. In dit proefschrift worden juist deze vragen geadresseerd door het presenteren van een overzicht van de ontwikkelingen in deep-learning in het beeldbewerkingsdomein en daarnaast worden bijdragen geleverd waarin de vooruitgang in model interpretatie en betrouwbaarheid wordt behandeld. Het onderzoek beschreven in dit proefschrift begint vanuit het perspectief van een breed overzicht gevolgd door specifieke methodologische verkenningen. Tezamen vormen de Hoofdstukken de basis voor transparante en robuuste systemen voor computer visie.

In **Hoofdstuk 1** behandelen we de begrippen die van belang zijn voor dit proefschrift. Vervolgens introduceren we de onderzoeksvragen die de basis zijn voor de inhoud van de Hoofdstukken 2 tot en met 5.

In **Hoofdstuk 2** wordt een uitgebreid literatuuroverzicht gepresenteerd over Deep Learning modellen die het werkveld van beeldbewerking domineren. In dit overzicht zijn het uitgangspunt de wiskundige principes die worden gedeeld door convolutionele neurale netwerken en vision-transformers, daarbij beoordelen we hoe deze modellen in parallel zijn geëvolueerd en waar eventuele convergentie wordt gevonden.

In **Hoofdstuk 3** bouwen we verder op het fundament van het voorgaande hoofdstuk en behandelen we interpreteerbaarheid van een kerntaak in computer visie, i.e. semantische segmentatie. We leggen de nadruk op het mechanisme van cross-attentie dat we herformuleren als een probabilistisch model en laten zien dat deze aanpak kan worden

geïnterpreteerd als een k-means clustering in de representatie van de kernmerkruimte. Dit nieuwe perspectief verheldert de veel gebruikte cross-attentie component en het voorziet tevens in een solide basis voor begrip van het gedrag van deze component onder de zogenaamde conditie van volledige supervisie.

In **Hoofdstuk 4** vervolgen we ons onderzoek en bestuderen we het uitdagende “weakly supervised” segmentatie waarbij annotaties op basis van gelabelde pixels niet beschikbaar zijn. We analyseren hoe in afwezigheid van gedetailleerde labels de kenmerkruimte tot stand komt en introduceren een probabilistisch kader waarmee missende achtergrondinformatie door contrastief leren kan worden gecorrigeerd. In samenhang laten de **Hoofdstukken 3 en 4** zien dat een geïntegreerde probabilistische weergave attentiegebaseerde modellen kan verklaren en verbeteren onder verschillende regimes van supervisie. Hierdoor wordt de interpreteerbaarheid van deze modellen aanzienlijk vergroot.

In **Hoofdstuk 5** gebruiken we onze technische bijdrages in een veeleisende klinische toepassing: de analyse van microscoopbeelden van nier-biopten. De uitwerking van onze technische bijdrages resulteert in een betrouwbare en robuuste aanpak. In transplantatiegeneeskunde is het monitoren van een getransplanteerde nier cruciaal, de status van de nier wordt aan de hand van de micro-anatomie in een biopt bestudeerd. De anatomische structuren zitten dicht opeengepakt, vertonen een grote variatie in vorm en grootte en zitten vaak aaneengeplakt. Om een automatische analyse mogelijk te maken introduceren we een generatief model van instantie segmentatie waarbij diffusie modellen worden gecombineerd met vision-transformers. De architectuur van onze methode is expliciet ontworpen om onbalans in de klassen te beperken, objecten te herkennen bij wisselende vergroting en desambiguering van aaneengeplakte structuren door het benadrukken van randinformatie. Door extensieve experimenten kunnen we de robuustheid van onze methode vaststellen en illustreren we de potentie voor het inzetten in de klinische praktijk.

In **Hoofdstuk 6**, tenslotte, vatten we de resultaten en conclusies van ons onderzoek samen en beantwoorden we de onderzoeksvragen. Met de resultaten als basis schetsen we een perspectief voor toekomstige ontwikkelingen in het veld.

Alles bij elkaar biedt dit proefschrift een reflectie over hoe een macroscopisch begrip van de evolutie van een model (**Hoofdstuk 2**) informatie geeft over specifiek onderzoek in interpreteerbaarheid (**Hoofdstuk 3-4**) en betrouwbaarheid (**Hoofdstuk 5**). Daarmee laat het zien hoe onze bijdragen tezamen vooruitgang bieden aan het doel

computer vision systemen te bouwen die krachtig zijn en tegelijkertijd ook begrijpelijk en betrouwbaar. Door het overbruggen van theoretische fundamenteen naar praktische oplossingen, tracht dit proefschrift de weg te bereiden door het implementeren van deep-learning modellen die betrouwbaar zijn onder operationele omstandigheden, waar transparantie en afhankelijkheid even belangrijk zijn als nauwkeurigheid.

Curriculum Vitae

Zhan Xiong was born in 1988 in Chongqing, China. He began his bachelor's degree in computer science at Southwest University in 2006 and completed his master's degree in Computer Vision there in 2010. During his master's studies, he worked in the Multimedia Lab under Prof. Guoqiang Xiao on medical image analysis, developing an interest in applied mathematics.

In 2013, he received a Chinese Scholarship Council scholarship and started his PhD at Leiden University under Prof. Fons Verbeek. His research combined image analysis and bioinformatics, covering machine learning, image reconstruction, algorithm design, and data analysis. He later shifted toward deep learning and model interpretability, strengthening his mathematical foundation in areas such as calculus, linear algebra, probability, and Bayesian inference.

During his PhD at Leiden University, he took courses including Communication in Science, Effective Communication, Time Management, and Scientific Conduct. He aims to develop interpretable and controllable deep learning systems that integrate theory with practice.

Author Publication List

Zhan Xiong and Fons J. Verbeek. Segmentation of Zebrafish Larva Inhomogeneous 3D Images Using the Level-Set Method. In David L. Kao, Thomas Wischgoll, and Song Zhang, editors, *Visualization and Data Analysis 2016, San Francisco, California, USA, February 14-18, 2016*, pages 1–9. Society for Imaging Science and Technology, 2016.

Yuanhao Guo, Zhan Xiong, and Fons J. Verbeek. An Efficient and Robust Hybrid Method for Segmentation of Zebrafish Objects from Bright-Field Microscope Images. *Machine Vision and Applications*, 29(8):1211–1225, 2018.

Zhan Xiong and Fons J. Verbeek. Weakly Supervised Semantic Segmentation via Contextual Cross-Attention. In *9th International Conference on Frontiers of Signal Processing, ICFSP 2024, Paris, France, September 12-14, 2024*, pages 83–92. IEEE, 2024

Zhan Xiong, Junling He, Pieter Valkema, Tri Q. Nguyen, Maarten Naesens, Jesper Kers, and Fons J. Verbeek. Advances in Kidney Biopsy Lesion Assessment Through Dense Instance Segmentation. *Artificial Intelligence in Medicine*, 164:103111, 2025.

Junling He, Pieter A. Valkema, Jingmin Long, Jia Li, Sandrine Florquin, Maarten Naesens, Priyanka Koshy, Tri Q. Nguyen, Soufian Meziyerh, Aiko P. J. de Vries, Onno J. De Boer, Fons J. Verbeek, Zhan Xiong, and Jesper Kers. Transformer-Based Multiclass Segmentation Pipeline for Basic Kidney Histology. *Scientific Reports*, 15(1):38957, 2025.

Zhan Xiong, Xiaoqin Tang, and Fons J. Verbeek. VGMEA: Interpretable Cross-Attention Mechanisms in Semantic Segmentation via Variational Gaussian Model Mixture. *ArXiv Preprint*, 2026.

Zhan Xiong, Xiaoqin Tang, and Fons J. Verbeek. A Survey of Deep Learning in Image Processing: Developments, Applications and Challenges. *Artificial Intelligence Review (Under Review)*, 2026.

Acknowledgments

The pursuit of a PhD has been, for me, a mirror of self-reflection. It revealed not only my own unrealistic arrogance, blind optimism, and moments of utter frustration, but also an indomitable resilience and an enduring patience I never knew I had. My journey began in September 2013, when I first walked into the LIACS building in Leiden, the Netherlands. Since then, time has passed as quietly as a flowing stream, carrying me along a winding path toward the end of this doctoral journey, not in haste, but with a quiet sense of completeness. Only now, as I set down this closing summary of an extraordinary experience, can I finally say with certainty: it has been one of the most significant chapters of my life. It has made me more grounded, taught me to balance ideals and reality, and guided me to build, patiently and pragmatically, a path in the real world for what I once only dreamed. None of this would have been possible without the support of my supervisors, colleagues, friends, and family. It is with my deepest gratitude that I share this moment with them.

First and foremost, I thank my supervisor, Prof. Fons J. Verbeek, for his mentorship and constant support. Beyond technical expertise in microscopy, algorithms, and bioinformatics, he taught me to think independently, question critically, remain humble, and collaborate respectfully across cultures. His guidance and example have had a lasting impact on both my research and personal development.

I am also grateful to my co-supervisor, Dr. Xiaoqin Tang, for her support in the final stage of my thesis. She provided a clear and realistic writing schedule, kept me accountable to deadlines, and significantly accelerated my progress. She not only improved the clarity and coherence of my work but also helped me stay focused on the main ideas.

I am also deeply grateful to my colleagues and staff at LIACS. You welcomed me into a new environment with warmth and openness, making LIACS feel like home long

Acknowledgments

before I realized it had become one. The list of those I wish to thank is far too long to capture here fully. It includes, among many others, Katy, Lu, Sacha, Rafael, Laura, Fuyu, Xiaoqin, Yuanhao, Jingmin, Jia, Bei, Wanbin, Danyi, Yi, Shima, Irene, Ana, Solomiia, Erick, Erwin, Song, Yanming, Yu, Punan, and Yuzhi. My heartfelt thanks also go to my office-mates Enrique, Mohamed, Feibo, and Chen. The moments we shared—coffee breaks in the corridor, lunch conversations in the canteen, academic discussions by the whiteboard—always reminded me that I was working alongside people who truly understand the struggles of a PhD. You listened to my half-formed ideas, offered thoughtful feedback, and helped me see my own work more clearly through your questions and insights. I would also like to thank Marloes van der Nat for her help with my studies at Leiden University. She has helped me solve many problems regarding courses, residence permits, graduation logistics, and more. This journey would have been far lonelier and far less joyful without you.

I owe a deep debt of gratitude to my friends, who have walked with me through every season of this journey. To Guangchao and Qiuyin, who helped me board the plane and carried the same uncertainties and hopes from China to Schiphol Airport. To Yuanhao, Song, Enrique, Jing & Liang, Ruobing & Hexi, and Boyang, who, like senior siblings, opened their doors and their hearts to us newcomers. To Fuyu, my former flatmate, who turned four walls into a home. To Yinlong, Yun, Zhenyu & Yaojin, Wenbo, Quanchi, and Rui, my companions in frustration and laughter, always ready to trade complaints over dinner. And to Lanpeng & Xue, Xinru & Kaihua, Yudan & Weichun, and Gangqi, who celebrated every small victory with me and reminded me why this journey matters. You sustained me when the path felt long, and you made the journey not just bearable, but beautiful. To all my friends, you have been my fellow travelers on this winding road and made the journey not just bearable, but beautiful. Thank you.

Last, and most importantly, my deepest gratitude goes to my family. To my parents: your love has never been conditional, and your belief in me has never wavered. I know that these years have been as hard on you as they have been on me, perhaps harder, watching from afar, carrying worries you never voiced during many sleepless nights. Thank you for tolerating my stubborn or even silly decisions, for indulging me as I followed a path you could not see, and for always being there when I return. And to my beloved wife, Dr. Junling He, who left behind a familiar life, a career, and the comfort of home to accompany me to a foreign country, changing her planned life path to pursue her own medical doctorate with equal courage and determination. You have

Acknowledgments

walked this path beside me. Your trust, your resilience, and your quiet strength have been my anchor. In the moments when I doubted myself, it was your belief in me and the future we are building together that kept me going. This journey would have been impossible without you.

Zhan Xiong
19 May 2026
Leiden