



Universiteit
Leiden
The Netherlands

Exploring deep learning for data-efficient and generalizable visual recognition

He, K.

Citation

He, K. (2026, September 22). *Exploring deep learning for data-efficient and generalizable visual recognition*. Retrieved from <https://hdl.handle.net/1887/4310276>

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/1887/4310276>

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

Exploring Deep Learning for Data-Efficient and Generalizable Visual Recognition

Kai He

Copyright © 2026 Kai He, All Rights Reserved

ISBN: 978-90-9043099-7

Printed by UFB Grafisch centrum, The Netherlands

Cover design: Kai He

Exploring Deep Learning for Data-Efficient and Generalizable Visual Recognition

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 22 september 2026
klokke 13.00 uur

door

Kai He

geboren te Yangzhou, China
in 1995

Promotor: Prof.dr. M.S.K. Lew
Co-promotor: Dr. E.M. Bakker

Promotiecommissie: Prof.dr. T.H.W. Bäck
Prof.dr. M.M. Bonsangue
Prof.dr. A. Plaat
Prof.dr. M.R. Spruit
Prof.dr. S. Satoh (National Institute of Informatics, Tokyo)
Dr. K.J. Wolstencroft (Amsterdam UMC, Amsterdam)



Kai He was financially supported through the China Scholarship Council (CSC) to participate in the PhD programme of Leiden University. Grant number 202106870027.

The research in this thesis was performed at the LIACS MediaLab, Leiden University, The Netherlands, and we would like to thank the NVIDIA Corporation for the donation of GPU cards.

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Background and Related Work	5
1.2.1	Metric Learning	5
1.2.2	Feature Enrichment	6
1.2.3	Knowledge Transfer	8
1.3	Thesis Outline and Research Questions	10
1.4	Main Contributions	13
1.4.1	Task Contributions	13
1.4.2	Technical Contributions	14
2	Few-shot & meta-learning methods for image understanding: a survey	17
2.1	Introduction	18
2.2	The framework of few-shot image classification	19
2.2.1	Notation and definitions	19
2.2.2	Datasets	20
2.2.3	Evaluation process of few-shot image classification	21
2.3	Paradigms of meta-learning for few-shot image classification	22
2.3.1	Metric-based meta-learning	22
2.3.2	Model-based meta-learning	26
2.3.3	Optimization-based meta-learning	29
2.3.4	Other methods	32
2.4	Major Challenges and future directions	33
2.4.1	Limitations and challenges	36
2.4.2	Future directions	36
2.5	Chapter Conclusions	37
3	ProtoSimNet: Exploiting Prototype Similarity for Few-shot Learning	39
3.1	Introduction	40
3.2	Related Work	41
3.3	Proposed Method	44

3.3.1	Problem Formulation	44
3.3.2	Cosine Sharpening Transformation	44
3.3.3	Elastic Regularization	45
3.4	Experiments	47
3.4.1	Datasets and Implementation details	47
3.4.2	Evaluations	47
3.4.3	Ablative Analysis	48
3.5	Chapter Conclusions	53
4	Attention-guided Feature Pyramid Network for few-shot learning	55
4.1	Introduction	56
4.2	Related Work	60
4.2.1	Few-Shot Learning Paradigm	60
4.2.2	Attention Mechanisms for Feature Enrichment	60
4.3	Proposed Methods	61
4.3.1	Multi-Scale Feature Extraction	62
4.3.2	Feature Pyramid Construction	62
4.3.3	Simple Parameter-Free Attention Module (SimAM)	62
4.3.4	Cross-Layer Spatial Attention Module (CSAM)	63
4.3.5	Classification	63
4.4	Experiments	64
4.4.1	Benchmarks and Implementation Details	64
4.4.2	Experimental results	64
4.4.3	Ablation Study	65
4.5	Chapter Conclusions	66
5	Adaptive Multi-scale Feature Fusion Network for Few-shot Learning	69
5.1	Introduction	70
5.2	Related Work	71
5.3	Adaptive multi-scale feature fusion	72
5.3.1	Problem formulation and notions	72
5.3.2	Feature enrich module	72
5.3.3	Feature fusion module	73
5.4	Experiments	74
5.4.1	Datasets, implementation details and evaluation protocol	75
5.4.2	Experimental results and ablative analysis	76
5.5	Chapter Conclusions	77
6	Dual Selective Knowledge Transfer for Few-Shot Classification	79
6.1	Introduction	80
6.2	Related Work	82
6.3	Dual Selective Knowledge Transfer	83
6.3.1	Problem Formulation	83

6.3.2	Multi-task Learning	84
6.3.3	Dual Selective Knowledge Distillation	84
6.4	Experiments	86
6.4.1	Datasets and Implementation details	86
6.4.2	Evaluations	87
6.4.3	Ablative Analysis	89
6.5	Chapter Conclusions	93
7	Conclusions	95
7.1	Main Findings	95
7.2	Limitations	96
7.2.1	Algorithmic perspective	97
7.2.2	Practical perspective	97
7.3	Future Research Directions	98
7.3.1	Metric Learning	98
7.3.2	Feature Enrichment	98
7.3.3	Knowledge Transfer	99
	Bibliography	101
	List of Abbreviations	119
	English Summary	121
	Nederlandse Samenvatting	123
	Curriculum Vitae	127
	Publication List	129

Chapter 1

Introduction

1.1 Motivation

With increasing attention toward social safety and security, comprehensive visual surveillance infrastructures are deployed in public areas, e.g., transportation hubs, healthcare facilities and retail centers. On the one hand, manually processing extensive visual data generated by these systems is time consuming and will bring severe economic burdens. On the other hand, in some high-stakes areas, such as autonomous driving and rare disease diagnosis, the availability of labeled training samples is often limited, making conventional supervised learning approaches impractical. Hence, there is an urgent need for few-shot learning (FSL), which aims to generalize and adapt models to novel categories based on a handful of annotated instances. As illustrated in Figure 1.1, a standard FSL pipeline typically includes: 1) Collect and preprocess a labeled dataset of base classes; 2) Train a feature extractor on abundant base-class samples; 3) Construct episodes that mimic few-shot scenarios and train the meta-learner or metric head under few-shot constraints; 4) Given novel classes with only a few labeled support samples, adapt the model using metric-based matching or optimization-based tuning; 5) Evaluate query images from unseen classes to measure model adaptability and robustness. This thesis places particular emphasis on Components 2, 3 and 4.

Thanks to the rapid development of deep learning, FSL has achieved remarkable progress in recent years. Current FSL approaches have demonstrated impressive performance on standard benchmarks, where models are evaluated on novel classes given only a handful of labeled examples [1, 2]. In some constrained cases, the performance of these FSL methods even surpass human-level recognition accuracy [3, 4]. However, in real-world scenarios, data distributions are diverse, which is a big challenge for models' generalization ability and robustness. In this thesis, we aim to bridge the gap between standard benchmarks and real-world scenarios, and improve FSL models' performance by exploring three core research themes: metric learning, representation enhancement and knowledge transfer.

In Figure 1.2 the three themes are conceptually depicted.

- **Metric Learning** plays an essential role in machine learning systems, which aims to learn a highly discriminative embedding space so that the distances between similar samples are as close as possible and the distances between different categories are as large as possible. In general, metric learning significantly improves the performance in classification, clustering, and retrieval tasks. Inspired by this, we explore the similarities between class prototypes for FSL.
- **Feature Enrichment** focuses on refining learned features, boosting models' generalization capabilities and robustness. Through techniques such as attention mechanisms or multi-scale fusion, the enhanced representations help the

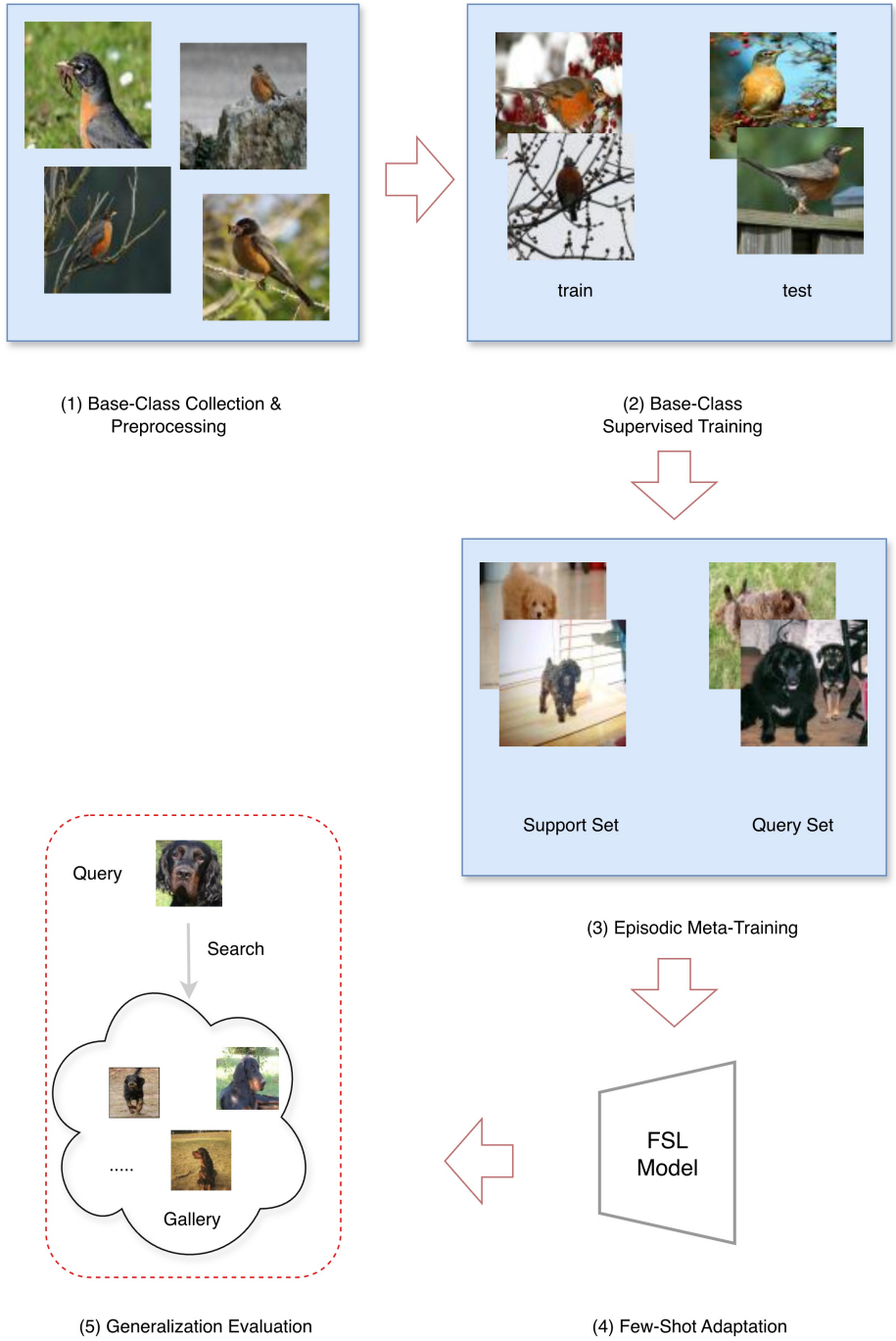


Figure 1.1: Five main stages in the pipeline of FSL.

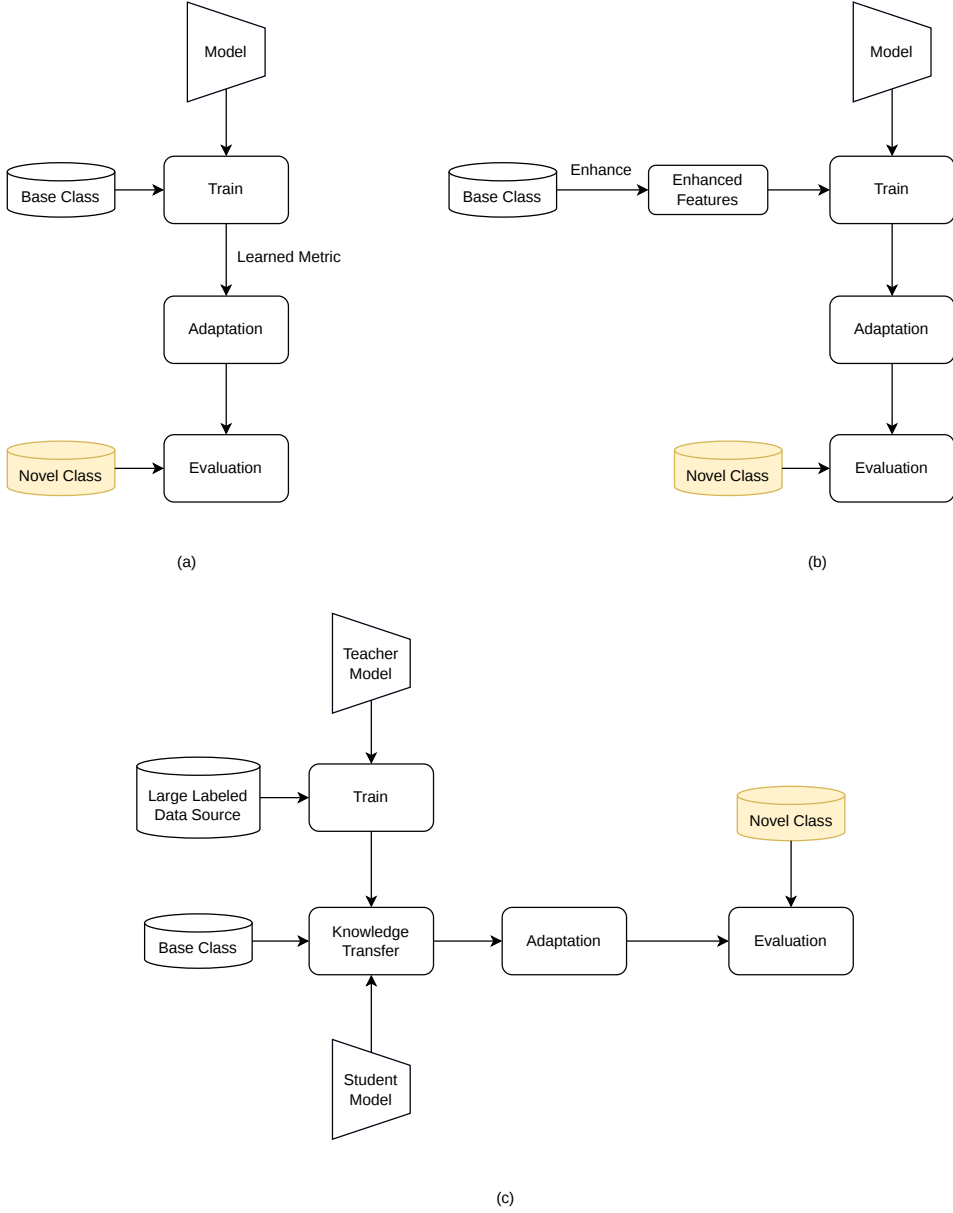


Figure 1.2: Conceptual illustration of the three research themes in this thesis, including (a) metric learning, (b) feature enrichment and (c) knowledge transfer.

model better capture subtle but discriminative characteristics, strengthening the model’s performance under challenging real-world conditions.

- **Knowledge Transfer** addresses the challenge of leveraging previously acquired knowledge to improve the model’s performances in new tasks. It seeks

to bridge the gap between domains by facilitating knowledge sharing, which is particularly effective when labeled data are scarce or expensive to collect. To the best of our knowledge, our work is the first study to explore the dual selective knowledge transfer framework for FSL, which aims to learn effective knowledge from both the teacher model and current samples

1.2 Background and Related Work

As one of the important machine learning applications, few-shot learning (FSL) has attracted extensive attention as a pivotal research direction of learning new concepts from very limited examples. Given a handful of labeled samples—the support set—for each new class, the goal of FSL is to build a model with good robustness and generalization ability by leveraging prior knowledge from related tasks.

In order to comprehensively explore FSL capabilities, over the past decades, a wide variety of benchmark datasets specific for FSL tasks have been proposed. These datasets cover different domains (e.g., natural images, medical scans, buildings) and vary in scale, complexity, and data distribution. The diversity among tasks and domains leads to enormous generalization challenges, often referred to as distribution or domain gaps. A detailed overview of these FSL benchmarks and their characteristics has been presented in the recent literature [5].

As discussed in the survey proposed by Song *et al.* [6], recent FSL researches have rapidly evolved beyond conventional meta-learning setups and shifted toward more realistic and open-world scenarios. This transition emphasizes the need for adaptable and scalable solutions, for example, 1) enhancing discriminability in low-data scenarios by using metric learning to optimize feature embeddings; 2) the use of representation enhancement to improve feature robustness and to mitigate overfitting; 3) the use of knowledge transfer to boost the model’s performance on new tasks by leveraging prior learned knowledge.

1.2.1 Metric Learning

Human beings inherently distinguish similarities and differences among observed instances, and effectively categorizing these objects based on these measurements. This innate ability significantly contributes to our ability for rapid learning and adaptation to new tasks. Analogously, machine learning algorithms also explore systematic strategies to evaluate similarities among samples, which is crucial to classification and retrieval tasks. Unlike traditional supervised learning methods that are trained on extensive datasets, metric learning designs specialized distance functions to classify samples effectively, which is especially valuable in scenarios where only limited labeled samples are available, such as FSL.

Metric learning has also been named similarity learning [7] and deep metric learning [8] in previous works. Existing methods focus on computer vision tasks like image classification [9], object detection [10] and person re-identification [11]. Metric learning requires the model to find an embedding space where similar instances are clustered together. The key challenge in metric Learning, especially within FSL scenarios, is accurately defining and learning meaningful measures from limited samples to ensure good generalization to novel classes. However, conventional metric learning methods are not suitable for FSL for various reasons:

- The number of training samples per class in FSL is much lower than that in conventional metric learning tasks, e.g., popular benchmarks such as CIFAR-100 [12] and ImageNet [13], which provides sufficient data for the model’s training. In contrast, benchmark for FSL scenarios like miniImageNet [14] or Omniglot [3] often provides as few as one to five examples per class, making it challenging for classification tasks on these new classes, as illustrated in Figure 1.3.
- FSL datasets are more imbalanced [15, 16, 17], that means there exist significant intra-class differences and subtle inter-class differences, which makes it difficult to learn reliable metrics that can effectively distinguish categories.
- Moreover, in real-world FSL scenarios, the training (base) classes and the testing (novel) classes follow different data distributions. This increases the difficulty of learning a transferable metric.

To address the above-mentioned difficult yet practical challenges, we propose a novel prototype similarity framework. In contrast to prior works that treat feature extraction, metric refinement, and regularization as independent components, our approach integrates a cosine sharpening transformation with elastic regularization into a unified representation learning framework. This joint design encourages the model to simultaneously refine decision margins, preserve the geometric consistency of the embedding manifold, and enhance robustness to subtle inter-class similarities. Extensive experiments on FSL benchmarks demonstrate that our approach effectively improves generalization to unseen categories and achieves state-of-the-art performance (Chapter 3).

1.2.2 Feature Enrichment

In order to adapt the model to novel classes, extracting discriminative features from the limited number of labeled samples plays an important role in FSL scenarios. To resolve the limitation of insufficient representation caused by scarce training samples, feature enrichment has been exploited to address these practical challenges. In recent FSL works, representation enhancement has been shown effective to extracting discriminative features from limited data, leading to better FSL performance

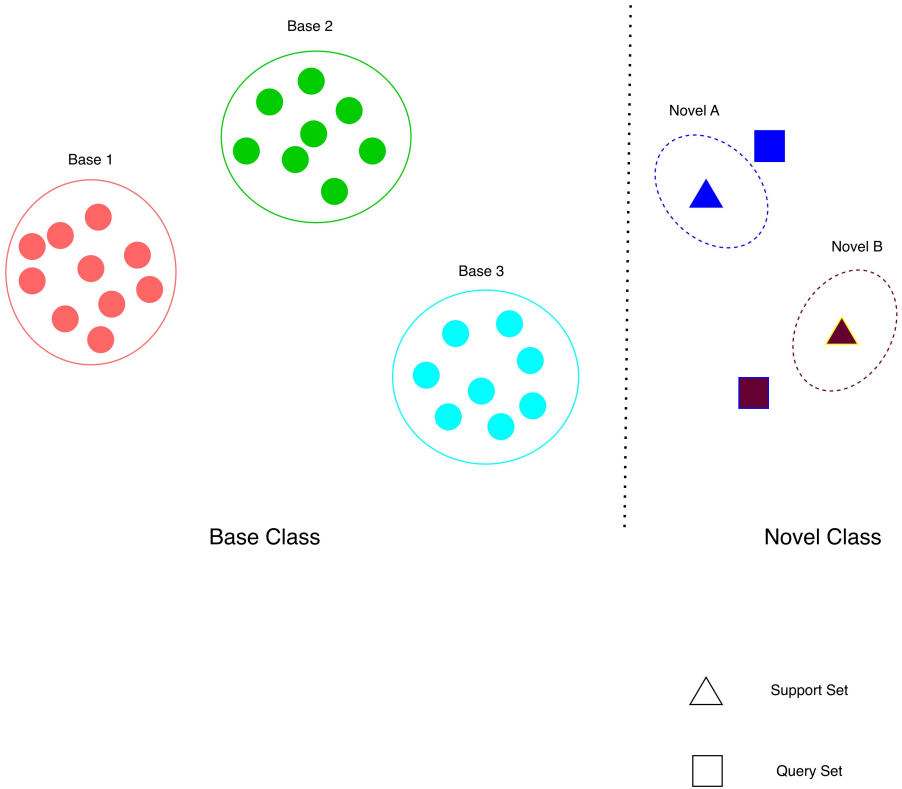


Figure 1.3: Metric learning challenge in FSL: for novel classes, the support set often consists of only 1-5 examples

Driven by the development of deep learning, current representation enhancement methods in FSL can be generally divided into two main branches. The first branch explores self-supervised learning to pretrain models on auxiliary tasks without manual annotations, thereby learning generalized and robust representations. For example, Chen *et al.* [18] proposed a contrastive learning framework that maximizes the agreement between augmented instances of the same image, while Lu *et al.* [19] demonstrated that self-supervised pretraining can outperform supervised pretraining in few-shot scenarios under appropriate conditions, such as deeper backbones, larger image sizes, and stronger data augmentations. Similarly, Gidaris *et al.* [20] introduced a rotation prediction task to learn semantic features that are applicable to FSL. The second branch focuses on feature refinement through attention mechanisms and meta-learning, effectively enhancing the discriminativeness of embeddings for FSL tasks. Tseng *et al.* [21] proposed a feature-wise transformation layer to simulate various feature distributions during meta-training, improving the model’s generalization ability to novel classes. Meanwhile, Hou *et al.* [22] introduced a cross-attention

module to dynamically highlight task-relevant features, reducing the interference by irrelevant background information. Recent work proposed by Oreshkin *et al.* [23] further demonstrated that task-adaptive feature modulation can significantly boost few-shot performance by aligning representations with task-specific requirements. Despite these notable advancements, existing methods still face two main challenges in FSL: 1) Feature overfitting, where enhanced representations generalize poorly to novel classes due to the scarcity of training samples; 2) Domain shift, where representations learned from base classes fail to adapt to novel classes following different data distributions. To mitigate these issues, recent studies have explored hybrid approaches that combine representation enhancement with metric learning [24] or data augmentation [25]. For example, Li *et al.* [26] propose adversarial feature hallucination networks to synthesize diverse training samples, while Wang *et al.* [27] employ information fusion rectification to rectify the distribution of the support set in novel classes.

Different from previous works that mainly focus on external data augmentation or simple feature aggregation strategies, we analyze the limitations that arise when few-shot models are forced to learn discriminative representations from extremely limited data. To address these limitations, we introduce an attention-guided feature pyramid network that explicitly enhances feature richness by integrating complementary information from multiple backbone layers through a learnable top-down fusion mechanism (Chapter 4). This design strengthens the discriminativeness of few-shot embeddings by preserving both local details and global semantic structure. In addition, we develop an adaptive multi-scale feature fusion framework that further enriches representations by expanding multi-branch feature diversity and dynamically re-weighting cross-scale responses according to their task relevance (Chapter 5). Together, these two directions form a comprehensive approach to feature enrichment in FSL, allowing models to form more discriminative, richer, and more transferable embeddings with very limited supervision.

1.2.3 Knowledge Transfer

The rapid development of deep neural networks has obtained advanced performance in many vision tasks. Meanwhile, it is still challenging to train models that can quickly adapt to new tasks with limited labeled examples. Knowledge transfer, aiming to leverage knowledge acquired from previous learned tasks or abundant labeled data to boost performance on new tasks, has emerged as a crucial paradigm in FSL. Given a handful of labeled samples in a target task, knowledge transfer seeks to utilize transferable knowledge from auxiliary domains or tasks to learn robust representations for novel tasks. The key challenge is how to filter useful knowledge, as presented in Figure 1.4.

Existing knowledge transfer approaches applied in FSL can be broadly divided into two main groups. The first group focuses on parameter transfer, where knowledge

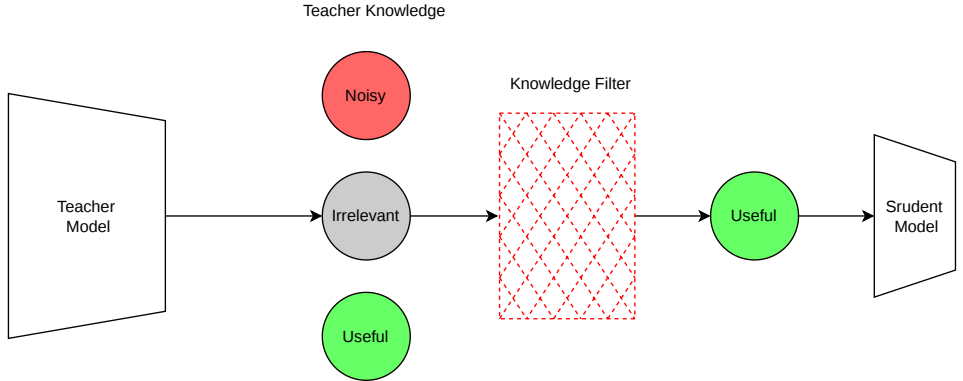


Figure 1.4: Knowledge Transfer challenge.

is transferred through the weights of pre-trained models. These approaches mainly employ meta-learning frameworks to learn good initialization parameters that can quickly adapt to new tasks. For example, Finn *et al.* [2] proposed Model-Agnostic Meta-Learning (MAML), which learns model parameters that can be fine-tuned with a small number of gradient steps on new tasks. Antoniou *et al.* [28] extended this idea by incorporating a meta-learning optimizer to improve the adaptation efficiency. Despite their effectiveness, these methods often require to tune the adaptation process carefully and may suffer from overfitting when there is a large domain gap between base and novel classes. In contrast, the second group [12,20,21,22,23] emphasizes feature-level knowledge transfer, which attempts to learn transferable representations that generalize well to novel classes. Jamal *et al.* [29] proposed a task-agnostic meta-learning approach that explicitly separates knowledge transfer from task-specific adaptation, enabling more effective cross-task generalization. To address severe domain shift in cross-domain few-shot learning, Paeedeh *et al.* [30] introduce ADAPTER, which uses a transformer with bidirectional cross-attention to learn transferable features between base-domain and target-domain representations.

Despite these advances, knowledge transfer in FSL still faces several key challenges: The semantic gap between base and novel classes may lead to negative knowledge transfer when the knowledge is not properly aligned; Limited supervision in novel tasks makes it difficult to preserve the discriminative power of transferred knowledge; The diversity of potential novel tasks requires the transfer mechanism to be flexible and adaptive. To address these challenges, recent works have explored more sophisticated knowledge transfer strategies. Cheraghian *et al.* [31] propose a semantic-aware knowledge distillation framework for few-shot class-incremental learning, which leverages word embeddings and an attention mechanism over multiple visual embeddings to align visual and semantic representations and mitigate

catastrophic forgetting. Chen *et al.* [32] develop a knowledge graph transfer network to explicitly model relationships between base and novel classes.

Different from previous approaches, we introduce a dual selective knowledge transfer (DSKT) framework that explicitly preserves the most transferable components of the teacher’s representation while suppressing task-irrelevant or noisy signals, as detailed in Chapter 6. DSKT performs selective distillation from both the perspective of global semantic knowledge and the task-specific representations extracted from the limited support samples. By selectively integrating the most beneficial components from both sources, DSKT enables the student model to maintain a more stable and discriminative embedding space when learning new categories with minimal annotated data.

1.3 Thesis Outline and Research Questions

In Section 1.2, we have briefly reviewed the recent progress from three research directions: 1) metric learning; 2) feature enrichment and 3) knowledge transfer. Few-shot learning has become an increasingly significant topic in the image understanding area. For a better understanding of FSL, we first present a comprehensive review of existing researches related to FSL. Our review includes the main challenges, benchmarks and strategies specialized for model robustness in FSL tasks. Chapter 2 is based on the published journal paper [33]:

He, K., Pu, N., Lao, M., and Lew, M.S. “Few-shot and meta-learning methods for image understanding: a survey.” *International Journal of Multimedia Information Retrieval*, 2023.

Although deep learning has driven state-of-the-art results in many FSL benchmarks, it still faces many limitations and challenges, such as sensitivity to label scarcity, reliance on large-scale pre-training, limited interpretability of learned metrics, and difficulties in generalizing to highly diverse or shifting task distributions. There remains considerable space to further improve the flexibility, robustness, and theoretical understanding of FSL models. In the following research chapters (Chapters 3–6), we present novel approaches to address the research questions (RQs) and challenges in terms of three research themes, including more discriminative metric formulations, enriched feature representations under data scarcity, and more effective knowledge transfer strategies. In Chapter 7, we summarize our main findings, limitations, potential solutions, and promising directions for future research.

- In Chapter 3, we conduct an in-depth investigation into the fundamental limitations of current metric learning frameworks in FSL scenarios. Through extensive empirical analysis, we find two primary reasons: (1) Most existing

methods employ static metric spaces that fail to dynamically accentuate discriminative features when presented with novel categories. (2) The lack of training samples make the model prone to overfitting and unstable calibration across episodes. These findings lead to our first research question **RQ 1: How to design a unified framework that simultaneously enhances metric discriminability and ensures training stability?** To address this challenge, we introduce a novel approach that strategically combines cosine sharpening transformation with elastic regularization mechanism. The cosine sharpening transformation operates as an adaptive feature recalibration module that dynamically sharpens angular decision boundaries between similar categories, effectively amplifying subtle discriminative signals that are crucial for few-shot identification. Concurrently, our elastic regularization framework provides multi-scale constraint adaptation, allowing the model to maintain representation stability across episodes while preserving the flexibility necessary for rapid task-specific adaptation. This dual mechanism enables our framework to achieve superior performance on both initial adaptation and cross-episode generalization, effectively bridging the gap between metric sensitivity and training stability. This work is based on the accepted paper:

He, K., Pu, N., Bakker, E. M., and Lew, M. S. “ProtoSimNet: Exploiting Prototype Similarity for Few-shot Learning.” accepted by the Neurocomputing journal, 2026.

- In order to enhance the model’s capability for rapid adaptation under limited supervision, our investigations in Chapter 2 and 3 primarily focus on improving feature robustness through metric-based and meta-learning paradigms. In Chapter 4, however, as we further explored the limitations of existing FSL techniques, we observed that many conventional strategies—especially those relying heavily on single-scale representations—tend to impose overly rigid constraints on feature adaptation. Such constraints restrict the model’s adaptation to the diverse scale distributions across different episodes. This observation leads us to the second research question **RQ 2: How to develop a flexible multi-scale representation framework that can dynamically reconcile feature hierarchies while preserving discriminative capability under limited supervision?** Given the inherent multi-scale nature of visual recognition tasks, we propose to exploit an attention-guided feature pyramid construction as the fundamental architecture. However, traditional feature pyramid networks typically rely on fixed fusion strategies and extensive parameterization, which are particularly prone to overfitting in few-shot tasks. To address these challenges, we propose a novel attention-guided feature pyramid network (AFPN) that integrates two innovative components: a simple,

parameter-free attention module (SimAM) for efficient channel-level feature refinement, and a cross-layer spatial attention mechanism (CSAM) for adaptive multi-scale feature fusion. This integrated approach enables our framework to maintain architectural simplicity while achieving remarkable adaptability across varying scale distributions and domain characteristics. This work is based on the published conference paper [34]:

He, K., Bakker, E. M., and Lew, M. S. “Attention-guided Feature Pyramid Network for few-shot learning.” Benelux Conference on Artificial Intelligence and Machine Learning, 2024.

- In Chapter 5, we aim to investigate the role of multi-scale feature representation learning in addressing the limitations of few-shot classification by answering the third research question **RQ 3: How can we effectively enhance feature richness and discriminativeness under extremely data scarcity?** As discussed in Section 1.2.1, most contemporary FSL approaches mainly focus on either metric optimization or meta-learning strategies. While these methods have demonstrated considerable successes in standard benchmark scenarios, they often overlook the fundamental challenge of insufficient feature diversity caused by limited training samples, particularly when dealing with fine-grained categories or significant intra-class variations. To address this gap, we introduce a novel Adaptive Multi-scale Feature Fusion (AMFF) framework specifically designed for FSL. AMFF integrates two key components: a Feature Enrichment Module (FEM) that expands feature diversity via multi-branch convolutional extraction, and a Feature Fusion Module (FFM) that adaptively reweights and aggregates multi-scale representations to enhance discriminative capability. These mechanisms allow the model to learn more diverse visual representations and stronger category-discriminative information from only a few labeled samples. This chapter is based on the published conference paper [35]:

He, K., Bakker, E. M., and Lew, M. S. “Adaptive Multi-scale Feature Fusion Network for Few-Shot Learning.” International Conference on Intelligent Computing, 2025.

- Although enhancing representation learning can improve a model’s ability to recognize new categories from limited data, conventional feature extractors often fail to obtain good generalizable feature representations for FSL as the representations learned from limited samples may not adequately capture the diversity that required to distinguish novel classes. In Chapter 6, we therefore introduce our fourth research question **RQ4: How can transferable**

knowledge be selectively leveraged to build more generalizable feature representations for few-shot tasks? Motivated by the effectiveness of teacher-student learning paradigms in transferring learned knowledge, we believe that a well-designed selective transfer mechanism can enable models to learn more robust and discriminative features. To this end, we propose a Dual Selective Knowledge Transfer (DSKT) framework that consists of two stages. First, we employ a multi-task pre-training methodology to develop a teacher model with robust, transferable feature representations; Second, we implement a dual selective knowledge distillation mechanism that enables the student model to intelligently integrate knowledge from both the teacher model and current task-specific samples. This selective knowledge transfer effectively strengthens the student model’s capability to generalize to unseen categories. This chapter is based on the published journal paper [36]:

He, K., Pu, N., Lao, M., Bakker, E. M., and Lew, M.S. “Dual selective knowledge transfer for few-shot classification.” Applied Intelligence, 2023.

- Finally, Chapter 7 summarizes the main findings from the research of this thesis. Also, we discuss the limitations and potential solutions, and point out directions for future research.

Additionally, this thesis draws on insights and experiences from the related work in other publications during my PhD studies:

Lao, M., Pu, N., Liu, Y., **He K.**, Bakker, E. M., and Lew, M. S. “COCA: COllaborative CAusal Regularization for Audio-Visual Question Answering.” AAAI Conference on Artificial Intelligence, 2023.

1.4 Main Contributions

The contributions of this thesis can be summarized and characterized as: 1) Task Contributions; and 2) Technical Contributions.

1.4.1 Task Contributions

A new and challenging FSL task. We introduce a new and challenging FSL task that specially emphasizes recognition under multi-scale and structural variations, which is often overlooked in existing FSL benchmarks. Unlike common FSL settings which assume relatively stable object scales, we require models to generalize to novel categories whose visual characteristics may shift dramatically across spatial resolutions. This setting better reflects real-world scenarios, where class instances

frequently face scale fluctuations, local distortions, and fine-grained discriminative differences, yet only a few labeled examples are available. Different from conventional FSL tasks that are conducted on fixed-resolution features, our setup gives rise to several additional challenges:

1. Large intra-class scale variation disrupts the stability of feature representations learned from scarce samples.
2. Subtle differences between classes require models to extract fine-grained local features, where conventional single-scale baselines are unable to capture effective representations..
3. Imbalanced scale distributions across categories exacerbate data scarcity in FSL.

To achieve a comprehensive analysis of these challenges, we develop a task formulation designed to assess FSL models under substantial multi-scale variations—a factor that is typically underrepresented in existing benchmarks. Conventional few-shot settings tend to normalize object scales, either implicitly through dataset bias or explicitly through preprocessing, which fails to address the fundamental difficulties introduced by scale inconsistencies. In contrast, our formulation treats multi-scale variability as a core aspect of the problem and examines how models deal with diverse feature resolutions when supervision is extremely limited. This setting allows us to further explore several questions: how a model can combine information from different hierarchical feature levels, to what extent scale shifts degrade generalization to novel classes. Overall, this task aims to encourage the development of more adaptive and scale-aware FSL models.

1.4.2 Technical Contributions

From Chapter 3 to Chapter 6, we develop new approaches based on deep learning to address the research questions regarding the three themes. The key contributions in these approaches are listed below.

A novel metric learning method for FSL. Previous conventional metric spaces for FSL often fail to form clear decision boundaries when only a few training examples are available. In addition to the intrinsic data scarcity, FSL also faces challenges arising from high feature similarity between different categories and limited separability in low-dimensional embeddings. These limitations make metric-based methods difficult to obtain good performance in fine-grained classification tasks. To address this issues, we propose a novel prototype similarity framework for FSL tasks. To map these samples into a more discriminative embedding space, our framework applies a cosine sharpening transformation to enlarge inter-class margins during episodic training. Furthermore, an elastic regularization component is integrated to prevent overfitting due to limited support examples. The numerous experiments demonstrate the effectiveness of our framework in FSL tasks.

An effective attention-guided pyramid fusion method for FSL. FSL endures challenges due to the extremely limited amount of annotated data, making it difficult for models to form discriminative representations-especially when the visual characteristics that distinguish categories arise from different spatial scales. In fine-grained FSL scenarios, a single global (image-level) embedding can be insufficient: it tends to suppress subtle local cues, making class boundaries ambiguous [37]. To mitigate these issues, we propose an attention-guided feature pyramid network (AFPN) that explicitly leverages multi-level feature hierarchies to construct robust few-shot representations. AFPN employs a learnable attention mechanism to regulate the contribution of features originating from different depths of the backbone and to highlight structural patterns that are critical for class discrimination under few-shot supervision. A top-down fusion process then aggregates these heterogeneous representations into a coherent feature space with enhanced discriminability. The experimental results demonstrate AFPN significantly improves generalization to novel categories, and achieves a new state-of-the-art performance.

A new multi-scale feature fusion framework tailor-made for FSL. Unlike traditional recognition tasks where abundant labeled data enables models to easily obtain rich feature diversity, FSL tasks are conducted under severe data scarcity, making it difficult to learn discriminative and robust representations. Beyond the typical challenges caused by intra-class variability or fine-grained visual differences, few-shot models also face an additional limitation caused by insufficient feature diversity, making it particularly difficult for models to capture comprehensive visual patterns. To address these challenges, we propose an adaptive multi-scale feature fusion (AMFF) framework designed to systematically enrich and refine feature representations learned from only a handful of labeled examples. Specifically, we introduce a feature enrichment module (FEM) that leverages a multi-branch convolutional design to expand feature diversity by capturing complementary representations. Complementing this, a feature fusion module (FFM) adaptively recalibrates and aggregates the extracted multi-scale features with learned importance weights, enabling the network to emphasize scale-consistent and discriminative information. The experimental validation demonstrates that both components of our framework contribute significantly to its overall performance,

A new knowledge distillation method for FSL. Through an in-depth analysis of why conventional few-shot learning models struggle to generalize, we observe that feature extractors trained under limited supervision often fail to obtain robust representations and to transfer effectively to unseen classes. This issue becomes more serious when novel categories exhibit fine-grained variations that cannot be inferred from only a few labeled samples. Motivated by this limitation, we propose a dual selective knowledge transfer (DSKT) framework specifically designed to rethink the knowledge distillation process for few-shot scenarios. DSKT introduces a two-stage learning paradigm that explicitly separates the acquisition of broad visual knowledge from the task-specific adaptation process. Benefiting from its selective

1. INTRODUCTION

transfer strategy, DSKT effectively alleviates the limitations between preserving generalized visual understanding and adapting to new categories with very limited data. The numerous experiments demonstrate the effectiveness of our DSKT in FSL tasks.

Chapter 2

Few-shot & meta-learning methods for image understanding: a survey

In the previous chapter we have introduced the background and research questions for this thesis. Improving the model’s generalization with limited data is one of the main challenges in few-shot learning (FSL), and is strongly related to our proposed approaches described in Chapter 3-6, we give a comprehensive survey of recent advances that address FSL performance beyond predicting accuracy.

State-of-the-art deep learning systems (e.g. ImageNet image classification) typically require very large training sets to achieve high accuracies. Therefore one of the grand challenges is called few-shot learning where only a few training samples are required for good performance. In this survey, we illuminate one of the key paradigms in few-shot learning called meta-learning. These meta-learning methods, by simulating the tasks which will be presented at inference through episodic training, can effectively employ previous prior knowledge to guide the learning of new tasks. In this chapter, we provide a comprehensive overview and key insights into the meta-learning approaches and categorize them into three branches according to their technical characteristics, namely metric-based, model-based and optimization-based meta-learning. Due to the major importance of the evaluation process, we also present an overview of current widely-used benchmarks, as well as performances of recent meta-learning methods on these datasets. Based on over 200 papers in this survey, we conclude with the major challenges and future directions of few-shot learning and meta-learning.

This chapter is based on the following publication [33]:

- **He, K.**, Pu, N., Lao, M., and Lew, M.S. “Few-shot and meta-learning methods for image understanding: a survey.” International Journal of Multimedia Information Retrieval, 2023.

2.1 Introduction

Image classification [38, 39] is an important application in computer vision [40, 41] and machine learning [42, 43]. With the continuous development of deep learning [44, 45, 46], recent years have witnessed great breakthroughs in this area [47, 48]. However, such success relies on a huge amount of data [49, 50] (usually in the order of millions), which is difficult and time-consuming to obtain in the real world. In order to reduce the data requirement, there has been growing interest in small-sample image classification [51, 52, 53], such as few-shot classification [54, 55, 56], which learns a classification rule from few (1-5) labeled samples.

A core challenge in few-shot image classification is to alleviate the susceptibility of models to overfitting under a few-data regime [57, 58, 59]. To address this problem, researchers have proposed several promising approaches, such as transfer learning [60, 61], meta-learning [1, 2, 62] and data augmentation [63, 64, 65]. In transfer learning, a model is first trained on a source domain where abundant source data is available. Then this trained model is fine-tuned [66, 67, 68] on another target domain with few labeled target samples. The learnt prior knowledge can be transferred from source tasks to target tasks during this process. Meta-learning, or learning to learn, has emerged as one of the prominent approaches for few-shot learning. It is proposed to train a meta-learner which can quickly generalize to new tasks with few examples [69, 70, 71, 72]. A meta-learning procedure also involves learning at two levels, within and across tasks. Meta-learning approaches simulate the tasks that will be presented at inference through episodic training [73, 74, 75], enabling the generalization ability of meta-learner within minor adaption steps. Data augmentation methods are often used as pre-processing in few-shot learning (FSL). In order to solve the problem of insufficient training data, they introduce various kinds of existing data variance for the model to capture. For image classification, one commonly-used method is deformation [76, 77, 78, 79], including horizontal flipping, cropping and rotation. Besides these, more advanced methods, such as generating training samples and pseudo labels [80, 81, 82], are also an important part of data augmentation,

In this chapter, we present a survey of recent meta-learning methods for few-shot image classification. Meta-learning focuses on learning prior knowledge from previous tasks which can bring efficient downstream learning to new tasks. This learning mechanism enables models to learn new concepts quickly where only few samples are available. Meta-learning deserves special attention as it is an essential part of few-shot image classification and it has also demonstrated outstanding performance on benchmark datasets [83, 84]. To be specific, in this survey we divide meta-learning into three categories according to the different mechanisms, namely metric-based, model-based and optimization-based methods [85, 86, 87, 88].

A number of surveys on FSL have been proposed. In 2018, Shu *et al.* [52] provided an early survey on small sample learning, discussing approaches for different

scenarios (zero-shot learning [89, 90, 91] and FSL) and tasks (image classification, visual question answering [92, 93, 94] and object detection [95, 96, 97]). Wang *et al.* [98] conducted a comprehensive review in 2021, which provides a formal definition of FSL and distinguishes it from other machine learning problems, exploring FSL from a fundamental viewpoint of error decomposition in supervised learning. Li *et al.* [99] published another comprehensive review on FSL in 2021, which is entirely focused on meta-learning and review literature [100, 101, 102, 103] over a long period in this area. There is another review on few-shot image classification [104] published in 2023, which is fully devoted to metric learning methods [105, 106, 107]. Compared with these surveys [52, 98, 99, 104], our review presents an up-to-date survey of meta-learning approaches for few-shot image classification and provides a thorough analysis of these different kinds of methods to better understand their individual strengths and limitations.

2.2 The framework of few-shot image classification

2.2.1 Notation and definitions

In this section, we first present a brief introduction to few-shot learning and meta-learning, then provide the notation and unified definitions of few-shot image classification [108, 109, 110].

Few-shot learning is a surprising research area that focuses on learning patterns from a set of data (base classes) and then adapting to a disjoint set (novel classes) with limited training samples. Few-shot image classification is the one with most attention and researches. As the most popular approach for few-shot learning, meta-learning organizes the learning process into two phases, called meta-training and meta-testing. During each phase, the meta-training set or meta-testing set is split into multiple episodes. Each episode samples from the task distribution and is further divided into a small training set and a testing set.

In the standard few-shot image classification setting, two distinct datasets are involved, namely the base dataset $D_{base} = \{(x_i, y_i); x_i \in X_{base}, y_i \in Y_{base}\}_{i=1}^{N_{base}}$ and the novel dataset $D_{novel} = \{(x_i, y_i); x_i \in X_{novel}, y_i \in Y_{novel}\}_{i=1}^{N_{novel}}$, where x_i represents the original feature vector of the i -th image and y_i is the corresponding class label; N_{base} and N_{novel} denote the total numbers of instances in D_{base} and D_{novel} , respectively. The base dataset is an auxiliary dataset that is used to train the classifier to learn some prior or shared knowledge and the novel dataset is used for the classifier to perform new classification tasks. Note that D_{base} and D_{novel} are disjoint, which means $Y_{base} \cap Y_{novel} = \emptyset$. In order to train and test the classifier, the D_{novel} is usually split into the support set D_S and the query set D_Q and they share the same label space.

Definition 1. The few-shot image classification task aims to learn a classifier from D_{base} and D_S to correctly classify the samples in D_Q . It is generally termed as an N -way K -shot problem, where N and K denote the number of classes and instances in D_S , respectively. If $K = 1$, it becomes a one-shot image classification task; and if $K = 0$, then the task is called zero-shot classification.

Definition 2. A few-shot image classification task is called cross-domain few-shot image classification when the base dataset and the novel dataset are from two different domains, i.e., $X_{base} \neq X_{novel}$.

2.2.2 Datasets

In this section, we briefly introduce several well-known datasets for few-shot image classification. According to different data types, we categorize them into simple image dataset (Omniglot [111]), complex image dataset (MiniImageNet [14], TieredImageNet [62], CIFAR-FS [112] and FC100 [23]) and special image dataset (CUB-200 [113, 114]). Among these datasets, CIFAR-FS and FC100 are considered more difficult as the resolution of images from the two datasets is 32×32 . It's more challenging for models to extract useful information from low-resolution images. Statistics of these datasets and popular experimental settings are summarized below. We also present some sample images from these benchmark datasets in Figure 2.1.

Omniglot is one of the most frequently used benchmarks for evaluating few-shot image classification algorithms. It contains 1623 handwritten characters collected from 50 different alphabets. Each character consists of 20 samples, drawn by different human subjects. This dataset is usually augmented by the rotations in multiples of 90 degrees, 1200 characters are used for training and the rest for evaluation.

MiniImageNet and TieredImageNet are two mini versions of the large ImageNet dataset [13]. MiniImageNet is composed of 60,000 color images from 100 classes, with 600 images in each class. Following the widely-used splitting protocol proposed by Revi and Larochelle [115], 64 classes are used for training, 16 classes for validation and 20 classes for evaluation. TieredImageNet is another larger subset of ImageNet with a hierarchical structure. It contains 779,165 images from 34 high-level categories (or 608 classes), which are further split into 20 base categories (351 classes), 6 validation categories (97 classes), and 8 novel categories (160 classes).

CIFAR-FS and FC100 are two widely-used datasets derived from CIFAR-100 [12]. CIFAR-FS is constructed from 100 classes with 600 images per class. The 100 classes are split into 64, 16 and 20 classes for training, validation, and evaluation, respectively. FC100 also contains 100 classes, which are further divided into 20 super-categories, with five classes in each super-categories. FC100 is split into 12 base, 4 validation, and 4 novel super-categories.

CUB-200 is a fine-grained dataset consisting of 200 bird species. The CUB-200 dataset has two versions, while the initial version was proposed in 2010 [114] which



Figure 2.1: Sample images of benchmark datasets widely-used for few-shot image classification.

includes 6033 images and is extended to 11,788 images in 2011 [113]. The CUB-200-2010 dataset is often split into 130 base, 20 validation, and 50 novel classes [116]; while the CUB-200-2011 dataset is divided into 100 classes for training, 50 classes for validation, and 50 classes for testing [55].

MiniImageNet $\rightarrow$ CUB is a dataset designed for cross-domain few-shot image classification [21, 117, 118, 119]. MiniImageNet plays the role of the base dataset, while 50 classes of CUB-200-2011 are used for validation, and the remaining 50 classes serve for evaluation.

2.2.3 Evaluation process of few-shot image classification

In this section, we present a general procedure [22, 120, 121, 122, 123] to evaluate a classifier’s performance on N -way K -shot image classification problems in Algorithm 1. The whole evaluation process is composed of E episodes. In each episode, we first randomly select N classes from the novel label space with K samples in each class to form a support set D_S , and M examples from the remaining samples of those N classes to compose a query set D_Q . A final classifier can be obtained based on the base dataset D_{base} and support set D_S , which is used to predict labels of samples in

D_Q . We use $acc^{(e)}$ to denote the classification accuracy in the e -th episode, and the performance of a learning algorithm can be measured by the averaged classification accuracy over all episodes.

Algorithm 1: Evaluation process of N -way K -shot image classification

Input: D_{base} ; D_{novel} ; the total number of episodes E and a classifier f .

. **for** $e = 1, \dots, E$ **do**

 Randomly select N classes from D_{novel} ;

 Randomly select K samples from each class to compose the support set $D_S^{(e)}$;

 Randomly select M samples from the remaining samples of each class to compose the query set $D_Q^{(e)} = \{(X^{(e)}, Y^{(e)})\}$;

 Use the classifier f to predict labels of the samples from D_Q :

$\hat{Y}^{(e)} = f(X^{(e)} | D_{base}, D_S^{(e)})$;

 Record accuracy $acc^{(e)} = \frac{1}{M} \sum_{j=1}^M 1 \left[\hat{Y}^{(e)} = Y^{(e)} \right]$;

end

Output: mean accuracy $\frac{1}{E} \sum_{e=1}^E acc^{(e)}$.

2.3 Paradigms of meta-learning for few-shot image classification

The goal of meta-learning for few-shot image classification [124, 125, 126, 127, 128] is to enable models, especially deep neural networks, to perform well on new tasks when only few samples are available. With the rapid development of few-shot learning [129, 130, 131], a number of meta-learning approaches [132, 133, 134] have been proposed. In this section, we provide a comprehensive overview of recent meta-learning studies and their advances. In order to let beginners better understand, we follow the main trend and still categorize meta-learning into metric-based, model-based and optimization-based methods. Besides, we also present other few-shot learning methods to make a comparison. Figure 2.2 shows an overview of few-shot image classification.

2.3.1 Metric-based meta-learning

Metric-based meta-learning methods [135, 136, 137, 138] aim to learn a distance metric, which can effectively measure the similarity among samples, ensuring it is optimal for new learning tasks. For few-shot image classification problems, the learned metric should follow the rule that samples from the same (or different) class should have a small (or large) distance.

Siamese Network is one of the most widely-used metric-based methods for one-shot image classification. The term "Siamese" was first proposed for signature verification

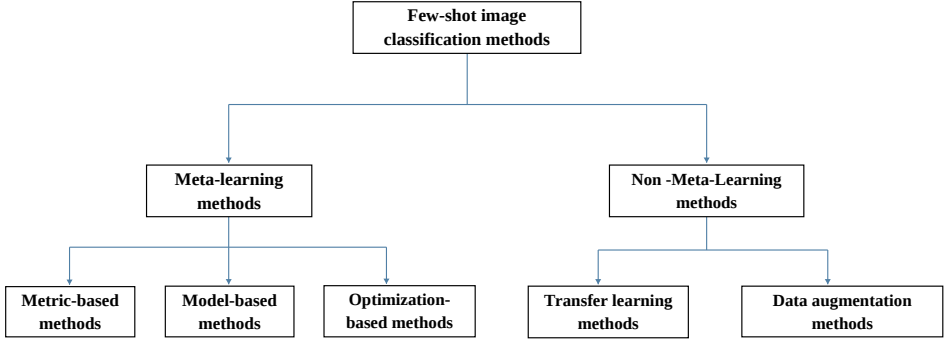


Figure 2.2: An overview of few-shot image classification approach.

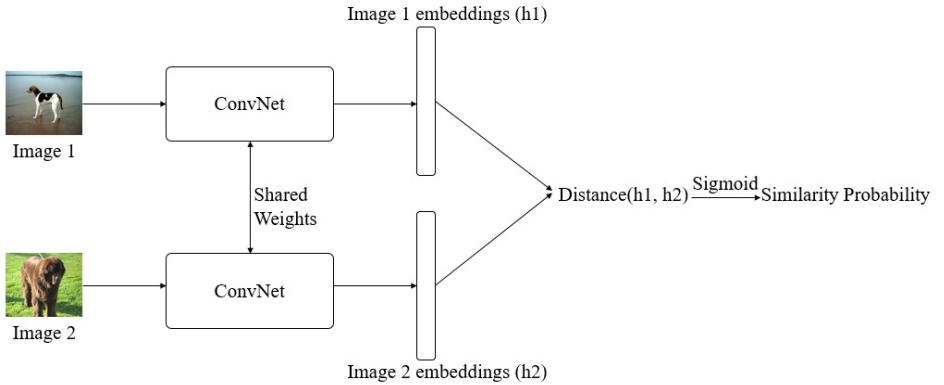


Figure 2.3: Siamese Network Architecture.

[139] and the principal structure of the Siamese Network was introduced for the fingerprint similarity estimation problem [140]. In 2015, Koch *et al.* [141] adopted a pair of identical VGG-styled [39] convolutional layers with shared weights to extract high-level features from two input images and calculate the weighted L_1 distance between the two feature vectors. The network finally outputs a score, representing the probability that the two images belong to the same class. The architecture of Siamese Network is shown in Figure 2.3. Wang *et al.* [142] proposed an Attention Based Siamese Network, which exploits an attention kernel function to measure the similarity between two feature vectors. To bridge the gap between one-shot image recognition [143, 144, 145] and regular classification, Lungu *et al.* [146] proposed a Multi-Resolution Siamese Network, which mixes different kernel size streams into one layer and adopts a hybrid training mechanism.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

As another powerful metric-based meta-learning method, Matching Network [14] uses different networks to encode support and query images. For support images embedding, a bidirectional Long-Short Term Memory (LSTM) [147] is used in the context of the support set D_S ; for query images embedding, an LSTM with an attention kernel is taken to enable the dependency on D_S , where the attention kernel [148, 149] is used to compute cosine similarities between support and query images and then normalize the similarities through a softmax function. Matching Network’s output is defined as a sum of the labels (one-hot encoded) of support images weighted by the attention kernel. In 2019, Mai *et al.* [150] proposed an Attentive Matching Network (AMN), introducing a feature-level attention mechanism to pay more attention to the features that can better reflect the inter-class differences and a complementary cosine loss function for optimization.

The initial Prototypical Network was proposed by Snell *et al.* [1] based on the hypothesis that there exists an embedding space where each class can be represented by a unique prototype, and all samples are supposed to cluster around their corresponding prototypes. Figure 2.4 shows the architecture of a Prototypical Network. A simple convolutional neural network with 4 layers is exploited to extract features, and the prototype of each class is defined as the mean value of feature embeddings from the support samples belonging to that class. The squared Euclidean Distance is employed as a distance metric, calculating the distance between query embeddings and each class prototype. Build on this, Li *et al.* [151] proposed a Covariance Metric Network (CovaMNet), using the covariance matrix of embedding vectors to represent the class prototype and also apply a covariance-based metric to measure the similarity between the query sample and the class prototype. Wang and Zhai [152] proposed a Prototypical Siamese Network (PSN), adding a prototype module in Siamese Network to obtain high-quality prototype representations of each class.

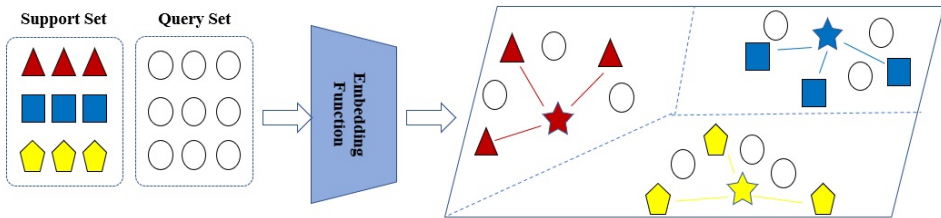


Figure 2.4: Prototypical Network Architecture.

Relation Network [24] is the first study that employs a neural network to estimate a similarity score of feature embeddings rather than manual computation. This model consists of two main components: an embedding module and a relation module. The embedding module is composed of convolutional blocks, mapping input images into an embedding space; and the relation module builds on two convolutional blocks and two fully-connected layers, calculating a relation score between each query and

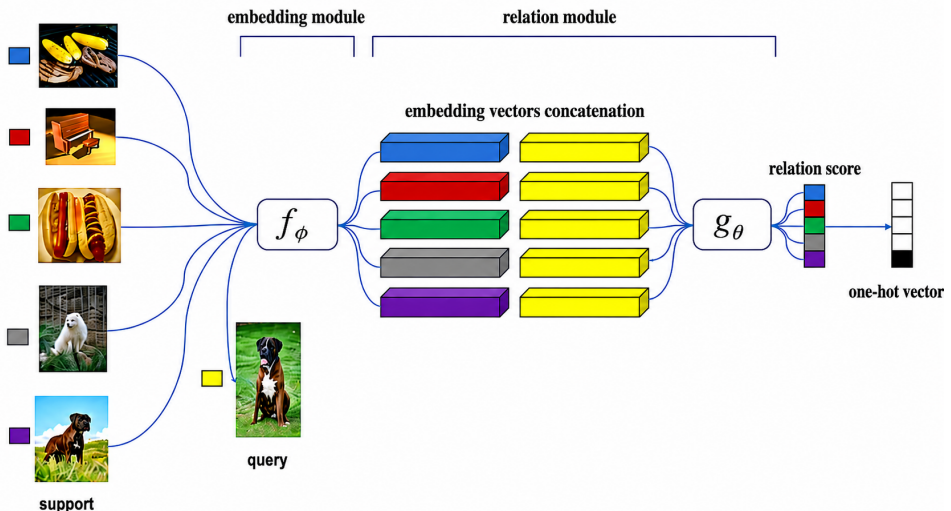


Figure 2.5: Relation Network Architecture. We use different colors to represent samples from different classes, where f_ϕ and g_θ denote embedding module and relation module, respectively

support image (or a class prototype when the number of support samples is more than one). Note that the feature embeddings of support and query images need to be concatenated together before they are fed into the relation module. The architecture of Relation Network is presented in Figure 2.5. In order to obtain discriminative features for fine-grained image classification [153, 154, 155], the subsequent work [156] proposed a Bi-Similarity Network (BSNet), which combines an extra cosine module with the existing similarity measure as a new relation module, generating a more compact feature space by forcing features to adapt to the new relation module.

In order to get optimal matching image regions, Zhang *et al.* [37] proposed a Deep-EMD algorithm, which adopts the Earth Mover’s Distance (EMD) [157, 158, 159] as a distance metric to calculate the similarity. They introduce a cross-reference mechanism to produce the weights of elements in the EMD formulation and embed the EMD layer into the network for end-to-end training. Motivated by this, Xie *et al.* [160] proposed a deep Brownian Distance Covariance (DeepBDC) approach, which applies the BDC metric for few-shot learning. To learn discriminative feature representations, Afrasiyabi *et al.* [161] proposed a Mixture-based Feature Space Learning (MixtFSL) approach, learning both the feature representations and the mixture model via an online manner. Different from those few-shot classification methods that extract a single feature vector from each image, Afrasiyabi *et al.* [162] held the view that a set-based representation can build a richer and more robust representation of images from base classes. To do so, they proposed a Matching Feature

Sets method which embeds self-attention modules in between convolutional blocks and introduces set-to-set metrics for evaluation. We summarize those introduced metric-based meta-learning approaches in Table 2.1.

2.3.2 Model-based meta-learning

With the goal of fast learning, model-based methods [163, 164] mainly focus on model architectures, adjusting model parameters based on presented tasks. There are several frequently used architectures in model-based methods, such as Convolutional Neural Networks (CNNs) [165], Recurrent Neural Networks (RNNs) [166, 167], Long Short-Term Memory (LSTM) [168], etc. According to the model architecture types, these model-based methods are further separated into memory-based, rapid-adaptation-based and miscellaneous models.

Memory-Augmented Neural Network(MANN) is a famous memory-based method proposed by Santoro *et al.* [169], which aims at improving task adaptation by utilizing the Neural Turing Machine (NTM) [170, 171, 172]. NTM is a neural network that integrates an external memory component during its learning process, enabling it has access to retrieve previously stored information. To be specific, NTM consists of a controller, interacting with an external memory module via a number of read and write heads. The NTM scheme is shown in Figure 2.6. In MANN, a new addressing mechanism namely least recently used access (LRUA) [169] is proposed, writing memories to either the least used memory location or the most recently used memory location. Through the stored information of a coupled representation-class label in the external memory, MANN can access them for later classification. Tran *et al.* [173] proposed a Memory Augmented Matching Network (MAMN), which combines MANN and Matching Network. In MAMN, to reduce the biased on class prototypes caused by data distribution skew, weighted class prototypes are introduced by incorporating the distances of class-wise samples. As another memory-based meta-learning method, Memory Matching Network (MM-Net) [174] incorporates the memory module extracted from Key-Value Memory Network [175] into Matching Network. Different from traditional one-shot learning methods, MM-Net encodes and generalizes the whole support set into memory slots and can generate a unified model regardless of the number of shots and categories.

Meta Network (MetaNet) [176] is a model that designed with specific architecture and training process for rapid adaption across tasks. Meta Network contains a base learner, a meta-learner and an external memory. It performs a generic knowledge acquisition in a meta space and shifts its inductive biases via fast parameterization for rapid generalization. Conditional Shifted Neurons (CSNs) [177] is a generic neural mechanism designed for fast adaption, which is able to extract conditional information and generate conditional shifts for prediction during the meta-learning process. Compared with previous works [169, 176, 178], CSNs is more efficient computationally as the number of neurons is usually much smaller than that of weight

2.3 Paradigms of meta-learning for few-shot image classification

Table 2.1: A summary of presented metric-based meta-learning approaches.

Approach	Author	Year	Brief Description
Siamese Network [141]	Koch <i>et al.</i>	2015	Exploit a pair of identical networks with shared weights to extract features and calculate the distance between feature vectors.
Attention Based Siamese Network [142]	Wang <i>et al.</i>	2018	Use an attention kernel function to measure the similarity.
Multi-Resolution Siamese Network [146]	Lungu <i>et al.</i>	2020	Mix different kernel size streams into one layer.
Matching Network [14]	Vinyals <i>et al.</i>	2016	Use different networks to encode support and query images.
AMN [150]	Mai <i>et al.</i>	2019	Introduce a feature-level attention mechanism.
Prototypical Network [1]	Snell <i>et al.</i>	2017	Propose class prototypes and calculate the distances between query samples and class prototypes.
CovaMNet [151]	Li <i>et al.</i>	2019	Apply the covariance matrix of embedding vectors to represent the class prototype and use a covariance-based metric to measure the similarity.
PSN [152]	Wang and Zhai	2020	Add a prototype module in Siamese Network.
Relation Network [24]	Sung <i>et al.</i>	2018	Employ a neural network to estimate a similarity score of feature embeddings.
BSNet [156]	Li <i>et al.</i>	2021	Combine an extra cosine module with the existing similarity measure as a new relation module.
DeepEMD [37]	Zhang <i>et al.</i>	2020	Adopt the Earth Mover’s Distance (EMD) as a distance metric to calculate the similarity.
DeepBDC [160]	Xie <i>et al.</i>	2022	Apply the Brownian Distance Covariance (BDC) metric.
MixtFSL [161]	Afrasiyabi <i>et al.</i>	2021	Learn the feature representations and the mixture model jointly in an online manner.
Matching Feature Sets [162]	Afrasiyabi <i>et al.</i>	2022	Embed self-attention modules between convolutional blocks and introduce set-to-set metrics.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

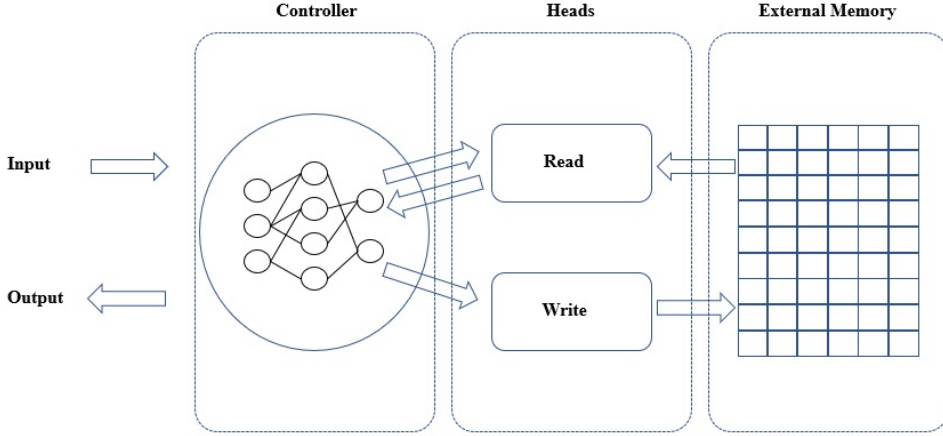


Figure 2.6: Neural Turing Machine scheme.

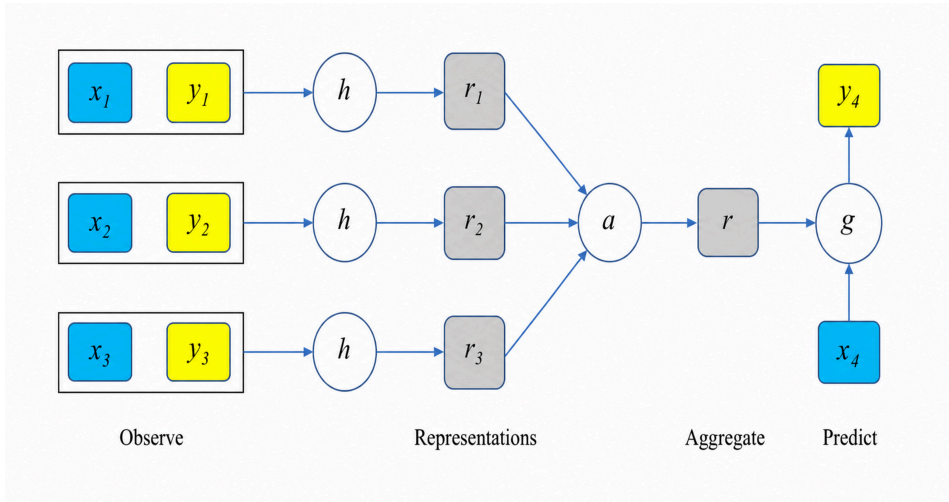


Figure 2.7: Conditional Neural Processes scheme. The encoder h maps each observed context pair (x_i, y_i) to an individual representation r_i . The permutation-invariant aggregation operator a combines these representations into a global context representation r . The decoder g then uses r together with the target input x_4 to predict the corresponding target output y_4 .

parameters. Moreover, CSNs can be integrated into various neural architectures, including CNNs and RNNs. Similar to MetaNet, CSNs contains a base learner, a meta learner and a memory module. During the description time, the meta learner extracts and employs conditional information to generate memory values for samples within a task; at the prediction phase, the meta learner generates query keys of query images by a key function for the purpose of getting the value of conditional shift.

Simple Neural Attentive Learner (SNAIL) [179] is a general model-based meta-learning architecture that incorporates temporal convolution and soft attention mechanism. The temporal convolution acts as high bandwidth memory access, and the soft attention enables access to specific pieces of information. This combination enables models to better leverage information from past experiences. Similar to SNAIL, Garnelo *et al.* [180] proposed Conditional Neural Processes (CNPs) which consists of a meta-learner and task-learner. The meta-learner generates a memory value by aggregating representations of the support set, and the task-learner makes predictions by processing the aggregated representations. Figure 2.7 shows the CNPs scheme. We also make a short summary of those model-based meta-learning approaches and present it in Table 2.2.

Table 2.2: A summary of presented model-based meta-learning approaches.

Approach	Author	Year	Brief Description
MANN [169]	Santoro <i>et al.</i>	2016	Employ a modified NTM to quickly assimilate new data into memory.
MAMN [173]	Tran <i>et al.</i>	2019	Combine MANN and Matching Network.
MM-Net [174]	Cai <i>et al.</i>	2018	Incorporate the memory module extracted from Key-Value Memory Network into Matching Network.
MetaNet [176]	Munkhdalai and Yu	2017	Perform a generic knowledge acquisition in a meta space and shift its inductive biases via fast parameterization.
CSNs [177]	Munkhdalai <i>et al.</i>	2018	Extract conditional information and generate conditional shifts for prediction during the meta-learning process.
SNAIL [179]	Mishra <i>et al.</i>	2018	Incorporate temporal convolution and soft attention mechanism.
CNPs [180]	Garnelo <i>et al.</i>	2018	Make predictions based on concise representations of seen classes.

2.3.3 Optimization-based meta-learning

Optimization-based meta-learning methods are an important vital branch in the field of few-shot image classification [181, 182, 183, 184, 185]. Basically, this kind of algorithm attempts to obtain a better initialization model or gradient descent direction by leveraging the meta-learning architecture, and optimizes the initialization parameters through episodic training, enabling an optimization procedure to work on a small number of training samples. Optimization-based methods generally contain a task-specific learner trained for a given task and a meta-learner trained on distributions of tasks.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

In 2017, Finn *et al.* [2] proposed Model-Agnostic Meta-Learning (MAML), the first algorithm for learning an initialization. The key idea of MAML is to enable the model’s parameters to adapt fast to new unseen tasks through a gradient-based learning rule. During the meta-training phase, MAML attempts to update the task-specific parameters and the global initialization jointly in an iterative manner. The MAML scheme is presented in Figure 2.8. The main contribution of MAML is its compatibility in different application domains, not only in classification, but also in regression [186, 187, 188] and reinforcement learning [189, 190, 191]. To address the limitation of neural networks that are trained with gradient-based optimization on few-shot learning tasks [84, 192, 193], Ravi and Larochelle [115] proposed an LSTM-based meta-learner to learn both the exact task-specific optimization of a classifier, as well as good initialization values for the parameters of task-specific learner.

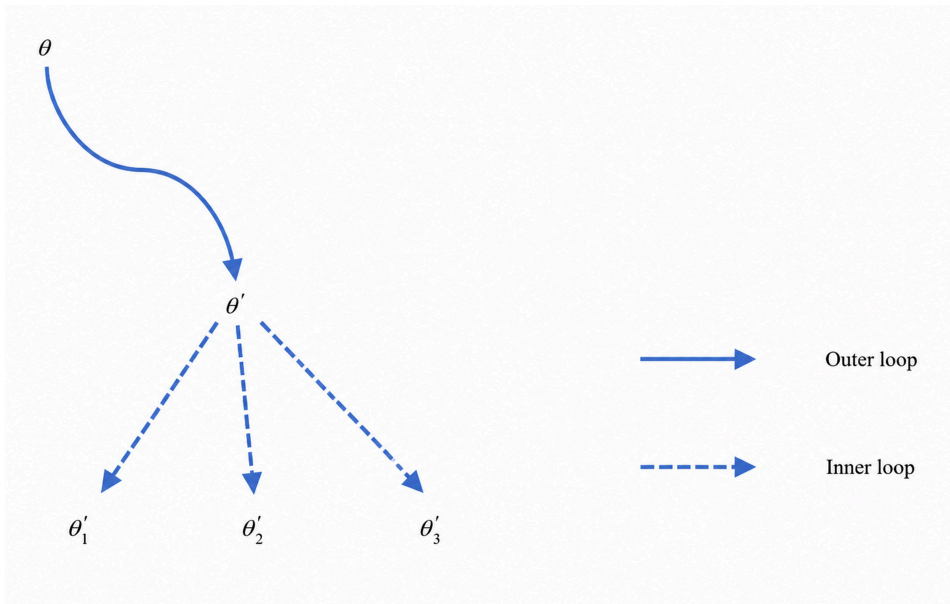


Figure 2.8: Model-Agnostic Meta-Learning scheme. The outer loop updates the shared model parameters from θ to θ' . The inner loop adapts the model separately to each sampled task, producing task-specific parameters θ'_i .

By taking ideas from Prototypical Network and MAML, Triantafillou *et al.* [194] proposed Proto-MAML, incorporating the advantages of both the former’s simple inductive bias and the latter’s flexible adaptation mechanism. As an extension to MAML, CAVIA [195] divides the model into parameters and task-specific context parameters which are shared across tasks. Compared with MAML, CAVIA is less prone to meta-overfitting and easier to parallelize. To address the issue that meta-learning models would be too biased toward existing tasks and lead to poor generalization, Jamal and Qi [29] proposed a Task Agnostic Meta-Learning (TAML) algorithm, where two approaches are exploited to train a model unbiased over tasks.

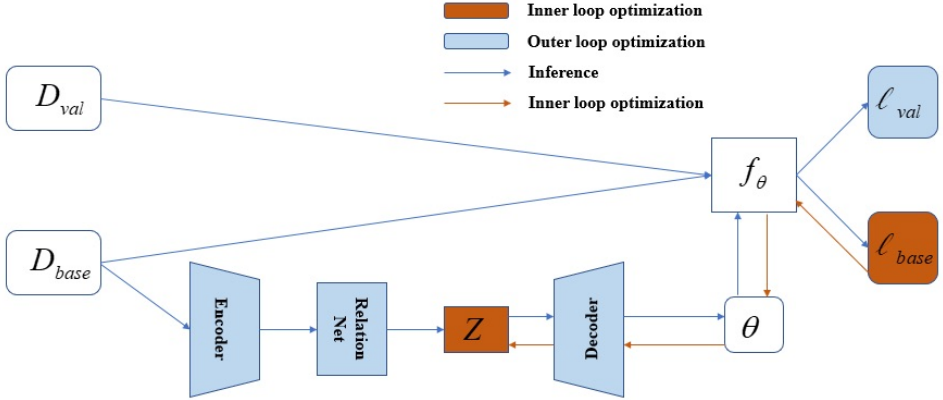


Figure 2.9: Latent Embedding Optimization scheme. Here, z denotes the latent embedding learned from D_{base} , θ denotes the learnable model parameters, and f_{θ} denotes the prediction function parameterized by θ . The model is optimized through an inner loop based on the base loss ℓ_{base} and an outer loop based on the validation loss ℓ_{val} .

In order to improve generalization performance, BaiK *et al.* [196] proposed a novel framework called Meta-Learning with Task-Adaptive Loss Function (MeTAL). Particularly, MeTAL learns a task-adaptive loss function through two meta-learners and can be applied to different MAML variants.

Wang *et al.* [197] introduced a new approach called task-aware feature embeddings for low-shot learning (TAFE-Net) which mainly concentrates on tuning task-specific feature embedding through the generic embedding of a meta-learner. TAFE-Net is composed of a meta-learner and a prediction network, where the task-aware feature embedding is obtained by utilizing the meta-learner to develop task-specific feature layers of the prediction network. Sun *et al.* [198] introduced a meta-transfer learner (MTL) method, which focuses on generating task-specific feature extractors by leveraging both meta-learning and transfer learning. In MTL, scaling and shifting operations are introduced on pre-trained feature embeddings to freeze the feature extractor. Besides, similar fine-tuning steps are taken in MTL as those in previous work [55]. This work also proposed a novel hard task meta-batch process that put more focus on hard tasks through sampling extra instances from the classes that the classifier failed.

Considering difficulties that exist in optimization on high-dimensional parameter spaces such as those faced by MAML [2], Rusu *et al.* [199] proposed an innovative algorithm called Latent Embedding Optimization (LEO) that learns a low-dimensional latent representation of model parameters and performs optimization-based meta-learning in this space. Similar to MAML, LEO also consists of an inner loop training where the task-specific values are learned and an outer loop training where global shared initializations are updated. To instantiate low-dimensional latent embedding

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

of model’s parameters, samples pass through a combination of an encoder and a relation network. The encoder is used to generate hidden codes from the support set. Then, these hidden codes are concatenated pairwise and fed into a relation network, leading to a probability distribution over latent codes in a lower dimension. Finally, the decoder produces task-specific initial parameters which are differentiable to backpropagate for adaptation. The LEO scheme is shown in Figure 2.9. We present a short summary of optimization-based meta-learning approaches in Table 2.3.

Table 2.3: A summary of presented optimization-based meta-learning approaches.

Approach	Author	Year	Brief Description
MAML [2]	Finn <i>et al.</i>	2017	Enable a model’s parameters can adapt fast to new unseen tasks through the gradient-based learning rule.
LSTM-based meta-learner [115]	Ravi and Larochelle	2017	Learn both the exact task-specific optimization of a classifier and good initialization values for the parameters of task-specific learner.
Proto-MAML [194]	Triantafillou <i>et al.</i>	2020	Combine Prototypical Network and MAML.
CAVIA [195]	Zintgraf <i>et al.</i>	2019	Divide the model into parameters and task-specific context parameters which are shared across tasks.
TAML [29]	Jamal and Qi	2019	Train a model unbiased over tasks.
MeTAL [196]	Baik <i>et al.</i>	2021	Learn a task-adaptive loss function through two meta-learners.
TAFE-Net [197]	Wang <i>et al.</i>	2019	Tune task-specific feature embedding through the generic embedding of a meta-learner.
MTL [198]	Sun <i>et al.</i>	2019	Generate task-specific feature extractors by leveraging both meta-learning and transfer learning.
LEO [199]	Rusu <i>et al.</i>	2019	Learn a low-dimensional latent representation of model parameters.

2.3.4 Other methods

Transfer learning involves leveraging knowledge learned from a related task to enhance learning in a new task [200, 201, 202, 203, 204]. In the few-shot image classification scenario, transferring knowledge from another network is a viable option when original data is too limited to train a deep neural network from scratch. Compared with meta-learning, the learning experience involved in transfer learning is much narrower. To address few-shot hyperspectral image classification problems, Qu *et al.* [205] applied the transfer learning scheme to extract learned intrinsic

representations from the same kind of objects in different domains. Tai *et al.* [206] proposed a novel few-shot transfer learning approach for synthetic aperture radar image classification, which uses a connection-free attention module to transfer features from a source network to a target network. Sun and Yang [207] proposed Trans-transfer Learning, a two-phase learning method for few-shot fine-grained visual categorization problems. In some cases, knowledge transfer may also fail when the source domain and target domain are not related to each other, even causing negative transfer. To address this problem, Liu *et al.* [208] proposed an Analogical Transfer Learning (ATL), following the analogy strategy to effectively control the occurrence of negative transfer.

Considering the fundamental problem in few-shot image classification that models are prone to overfitting caused by few training samples, many researchers proposed a number of data augmentation approaches [209, 210, 211] to improve sample diversity and prevent overfitting during training. Goodfellow *et al.* [212] proposed the well-known Generative Adversarial Nets (GAN), which contains a generator for generating similar images and a discriminator for distinguishing. Based on GAN, Mehrotra and Dukkipati [213] proposed to generate samples for specific tasks, enabling these generated samples more suitable for few-shot learning. Zhang *et al.* [214] proposed MetaGAN. To help the classifier learn a clearer decision boundary, MetaGAN involves GAN and part of the classification network during the training process. Li *et al.* [26] proposed Adversarial Feature Hallucination Network (AFHN), using conditional Wasserstein Generative Adversarial Network (cWGAN) to generate samples.

We present experimental results of recent meta-learning methods in Table 2.4 and Table 2.5. Table 2.4 shows performances of different approaches on Omniglot. Omniglot is a handwritten dataset with multiple handwriting styles, languages and stroke types, this diversity makes Omniglot suitable for training deep learning algorithms. From Table 2.4 we can find that most meta-learning approaches obtain over 98% accuracies on Omniglot. Table 2.5 shows experimental results on Mini-ImageNet and TieredImageNet. These two datasets contain images with different objects, scenes and lighting conditions, which can improve the model’s robustness. However, the limitations in dataset size and image quality may affect the model’s performance. Table 2.5 shows that DeepBDC [160] and Matching Feature Sets [162] achieved best results on both datasets.

2.4 Major Challenges and future directions

Although meta-learning methods have achieved promising performance in few-shot image classification, there remain some vital challenges that ought to be dealt with in the future. These existing issues and suggested future research directions are outlined here.

2. FEW-SHOT & META-LEARNING METHODS FOR IMAGE UNDERSTANDING: A SURVEY

Table 2.4: Accuracy results on Omniglot dataset reported in original papers, with mean accuracy (%) and 95% confidence interval. i: Metric-based; ii: Model-based; iii: Optimization-based

Model	Category	Omniglot 5-way	
		1-shot	5-shot
Siamese Network [141]	i	97.30	98.40
Attention Based Siamese Network [142]	i	99.60	99.80
Multi-Resolution Siamese Network [146]	i	-	-
Matching Network [14]	i	98.10	98.90
AMN [150]	i	99.44 ± 0.09	99.86 ± 0.06
Prototypical Network [1]	i	98.80	99.70
CovaMNet [151]	i	-	-
PSN [152]	i	-	-
Relation Network [24]	i	99.60 ± 0.20	99.80 ± 0.10
BSNet [156]	i	-	-
DeepEMD [37]	i	-	-
DeepBDC [160]	i	-	-
MixtFSL [161]	i	-	-
Matching Feature Sets [162]	i	-	-
MANN [169]	ii	82.80	94.90
MAMN [173]	ii	98.90	99.70
MM-Net [174]	ii	99.28 ± 0.08	99.77 ± 0.04
MetaNet [176]	ii	98.45	-
CSNs [177]	ii	98.42 ± 0.21	99.37 ± 0.28
SNAIL [179]	ii	99.07 ± 0.16	99.78 ± 0.09
CNPs [180]	ii	95.30	98.50
MAML [2]	iii	98.70 ± 0.40	99.90 ± 0.10
LSTM-based meta-learner [115]	iii	-	-
Proto-MAML [194]	iii	-	-
CAVIA [195]	iii	-	-
TAML [29]	iii	99.47 ± 0.25	99.83 ± 0.09
MeTAL [196]	iii	-	-
TAFE-Net [197]	iii	-	-
MTL [198]	iii	-	-
LEO [199]	iii	-	-

Table 2.5: Accuracy results on MiniImageNet and TieredImageNet datasets reported in original papers, with mean accuracy (%) and 95% confidence interval. i: Metric-based; ii: Model-based; iii: Optimization-based

Model	Category	MiniImageNet 5-way		TieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
Siamese Network [141]	i	-	-	-	-
Attention Based Siamese Network [142]	i	51.20	69.70	-	-
Multi-Resolution Siamese Network [146]	i	-	-	-	-
Matching Network [14]	i	46.60	60.00	-	-
AMN [150]	i	54.97 ± 0.77	71.84 ± 0.67	-	-
Prototypical Network [1]	i	49.42 ± 0.78	68.20 ± 0.66	-	-
CovaMNet [151]	i	51.19 ± 0.76	67.65 ± 0.63	-	-
PSN [152]	i	48.70	69.40	-	-
Relation Network [24]	i	50.44 ± 0.82	65.32 ± 0.70	-	-
BSNet [156]	i	-	-	-	-
DeepEMD [37]	i	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
DeepBDC [160]	i	67.83 ± 0.43	85.45 ± 0.29	73.82 ± 0.47	89.00 ± 0.30
MixtFSL [161]	i	64.31 ± 0.79	81.66 ± 0.60	70.97 ± 1.03	86.16 ± 0.67
Matching Feature Sets [162]	i	68.32 ± 0.62	82.71 ± 0.46	73.63 ± 0.88	87.59 ± 0.57
MANN [169]	ii	-	-	-	-
MAMN [173]	ii	49.80	66.50	-	-
MM-Net [174]	ii	53.37 ± 0.48	66.97 ± 0.35	-	-
MetaNet [176]	ii	49.21 ± 0.96	-	-	-
CSNs [177]	ii	56.88 ± 0.62	71.94 ± 0.57	-	-
SNAIL [179]	ii	55.71 ± 0.99	68.88 ± 0.92	-	-
CNPs [180]	ii	-	-	-	-
MAML [2]	iii	48.70 ± 1.84	63.11 ± 0.92	-	-
LSTM-based meta-learner [115]	iii	43.44 ± 0.77	60.60 ± 0.71	-	-
Proto-MAML [194]	iii	-	-	-	-
CAVIA [195]	iii	51.82 ± 0.65	65.85 ± 0.55	-	-
TAML [29]	iii	51.73 ± 1.88	66.05 ± 0.85	-	-
MeTAL [196]	iii	66.61 ± 0.28	81.43 ± 0.25	70.29 ± 0.40	86.17 ± 0.35
TAFE-Net [197]	iii	-	-	-	-
MTL [198]	iii	61.20 ± 1.80	75.50 ± 0.80	-	-
LEO [199]	iii	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09

2.4.1 Limitations and challenges

- **Data availability and computational complexity.** In image classification, a large dataset typically has a thousand (or more) categories. Meta-learning approaches also require a large amount of data and computational resources, but in few-shot scenarios, it is quite challenging to collect sufficient data. For deep testing of meta-learning we may need thousands of large datasets! These may also be very difficult and slow to process.
- **Model selection.** There is not a one-size-fits-all model so selecting an appropriate model is important. Model selection is more crucial in few-shot image classification scenarios as the model is prone to overfitting the training data. The model may perform well on the base set while lacking generalization on new tasks.
- **Transferability.** Meta-learning models can transfer learned knowledge between various tasks. The success of transferability depends on the similarity between the tasks. Sometimes new tasks may have significant differences from old ones, making it difficult to transfer learned knowledge effectively, such as cross-domain tasks.
- **Task dependence.** Most meta-learning approaches are designed to work for a specific set of tasks or domains. They may not perform well on new tasks or domains that are significantly different from the ones used during training. Improving meta-learning’s generalization ability can be a hard task.
- **Interpretability.** Interpretability is a critical aspect of neural approaches that refers to the ability to understand how a model works. Unfortunately, neural approaches can be extremely challenging to interpret, and thus difficult to understand how it learns to learn and make predictions or decisions. This issue can make it arduous to debug, diagnose and improve models’ performances.

2.4.2 Future directions

- **Enhancing generalized feature learning.** To address the main challenge in few-shot learning, learning from a handful of samples [6, 215, 216], meta-learning employs shared knowledge from previously experienced tasks for unseen tasks. However, in most existing meta-learning methods, researchers attempt to learn discriminative features via attention mechanisms, multi-task learning, data augmentation, and so on. One major research direction is developing new approaches for learning features that generalize better to new domains; and evaluation measures for assessment and selection of the learned features.

- **Practice of episodic training strategy.** In order to realize fast adaption to new tasks with limited samples, episodic training requires that each training episode should have the same number of classes and examples as the evaluation episode. But, this setting is prone to catastrophic forgetting [217, 218, 219] and leads to model underfitting in base classes. A number of approaches have been proposed to address this issue, further improving model performance on both base and novel classes remains a vital direction for future work.
- **Improving stability.** Despite the continuous improvement of meta-learning in few-shot image classification, one existing issue is that some meta-learning methods obtain state-of-the-art performance on special datasets, but perform not well on other benchmarks. For example, a metric-based meta-learning method named global class representation (GCR) [220] achieved great performance on Omniglot, but cannot compete with other non-metric-based methods on miniImageNet. Further exploration of stable models [221, 222] will be very valuable.
- **Cross-domain and multimodal meta-learning.** In principle, the base dataset D_{base} and novel dataset D_{novel} in few-shot learning can be from different domains [223, 224]. However, most model performances will decline when the difference between D_{base} and D_{novel} . Developing meta-learning methods on cross-domain performance can be one future research direction. Multimodal deep learning has also brought great opportunities to few-shot learning [225, 226, 227]. For example, Peng *et al.* [228] proposed a Knowledge Transfer Network (KTN), which combines semantic features and image features for few-shot image classification tasks. Therefore, how to design a more appropriate multimodal fusion method is a research trend in few-shot image classification.

2.5 Chapter Conclusions

This chapter presents a survey comprised of over 200 papers on recent few-shot learning and meta-learning research for image understanding. Based on the research literature, we introduce the general approaches for few-shot learning and then turn to one of the key approaches called meta-learning. We separate existing meta-learning methods into three important categories: metric-based, model-based and optimization-based methods. We introduce both classical and state-of-the-art approaches in each category, and summarize the state-of-the-art. We also present the state-of-the-art performance of the literature approaches on well-known datasets. According to our study, we conclude with limitations, challenges and weaknesses for meta-learning and present promising directions of meta-learning from the perspectives of generalization, effectiveness and applicability.

Chapter 3

ProtoSimNet: Exploiting Prototype Similarity for Few-shot Learning

In the previous chapter we have introduced the background and research questions for this thesis. Starting with this chapter, we begin to answer the research questions with our proposed approaches. In this chapter, we address the question of how to design a unified framework for FSL that simultaneously enhances metric discriminability and ensures training stability (**RQ 1**).

Many few-shot learning (FSL) approaches exploit Euclidean Distance as the metric, as it provides a stable and well-understood geometric structure for prototype-based classification. However, cosine similarity has become increasingly popular in normalized embedding spaces due to its scale invariance and computational efficiency. Meanwhile, with limited supervision, cosine-based scoring often leads to overly smooth similarity distributions and weak angular decision boundaries, resulting in ambiguous predictions when distinguishing fine-grained novel categories. In this work, we explore a new prototype similarity network (ProtoSimNet) for FSL that emphasizes sharpening the angular geometry of the embedding space under data scarcity constraints. ProtoSimNet introduces a cosine sharpening transformation, combined with an elastic regularization, which together enhance discriminability by amplifying inter-class angular differences while stabilizing intra-class embedding consistency. Our method strengthens the reliability of cosine-based classification and improves the robustness of prototype matching when only a few labeled samples are available. Extensive experiments demonstrate our method outperforms other competitors in generalizing evaluations.

This chapter is based on the accepted paper:

- **He, K.**, Pu, N., Bakker, E. M., and Lew, M. S. “ProtoSimNet: Exploiting Prototype Similarity for Few-shot Learning.” accepted by the Neurocomputing journal, 2026.

3.1 Introduction

Over recent years, with the rapid development of big data, remarkable successes have been achieved in deep learning area for various computer vision tasks, such as image classification [229, 230], object detection [231, 232] and semantic segmentation [233, 234]. However, these deep learning models usually require a large number of labeled samples during the training process, which severely restricts its application in scenarios with limited data. Therefore, training a good model with only few labeled samples, named few-shot learning (FSL) [130, 235, 236], has received growing interest and has become an important topic in deep learning.

In general, these FSL methods can be roughly categorized into three types, metric-based [141], model-based [169, 176] and optimized-based methods [2, 115, 237]. Metric-based methods focus on finding a metric space, where query samples can be simply matched with the support set. Compared to other kinds of FSL methods, metric-based methods do not require to learn additional parameters once the metric is learned. Due to the promising performances, metric-based methods are taken as a breakthrough point and a series of studies have been proposed. In order to calculate the similarity between support and query samples, Sung *et al.* [24] introduced RelationNet which utilizes network to learn the relation scores between the support and query samples. Based on a memory module, cosine distance is applied as a metric in MatchingNet [14] to classify query samples. In order to achieve better classification performances, Hou *et al.* designed CAN [22] to highlight the proper regions of interest through an attention module. Besides, Zhang *et al.* [37] also proposed an algorithm named DeepEMD, which applies the Earth Mover’s Distance (EMD) as a similarity metric.

As a typical metric-based method, Prototypical Network achieves promising performance in few-shot classification. However, Euclidean prototypes can be sensitive to the distribution of support samples. When class prototypes are estimated from imbalanced or slightly distribution-shifted support sets, small perturbations in the support distribution may significantly shift the class prototypes in the embedding space and thus change the decision boundary, which potentially degrades generalization performance on new tasks. This motivates us to replace the Euclidean distance with another metric which can preserve stable decision boundaries.

In this paper, we propose a novel prototype similarity network (ProtoSimNet) for FSL tasks. ProtoSimNet consists of two complementary modules: a cosine sharpening transformation module and an elastic regularization module. 1) Cosine sharpening transformation module. We first introduce a prototype-level cosine sharpening transformation to refine similarity estimation. Specifically, we apply a non-linear sharpening function to the cosine similarities between query features and class prototypes, which preserves the original angular relationships while amplifying the gap

between high and low similarity scores. This operation effectively enlarges similarity margins and yields more distinct decision boundaries in the embedding space, thereby mitigating the sensitivity of prototype-based classifiers to imbalanced or mildly shifted support distributions. 2) Elastic regularization module. Although cosine sharpening transformation already brings noticeable improvements, we observe that it is still insufficient to fully address issues such as overfitting to base classes, label noise, and class imbalance. To further enhance robustness, we design an elastic regularization mechanism that explicitly shapes the feature space by encouraging compact within-class clusters and preventing excessive prototype drift across episodes. This regularization term is lightweight and can be seamlessly integrated into the episodic training objective. By jointly optimizing these two modules within ProtoSimNet, we obtain more robust feature representations and more stable decision boundaries, leading to substantial performance gains on both base and novel classes. The whole framework is built on a standard Conv4 backbone network and trained in an end-to-end episodic strategy. The comparison of different metrics is presented in Figure 3.1 and the overall pipeline diagram is shown in Figure 3.2.

The main contributions of our work can be concluded as follows: 1) We identify and systematically analyze a previously overlooked metric over-smoothing phenomenon in prototypical few-shot learning, where conventional distance metrics produce overly similar prototype scores and thus weak decision margins. To explicitly address this issue, we design a new metric-based framework, ProtoSimNet, specifically designed to alleviate metric over-smoothing and enhance the separability of prototypes. 2) We introduce a non-linear cosine sharpening transformation for prototype similarity, which induces an implicit margin without altering the classifier architecture. The transformation delivers stronger discriminability and improves model sensitivity to subtle inter-class differences. 3) We incorporate an elastic regularization that adaptively balances ℓ_1 sparsity constraints and ℓ_2 weight decay, effectively mitigating overfitting in data-scarce scenarios. This regularization strategy demonstrates particular efficacy in enhancing model robustness against class imbalance and label noise. 4) Extensive experiments on standardized FSL benchmarks demonstrate that our method achieves superior performance across various task configurations, including both 5-way 1-shot and 5-way 5-shot scenarios. The consistent improvements over conventional cosine and Euclidean distance metrics validate the effectiveness of our proposed approach.

3.2 Related Work

FSL Methods: Metric-based methods have witnessed great success due to their strong representation ability. These methods employ a learning-to-compare paradigm, matching features between support and query samples through a suitable metric function. As one of the most representative metric-based methods, Prototypical

3. PROTOSIMNET: EXPLOITING PROTOTYPE SIMILARITY FOR FEW-SHOT LEARNING

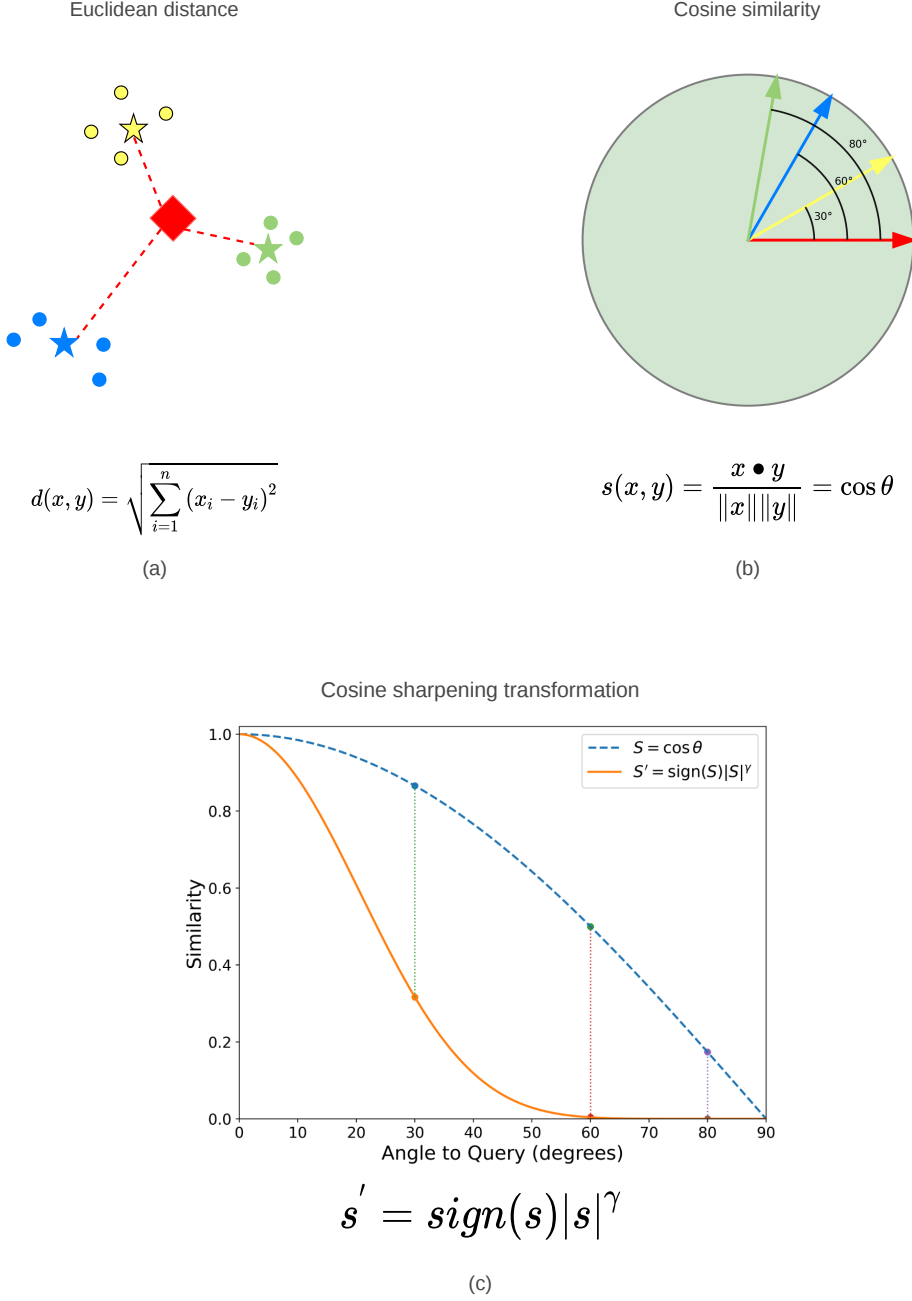


Figure 3.1: The comparison of different metrics. (a) Euclidean distance; (b) Cosine similarity; (c) Cosine sharpening transformation.

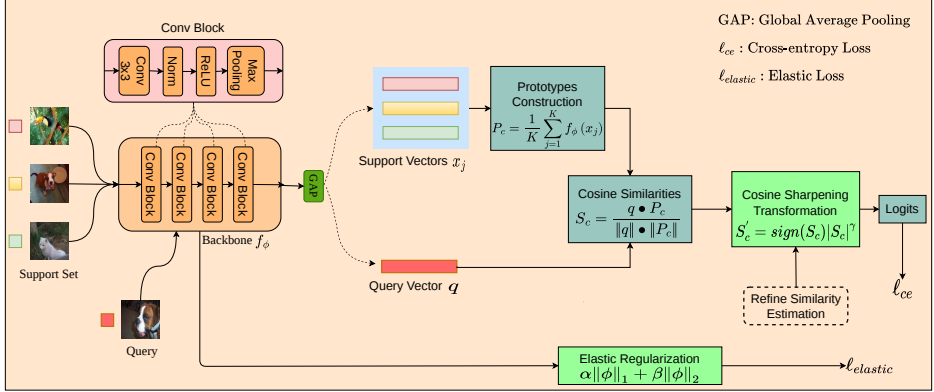


Figure 3.2: The overall pipeline diagram of our framework. The class prototypes, cosine similarities, cosine sharpening transformation, and elastic regularization are defined in Eq. 3.1, 3.2, 3.3, and 3.7, respectively. Their detailed implementations are illustrated in Figure 3.3

Network [1] adopts the mean of intra-class examples as the class prototype, classifying the query sample according to the Euclidean Distance between the query and class prototypes. Model-based methods encourage networks to quickly update the parameters through the design of the model structure. For example, Santoro *et al.* [169] explored a memory-augmented neural network (MANN) which uses an explicit storage buffer to learn unseen classes via an LSTM-based meta-learner. For rapid learning and generalization, Munkhdalai and Yu [176] introduced an advanced weight update scheme by generating fast and slow weights and incorporated them into a neural network. Optimized-based methods try to learn a model with a good initialization from base classes and then adapt the learned model to novel classes through few optimization steps. Model-Agnostic Meta learning (MAML) [2] aims to learn a good parameter initialization by involving the second-order gradients, enabling the model to rapidly adapt to new tasks with a few gradient steps. In order to obtain better generalization for novel classes, Lee *et al.* [237] proposed to utilize properties of linear classifiers to learn feature representations.

Prototypical Network: As one of the most representative methods among metric-based learning, Prototypical Network denotes the mean embeddings of each support class as class prototypes, and the evaluation rule is based on the Euclidean Distance to the nearest class prototype. Despite the outstanding performance of Prototypical Network, several studies have argued that the learned class prototypes may be biased and not efficient when given only few labeled examples. To alleviate the influence caused by biased data distribution, Ji *et al.* [238] proposed IPN, which uses a softmax function to generate weights for the intra-class feature vectors. A metric scaling method [23] is also adopted into the objective function to obtain a task-dependent metric space. Ye *et al.* [239] hold the view that backbone networks may accidentally select uncharacteristic features since classification tasks are agnostic to the current

model. Hence, they utilized a self-attention mechanism to the mean prototypes to yield task-specific prototypes to address the issue. Huang *et al.* [240] believe that a single prototype is insufficient to represent the characteristics of the class. Based on this point, they proposed to learn multiple prototypes of each class via a learnable squeeze and spatial excitation module (sSE) [241].

3.3 Proposed Method

To address the limitations of common similarity metrics in FSL, we propose a novel framework that integrates cosine sharpening transformation with elastic regularization. Our approach employs a unified episodic training strategy with a standard Conv4 backbone, eliminating the need for complex multi-stage optimization while enhancing the discriminative ability of prototype-based classification.

3.3.1 Problem Formulation

In the FSL paradigm, we take a base dataset as $D^{base} = \{x_i^{base}, y_i^{base}\}_{i=1}^{N^{base}}$ with labels $y_i^{base} \in C^{base}$, where both the number of classes $|C^{base}|$ and instances N^{base} are sufficiently large to learn transferable feature representations. For evaluation, we utilize a novel dataset $D^{novel} = (x_j^{novel}, y_j^{novel})_{j=1}^{N^{novel}}$ with $y_j^{novel} \in C^{novel}$, where $C^{base} \cap C^{novel} = \emptyset$ ensures disjoint class spaces between training and testing phases.

Both the training and testing process follow an episodic scheme, where each episode E_i consists of a support set $S_i = (x_i^{t_{support}}, y_i^{t_{support}})_{t=1}^{C \times K}$ and a query set $Q_i = (x_i^{t_{query}}, y_i^{t_{query}})_{t=1}^{C \times K'}$, containing C classes with K examples per class from D^{base} and K' examples per class from D^{novel} , respectively.

3.3.2 Cosine Sharpening Transformation

To enhance the metric’s sensitivity, we introduce a novel cosine sharpening transformation that nonlinearly amplifies the distinction between strong and weak similarities.

Given K support samples from the c -th class, based on the backbone f_ϕ , the c -th class prototype P_c can be calculated as:

$$P_c = \frac{1}{K} \sum_{j=1}^K f_\phi(x_j), \quad (3.1)$$

For any query embedding q , the traditional cosine similarity S_c between the query sample q and the class prototype P_c is computed as:

$$S_c = \frac{q \bullet P_c}{\|q\| \bullet \|P_c\|}, \quad (3.2)$$

where $\|\cdot\|$ is the Euclidean norm and $\bullet$ represents the standard dot product. The cosine similarity measures the angular alignment between the query representation and each class prototype, while reducing the influence of feature magnitude.

The cosine sharpening transformation exploits a signed power operation to reshape the cosine similarity distribution:

$$S'_c = \text{sign}(S_c) |S_c|^\gamma, \quad (3.3)$$

where $\gamma > 1$ is a sharpening factor. This transformation performs a non-linear contrast adjustment on the original cosine similarity as it nonlinearly enhances the contrast between highly relevant and marginally relevant pairs. Specifically, the sign function is introduced to preserve the sign of the original cosine similarity and γ controls the strength of the non-linear calibration. The sharpened similarities are subsequently calibrated through temperature scaling τ to form the final logits for classification:

$$\text{logits}_c = \tau S'_c, \quad (3.4)$$

where τ is a parameter that controls the temperature scaling of the distribution sharpness, and class posteriors follow can be obtained via the following softmax function:

$$p(y = c | x) = \frac{\exp(\text{logits}_c)}{\sum_{j=1}^N \exp(\text{logits}_j)}, \quad (3.5)$$

The query loss in an episode is defined as:

$$\ell_{ce} = -\frac{1}{|Q_i|} \sum_{(x,y) \in Q_i} \log p(y | x). \quad (3.6)$$

The whole workflow of cosine sharpening transformation is presented in Figure 3.3a. This formulation allows the module to emphasize discriminative similarity patterns without introducing excessive computational source, making it particularly suitable for data-scarce FSL scenarios.

3.3.3 Elastic Regularization

To alleviate the overfitting problem in data-scarce scenarios and to obtain robust feature representations, we further introduce an elastic regularization term into our framework. In FSL, the backbone is trained on a limited set of base classes. This often leads to unstable prototypes and poor generalization to novel classes when there exists imbalanced distribution in the base set. Elastic regularization aims to mitigate this problem by explicitly constraining the parameter space during episodic training.

Specifically, elastic regularization combines the complementary effects of ℓ_2 smoothness and ℓ_1 sparsity. The ℓ_1 component enables many parameters to become exactly or approximately zero, effectively performing feature selection by suppressing

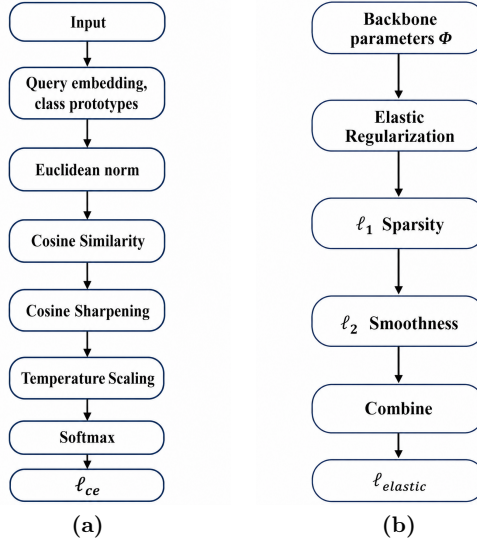


Figure 3.3: Module-level visualization. (a) Cosine sharpening module. (b) Elastic regularization.

uninformative or redundant feature dimensions in the embedding space. The ℓ_2 component, on the other hand, penalizes excessively large weights and promotes smoothness of the learned mapping, which stabilizes optimization and prevents the model from overfitting to noise or outliers in the support set. For the backbone parameters ϕ , the regularizer is:

$$\ell_{elastic} = \alpha \|\phi\|_1 + \beta \|\phi\|_2, \quad (3.7)$$

where $\|\phi\|_1 = \sum_{\phi_i \in \phi} |\phi_i|$, $\|\phi\|_2 = \left(\sum_{\phi_i \in \phi} \phi_i^2 \right)^{1/2}$. α and β are hyperparameters controlling the relative contributions of the ℓ_1 and ℓ_2 regularization terms. The workflow of elastic regularization is presented in Figure 3.3b.

Within the proposed ProtoSimNet framework, the elastic regularization term is optimized jointly with the standard episodic classification loss. This joint optimization gradually shapes the embedding space into one with tighter intra-class clusters and clearer inter-class separation, while also reducing redundancy across feature dimensions. As a result, the learned representations exhibit stronger robustness to label noise, class imbalance, and base/novel distribution mismatch, and the resulting prototypes become more stable across episodes and more discriminative on novel few-shot tasks. The overall objective of ProtoSimNet can be written as:

$$\ell = \ell_{ce} + \ell_{elastic}, \quad (3.8)$$

which enables tighter intra-class clusters and reduces redundancy in the embedding

space.

3.4 Experiments

In this section, we elaborate on the experimental configuration and evaluation. The description of the datasets and implementation details are first presented, followed by experimental results of our approach on popular benchmark datasets. Finally, an ablative analysis is presented.

3.4.1 Datasets and Implementation details

Datasets: We evaluate our approach on three widely used few-shot learning benchmarks: mini-ImageNet [14], tieredImageNet [62] and CUB-200 [113]. The mini-ImageNet dataset is a subset of ImageNet, consisting of 100 categories, each with 600 images. We adopt the data splitting protocol proposed in [115], where 64 classes are allocated for training, 16 classes for validation, and the remaining 20 classes for testing. The tieredImageNet dataset comprises 608 classes, which are further organized into 34 high-level categories. We use 20 categories (351 classes) for training, 6 categories (97 classes) for validation, and 8 categories (160 classes) for testing. The CUB dataset was originally introduced for fine-grained bird classification and contains 11,788 images from 200 classes. We use 100 classes for training, 50 classes for validation, and 50 classes for testing.

Implementation details: in this work, we use Conv4 as our backbone, which is composed of four convolutional blocks. Each block comprises a 64-filter 3×3 convolution, batch normalization layer, a ReLU layer and a 2×2 max-pooling layer. We use Adam with an initial learning rate of $1e-3$ and cut the learning rate in half every 2000 episodes. Each episode includes 60 classes, with 5 support samples and 5 query samples. Besides, we use random cropping, color jittering and random horizontal flip for data augmentation during the whole process.

3.4.2 Evaluations

miniImageNet and tieredImageNet. Table 3.1 presents a comparison between our approach and the state-of-the-art methods in FSL tasks on the two ImageNet-based benchmarks. Our method is denoted as ProtoSimNet. On both datasets and in both 1-shot and 5-shot scenarios, our approach yields state-of-the-art results. On miniImageNet, ProtoSimNet achieves 59.48% and 79.75% in 5-way 1-shot and 5-way 5-shot tasks, respectively. ProtoSimNet achieves improvements over the best-reported numbers by 0.52% under 5-way 1-shot and 4.58% under 5-way 5-shot settings. On tieredImageNet, ProtoSimNet achieves 59.71% and 79.40% in the 1-shot and 5-shot tasks, respectively. ProtoSimNet achieves improvements over the best-reported numbers by 1.51% under 5-way 1-shot and 3.02% under 5-way 5-shot settings.

3. PROTO SIMNET: EXPLOITING PROTOTYPE SIMILARITY FOR FEW-SHOT LEARNING

Table 3.1: Comparison of ProtoSimNet (Conv4-64) to prior works on miniImageNet and tieredImageNet datasets, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	miniImageNet 5-way 1-shot	miniImageNet 5-way 5-shot	tieredImageNet 5-way 1-shot	tieredImageNet 5-way 5-shot
ProtoNet [1]	Conv4-64	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	53.31 $\pm$ 0.89	72.69 $\pm$ 0.74
MatchingNet [14]	Conv4-64	43.56 $\pm$ 0.84	55.31 $\pm$ 0.73	–	–
RelationNet [24]	Conv4-64	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70	54.48 $\pm$ 0.93	71.31 $\pm$ 0.78
CovaMNet [151]	Conv4-64	51.19 $\pm$ 0.76	67.65 $\pm$ 0.63	54.98 $\pm$ 0.90	71.51 $\pm$ 0.75
DN4 [116]	Conv4-64	51.24 $\pm$ 0.74	71.02 $\pm$ 0.64	52.89 $\pm$ 0.23	73.36 $\pm$ 0.73
MetaOptNet [237]	Conv4-64	52.87 $\pm$ 0.57	68.76 $\pm$ 0.48	54.71 $\pm$ 0.67	71.79 $\pm$ 0.59
DeepEMD [37]	Conv4-64	51.72 $\pm$ 0.20	65.10 $\pm$ 0.39	51.22 $\pm$ 0.14	65.81 $\pm$ 0.68
RFS-Simple [110]	Conv4-64	55.25 $\pm$ 0.58	71.56 $\pm$ 0.52	56.18 $\pm$ 0.70	72.99 $\pm$ 0.55
RFS-Distill [110]	Conv4-64	55.88 $\pm$ 0.59	71.65 $\pm$ 0.51	56.76 $\pm$ 0.68	73.21 $\pm$ 0.54
FRN [122]	Conv4-64	54.87	71.56	55.54	74.68
DMN4 [242]	Conv4-64	55.77	74.22	56.99	74.13
ATL-Net [243]	Conv4-64	54.30 $\pm$ 0.76	73.22 $\pm$ 0.63	–	–
TALDS-Net [244]	Conv4-64	56.78 $\pm$ 0.37	74.63 $\pm$ 0.13	57.54 $\pm$ 0.71	75.79 $\pm$ 0.43
MCL [245]	Conv4-64	57.88	74.03	58.20	74.77
HTS [246]	Conv4-64	58.96 $\pm$ 0.18	75.17 $\pm$ 0.14	53.20 $\pm$ 0.22	72.38 $\pm$ 0.19
TADNet [247]	Conv4-64	56.14 $\pm$ 0.20	74.68 $\pm$ 0.15	57.88 $\pm$ 0.21	75.98 $\pm$ 0.17
DCRNet [248]	Conv4-64	56.35 $\pm$ 0.31	74.93 $\pm$ 0.17	58.04 $\pm$ 0.21	76.38 $\pm$ 0.23
ProtoSimNet	Conv4-64	59.48 $\pm$ 0.63	79.75 $\pm$ 0.42	59.71 $\pm$ 0.74	79.40 $\pm$ 0.47

CUB-200 dataset. As shown in Table 3.2, our method also achieves state-of-the-art performance on CUB-200 dataset. In the 5-way 1-shot and 5-shot settings, when using Conv4-64 as the backbone, ProtoSimNet improves by 0.33% and 0.42%, respectively, compared to the state-of-the-art method based on HTS [246]. To evaluate the cross-domain generalization of ProtoSimNet, we conducted experiments under the miniImageNet→CUB setting (see Table 3.3) and compared it with state-of-the-art methods. The model is trained on 64 base classes of miniImageNet and evaluated on 50 novel classes in the CUB-200 test set. Our ProtoSimNet demonstrates significant advantages in this cross-domain scenario. Using the ResNet12 backbone, ProtoSimNet maintains a lead in both 5-way 1-shot and 5-way 5-shot settings compared to other methods.

3.4.3 Ablative Analysis

Backbone Dependence. To address backbone dependence, we additionally evaluate our method on a stronger backbone, ResNet12. These new results are provided in Table 3.4. Experimental results indicate that the proposed ProtoSimNet is not tied to a specific backbone.

Benefits of Cosine Sharpening Transformation. To study the impact of cosine sharpening transformation, we evaluate our approach with ProtoNet and DeepEMD on miniImageNet. As shown in Table 3.1, compared to ProtoNet which uses Euclidean distance as the metric, ProtoSimNet obtains 10.06% and 11.55% improvements on 5-way 1-shot and 5-way 5-shot tasks, respectively. Compared to DeepEMD, ProtoSimNet also gains 7.76% and 14.65% improvements on 5-way 1-shot and 5-way 5-shot tasks, respectively. Furthermore, we make a comparison

Table 3.2: Comparison of ProtoSimNet (Conv4-64) to prior works on CUB-200 dataset, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	CUB-200 5-way	
		1-shot	5-shot
ProtoNet [1]	Conv4-64	37.36 $\pm$ 1.00	45.28 $\pm$ 1.03
MatchingNet [14]	Conv4-64	45.30 $\pm$ 1.03	59.50 $\pm$ 1.01
CSS [249]	Conv4-64	66.01 $\pm$ 0.90	81.84 $\pm$ 0.59
DN4 [116]	Conv4-64	46.84 $\pm$ 0.81	74.92 $\pm$ 0.62
MADN4 [250]	Conv4-64	57.11 $\pm$ 0.70	77.83 $\pm$ 0.40
HTS [246]	Conv4-64	67.32 $\pm$ 0.24	78.09 $\pm$ 0.15
BDLA [251]	Conv4-64	50.59 $\pm$ 0.97	75.36 $\pm$ 0.72
DLDA [252]	Conv4-64	55.12 $\pm$ 0.86	74.46 $\pm$ 0.65
KLSANet [253]	Conv4-64	66.70 $\pm$ 0.82	83.63 $\pm$ 0.28
FAFD-LDWR [254]	Conv4-64	65.45 $\pm$ 0.67	79.63 $\pm$ 0.33
ProtoSimNet	Conv4-64	67.65 $\pm$ 0.84	84.05 $\pm$ 0.49

Table 3.3: Cross-domain performance comparison of the proposed ProtoSimNet with state-of-the-art methods under miniImageNet $\rightarrow$ CUB-200 setting, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	miniImageNet $\rightarrow$ CUB-200	
		5-way 1-shot	5-way 5-shot
MIFN [255]	ResNet12	48.21 $\pm$ 0.60	65.33 $\pm$ 0.54
KLSANet [253]	ResNet12	48.16 $\pm$ 0.64	67.25 $\pm$ 0.61
FAFD-LDWR [254]	ResNet12	48.64 $\pm$ 0.42	67.36 $\pm$ 0.55
ProtoSimNet	ResNet12	49.38 $\pm$ 0.68	67.83 $\pm$ 0.53

with other cosine-metric methods. The results are presented in Table 3.5. Table 3.5 shows that our ProtoSimNet obtains the best performance on both 5-way 1-shot and 5-way 5-shot tasks on miniImageNet. From the results, we can infer the effectiveness of cosine sharpening transformation during the training process.

Benefits of ElasticNet Loss. To better evaluate the contribution of ElasticNet Loss, we train the model with different loss on miniImageNet. From Table 3.6, we can find that, for the model trained without ElasticNet Loss, it achieves 58.56% and 78.28% in 1-shot and 5-shot tasks, respectively. For the model trained with ℓ_1 or ℓ_2 regularization only, the experimental results demonstrate consistent improvements of varying magnitudes. In particular, the model trained with ElasticNet loss yields the most significant improvement over the prior settings. These results confirm the importance of ElasticNet Loss.

Module contribution. To investigate the contribution of each module, we evaluate our method with different modules. Figure 3.4 shows embedding visualization (t-SNE) on base and novel classes with different models. We can find that samples

3. PROTOSIMNET: EXPLOITING PROTOTYPE SIMILARITY FOR FEW-SHOT LEARNING

Table 3.4: Comparison of ProtoSimNet (ResNet12) to prior works on 5-way 5-shot tasks, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	miniImageNet	tieredImageNet
ProtoNet [1]	ResNet12	80.88 $\pm$ 0.41	84.06 $\pm$ 0.26
MatchingNet [14]	ResNet12	75.99 $\pm$ 0.15	80.60 $\pm$ 0.71
DN4 [116]	ResNet12	81.10	83.41
FEAT [239]	ResNet12	82.05	84.79
BML [256]	ResNet12	83.63 $\pm$ 0.29	85.49 $\pm$ 0.34
FRN [122]	ResNet12	82.83 $\pm$ 0.13	86.89 $\pm$ 0.14
TPMN [257]	ResNet12	83.44 $\pm$ 0.43	86.55 $\pm$ 0.63
DeepEMD [37]	ResNet12	82.41 $\pm$ 0.56	86.03 $\pm$ 0.58
MCL [245]	ResNet12	83.99	86.02
DMN4 [242]	ResNet12	83.52	85.72
DeepBDC [160]	ResNet12	84.46 $\pm$ 0.28	87.31 $\pm$ 0.22
TADNet [247]	ResNet12	84.23 $\pm$ 0.67	86.46 $\pm$ 0.15
Meta-OLE [258]	ResNet12	82.23 $\pm$ 0.67	85.51 $\pm$ 0.59
CORL [259]	ResNet12	83.03 $\pm$ 0.33	86.76 $\pm$ 0.52
LEADNET [260]	ResNet12	83.21 $\pm$ 0.14	86.60 $\pm$ 0.16
ProtoSimNet	ResNet12	84.70 $\pm$ 0.35	88.40 $\pm$ 0.37

Table 3.5: Comparison of ProtoSimNet (Conv4-64) with other cosine-metric methods.

Model	Metric	miniImageNet 1-shot	5-way 5-shot
DN4 [116]	cosine similarity	51.24 $\pm$ 0.74	71.02 $\pm$ 0.64
DMN4 [242]	cosine similarity	55.77	74.22
TCDSNet [261]	cosine similarity	57.14 $\pm$ 0.22	75.89 $\pm$ 0.35
LDCA [262]	cosine similarity	53.46 $\pm$ 0.63	75.06 $\pm$ 0.48
TALDS-Net [244]	cosine similarity	56.78 $\pm$ 0.37	74.63 $\pm$ 0.13
ProtoSimNet	cosine sharpening transformation	59.48 $\pm$ 0.63	79.75 $\pm$ 0.42

Table 3.6: ProtoSimNet (Conv4-64) performances on miniImageNet, with different regularization terms.

Loss Function	miniImageNet 1-shot	5-way 5-shot
ProtoSimNet (without Elastic Loss)	58.56 $\pm$ 0.64	78.28 $\pm$ 0.43
ProtoSimNet (with ℓ_1 regularization only)	58.68 $\pm$ 0.61	78.53 $\pm$ 0.41
ProtoSimNet (with ℓ_2 regularization only)	59.14 $\pm$ 0.61	78.77 $\pm$ 0.42
ProtoSimNet (with Elastic Loss)	59.48 $\pm$ 0.63	79.75 $\pm$ 0.42

from same classes tend to form tighter, more compact clusters, while different classes become more clearly separated after adding the cosine sharpening transformation and elastic regularization modules. We use the model without any module as the baseline, which uses a Conv4 as the backbone. To obtain better performance, the

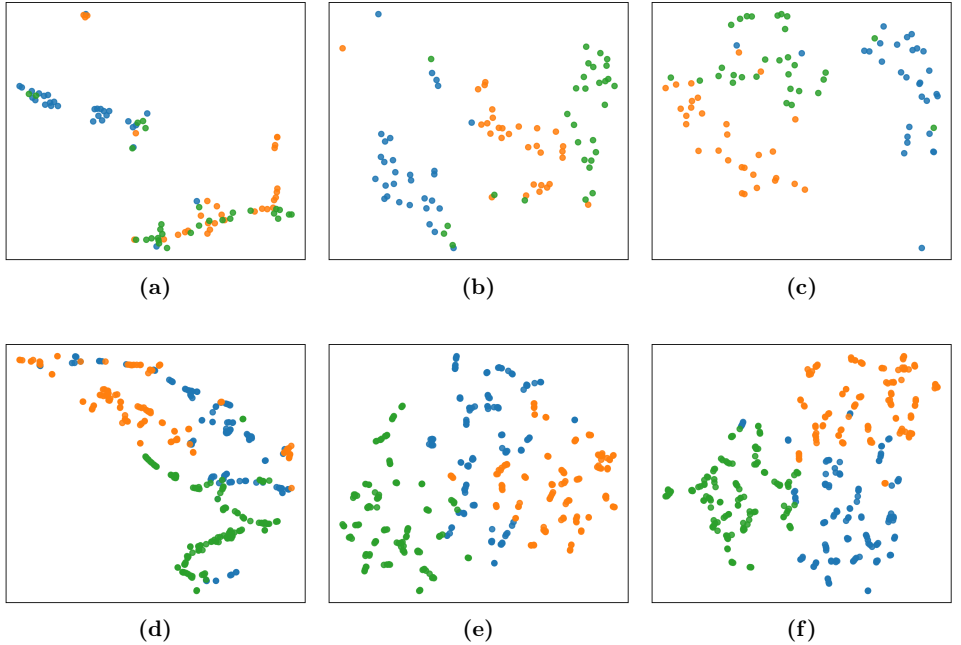


Figure 3.4: Qualitative embedding visualization (t-SNE) on base and novel classes. (a) Base Classes (Cosine Similarity). (b) Base Classes (Cosine Sharpening Transformation). (c) Base Classes (Cosine Sharpening Transformation + Elastic Regularization). (d) Novel Classes (Cosine Similarity). (e) Novel Classes (Cosine Sharpening Transformation). (f) Novel Classes (Cosine Sharpening Transformation + Elastic Regularization).

Table 3.7: ProtoSimNet (Conv4-64) performances on miniImageNet, with different module.

Model	miniImageNet 5-way	
	1-shot	5-shot
baseline	57.98 ± 0.65	77.56 ± 0.44
baseline + cosine sharpening transformation	58.56 ± 0.64	78.28 ± 0.43
baseline + cosine sharpening transformation + elastic regularization	59.48 ± 0.63	79.75 ± 0.42

baseline simulates FSL setting and takes 5 samples per class as the support set and another 5 samples as the query set in each episode during the training process. The cosine similarity is used as the metric for evaluation. From Table 3.7, we find that the baseline achieves 57.98% and 77.56% accuracies on miniImageNet 5-way 1-shot and 5-shot tasks, respectively. By adding the cosine sharpening transformation module, the model obtains 0.58% and 0.72% improvements. The elastic regularization helps the model further get 0.92% and 1.47% gains.

Variations of Hyper-parameters. There are totally four hyper-parameters in

3. PROTOSIMNET: EXPLOITING PROTOTYPE SIMILARITY FOR FEW-SHOT LEARNING

Table 3.8: ProtoSimNet (Conv4-64) performances on miniImageNet, with different γ values.

γ	miniImageNet 5-way	
	1-shot	5-shot
1.2	59.12 ± 0.63	78.49 ± 0.43
1.5	59.48 ± 0.63	79.75 ± 0.42
2	59.43 ± 0.64	79.47 ± 0.42
5	58.45 ± 0.64	79.27 ± 0.43
10	57.63 ± 0.63	79.05 ± 0.43

Table 3.9: ProtoSimNet (Conv4-64) performances on miniImageNet, with different α values.

α	miniImageNet 5-way	
	1-shot	5-shot
0.0001	58.15 ± 0.64	79.21 ± 0.41
0.0005	59.48 ± 0.63	79.75 ± 0.42
0.001	58.90 ± 0.66	78.62 ± 0.42
0.005	58.27 ± 0.62	75.27 ± 0.51
0.01	58.80 ± 0.67	71.46 ± 0.52
0.05	57.86 ± 0.64	59.49 ± 0.62

this work, τ , α , β and γ . τ is a temperature coefficient, γ is a sharpening factor, α and β are the weights of ℓ_1 and ℓ_2 regularization terms, respectively. We investigate the variants of hyper-parameters α , β and γ , where the default values of τ is set to 25.0. Table 3.8 shows the ProtoSimNet performance on miniImageNet by changing γ . It’s obviously that our model obtains over 59% accuracies on 5-way 1-shot tasks when γ range from 1.2 to 2, and then the performance decreases when γ increases from 2 till 10, which indicates the importance of sharpen scale. Table 3.9 shows the ProtoSimNet performance on miniImageNet by changing α . We observe that ProtoSimNet achieves over 58% accuracies on 5-way 1-shot tasks when α increases from 0.0001 till 0.01, while $\alpha = 0.05$ performs the lowest. Table 3.10 presents the ProtoSimNet performance on miniImageNet by changing β . When β ranges from 0.0001 till 0.001, ProtoSimNet can always obtains over 79% accuracies on 5-way 5-shot tasks and achieves the highest accuracy when $\beta = 0.0005$, while $\beta = 0.05$ performs the lowest, indicating a large value of β is not beneficial for the model. We also study ProtoSimNet performance with different values of τ . As shown in Table 3.11, when τ ranges from 10 to 30, ProtoSimNet always obtains over 58% accuracies on miniImageNet 5-way 1-shot tasks and over 76% accuracies on miniImageNet 5-way 5-shot tasks, respectively.

Table 3.10: ProtoSimNet (Conv4-64) performances on miniImageNet, with different β values.

β	miniImageNet 5-way	
	1-shot	5-shot
0.0001	58.94 ± 0.66	79.36 ± 0.43
0.0005	59.48 ± 0.63	79.75 ± 0.42
0.001	59.27 ± 0.65	79.27 ± 0.43
0.005	58.59 ± 0.62	77.77 ± 0.42
0.01	57.90 ± 0.63	77.34 ± 0.43
0.05	57.77 ± 0.62	76.86 ± 0.44

Table 3.11: ProtoSimNet (Conv4-64) performances on miniImageNet, with different τ values.

τ	miniImageNet 5-way	
	1-shot	5-shot
5	56.08 ± 0.63	72.02 ± 0.47
10	58.39 ± 0.62	76.58 ± 0.44
15	58.90 ± 0.63	79.18 ± 0.42
20	59.01 ± 0.64	79.13 ± 0.41
25	59.48 ± 0.63	79.75 ± 0.42
30	58.74 ± 0.65	79.20 ± 0.43

3.5 Chapter Conclusions

In this chapter, we have presented a comprehensive investigation into the limitations of traditional similarity metrics in few-shot learning scenarios, with particular focus on their sensitivity to data distribution shifts and limited discriminative ability when dealing with complex class relationships. In prototype-based few-shot learning, each class prototype acts like a “center of mass” computed from a few support examples, and classification is determined by which prototype is closest. If the support set is imbalanced or slightly shifted, that “center of mass” can move a lot, even when the underlying class has not changed—causing the separating boundary between classes to drift. Correspondingly, we introduce a novel prototype similarity metric framework that integrates two key innovations: a cosine sharpening transformation and an elastic regularization mechanism. The cosine sharpening transformation emphasizes angular similarity between features and encourages clearer separation in the similarity space, while elastic regularization further limits how freely prototypes can shift in response to small support perturbations. In combination, these contributions make the decision boundary less prone to drifting, improving generalization to new tasks.

Future work. We first plan to investigate adaptive mechanisms for automatically

3. PROTOSIMNET: EXPLOITING PROTOTYPE SIMILARITY FOR FEW-SHOT LEARNING

tuning the sharpening factor γ based on task characteristics and data distributions, and then extend the approach to more challenging cross-domain and incremental FSL scenarios. The proposed method provides a solid foundation for developing more effective and practical few-shot learning systems while maintaining the simplicity that makes prototype-based methods appealing.

Chapter 4

Attention-guided Feature Pyramid Network for few-shot learning

In the previous chapter, we have studied the influences of a good metric in few-shot learning (FSL) tasks. In this chapter, we turn our attention to strengthening the discriminative capability of backbone features by addressing the following question: how to develop a flexible multi-scale representation framework that can dynamically reconcile feature hierarchies while preserving their discriminative capability under limited supervision (**RQ 2**).

Conventional few-shot learning models are typically built on backbone networks that extract features from a single or fixed-resolution representation space. However, in real-world scenarios, models must adapt to diverse visual patterns and significant scale variations across different categories, making it inadequate for these approaches. In this chapter, we introduce an attention-based feature pyramid network (AFPV) to enhance the multi-scale expressiveness of FSL frameworks. AFPV integrates feature maps from multiple backbone layers through attention-driven fusion modules, enabling the model to selectively aggregate discriminative information across spatial scales while suppressing irrelevant or noisy information. This design improves the model’s ability to construct more stable and generalizable representations, even when training data are extremely limited. Experimental results show that: (1) AFPV significantly strengthens classification robustness across few-shot scenarios; (2) the improved representations transfer effectively to unseen categories.

This chapter is based on the following publication [34]:

- **He, K.**, Bakker, E. M., and Lew, M. S. “Attention-guided Feature Pyramid Network for few-shot learning.” Benelux Conference on Artificial Intelligence and Machine Learning, 2024..

4.1 Introduction

Few-shot learning (FSL) [24, 101, 227] has emerged as a significant challenge in machine learning area where models must rapidly adapt to new tasks with only limited labeled examples. In contrast to conventional deep learning paradigms which rely on large-scale annotated datasets, FSL models aim to efficiently extract and transfer generalizable representations from limited samples. This setting makes FSL particularly valuable in many real-world applications, such as medical diagnosis [263, 264], autonomous driving [265, 266], and remote sensing [267, 268], where data acquisition is time-consuming or infeasible. Over the past decades, various research efforts have been made to improve the representation learning ability under the FSL setting. Metric-based approaches such as Matching Networks [14] and Prototypical Networks [1] design distance metrics in the embedding space, while optimization-based frameworks like MAML [2] enable rapid adaptation to new tasks. In addition, auxiliary training strategies are also explored to alleviate overfitting risks and improve generalization ability in FSL. For instance, self-supervision techniques [269, 270] have been incorporated to generate additional pretext tasks from the limited samples, thereby enriching the feature learning process without requiring extra annotations. Similarly, knowledge distillation [110, 271] has been adapted to transfer knowledge from a large, well-trained teacher model to a compact student model, enhancing the latter’s generalization from few examples.

Among these FSL approaches, attention mechanism also attracts much interest in FSL as it enables models to emphasize discriminative regions while suppress irrelevant background features. For instance, Huang and Choi [272] designed SAPENet, which applies multi-head self-attention to recalibrate local feature maps and enhance prototype representations, thereby improving few-shot classification performance. Similarly, Hou *et al.* [22] introduced a cross-attention-based comparison module that constructs attention maps between support and query features to focus on task-relevant object regions while suppressing background clutter, achieving competitive or state-of-the-art results on benchmarks such as miniImageNet and CUB-200-2011. More recently, FEAT [239] proposed a Transformer-style attention module to learn task-specific embeddings in the feature space, and Yang *et al.* [273] further design a meta-tuned transformer which employs self-attention to jointly model relationships among patches, samples, and tasks, bringing significant improvements for few-shot classification on both in-domain and cross-domain benchmarks. However, these methods often treat features at a single scale, ignoring the hierarchical structure of visual representations that are crucial for distinguishing fine-grained or low-contrast categories in few-shot settings.

Building on these insights, we propose an Attention-Guided Feature Pyramid Network (AFPN) for few-shot image classification. AFPN aims to fully exploit hierarchical and spatial information by integrating attention mechanisms into a multi-scale feature learning framework. As presented in Figure 4.1, multi-scale features often

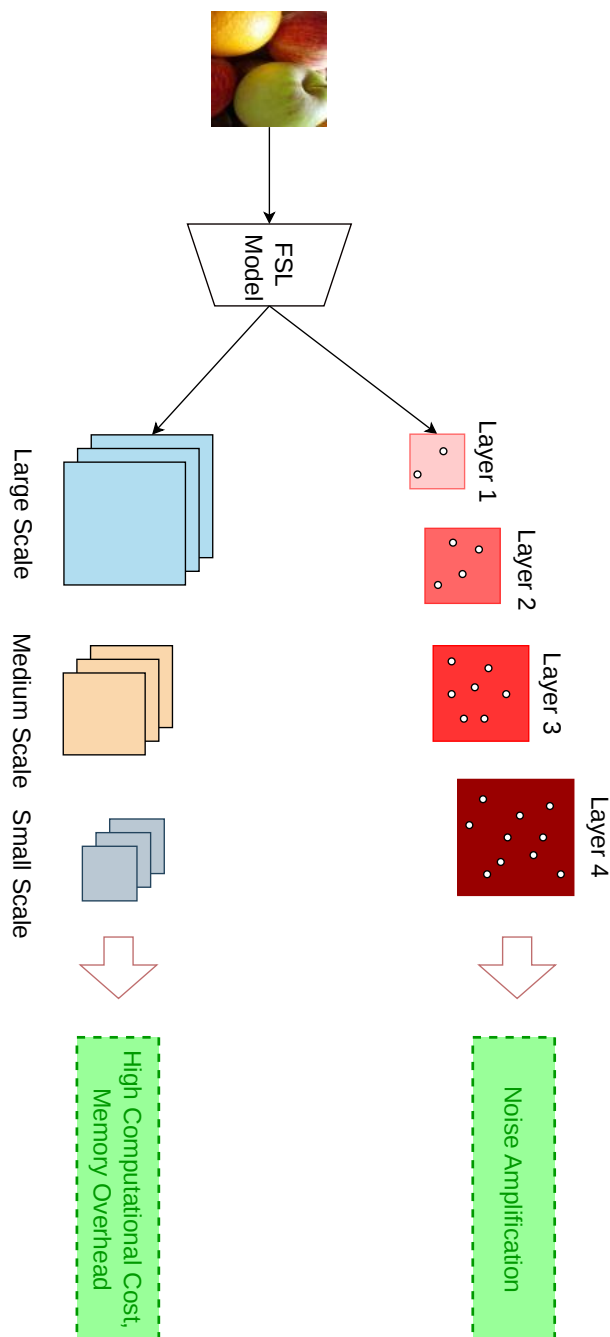


Figure 4.1: Multi-scale features challenge. Large-scale features incur higher computational cost, while deeper layers tend to amplify noise.

4. ATTENTION-GUIDED FEATURE PYRAMID NETWORK FOR FEW-SHOT LEARNING

face high computational cost and noise amplification challenges. Inspired by the success of Feature Pyramid Networks (FPNs) [274] in capturing both fine-grained and high-level semantic cues, AFPN integrates two distinct but complementary attention mechanisms to coordinate effective multi-scale feature interactions. Specifically, we incorporate a parameter-free SimAM attention module [275], which efficiently enhances feature discriminability without introducing additional learnable parameters. It adaptively measures neuron importance based on energy functions, enabling the model to emphasize informative regions while suppressing noisy activations across feature maps. Besides, we design a Cross-layer Spatial Attention Module (CSAM) that explicitly models the interdependencies between feature maps from different pyramid levels. By enabling cross-layer feature modulation, CSAM facilitates the propagation of high-level semantic information to lower-level features and, conversely, allows fine-grained spatial details to refine higher-level representations. The overall framework of AFPN is shown in Figure 4.2. The workflows of SimAM and CSAM are presented in Figure 4.3 and 4.4, respectively.

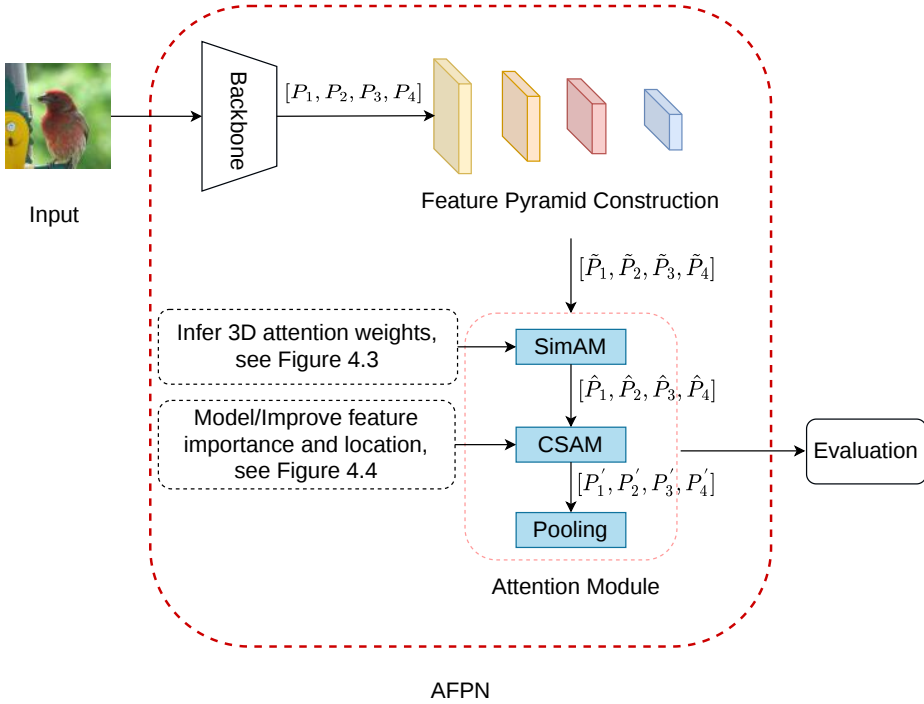


Figure 4.2: The overall framework of AFPN. The detailed implementations of SimAM and CSAM are presented in **Figure 4.3** and **Figure 4.4**

By combining these two attention mechanisms, AFPN achieves effective multi-scale feature interaction while maintaining computational efficiency. The joint operation

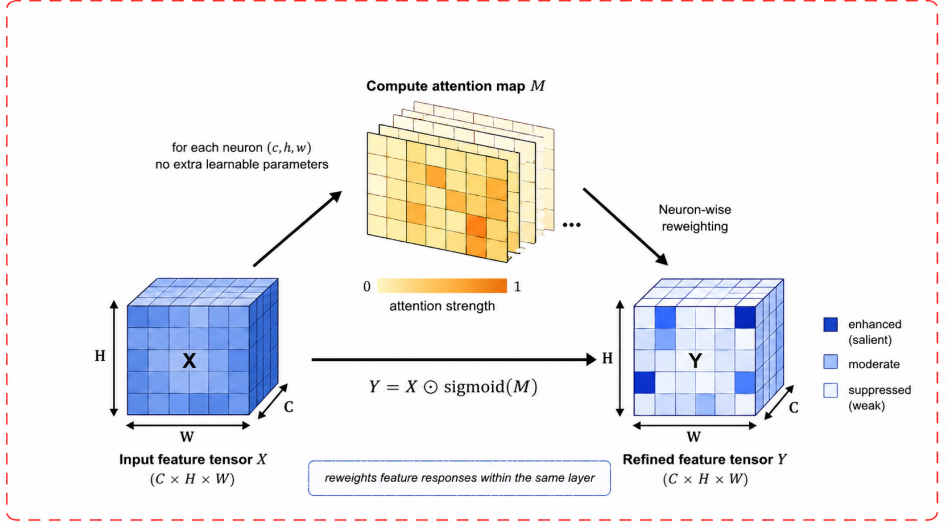


Figure 4.3: The workflow of SimAM

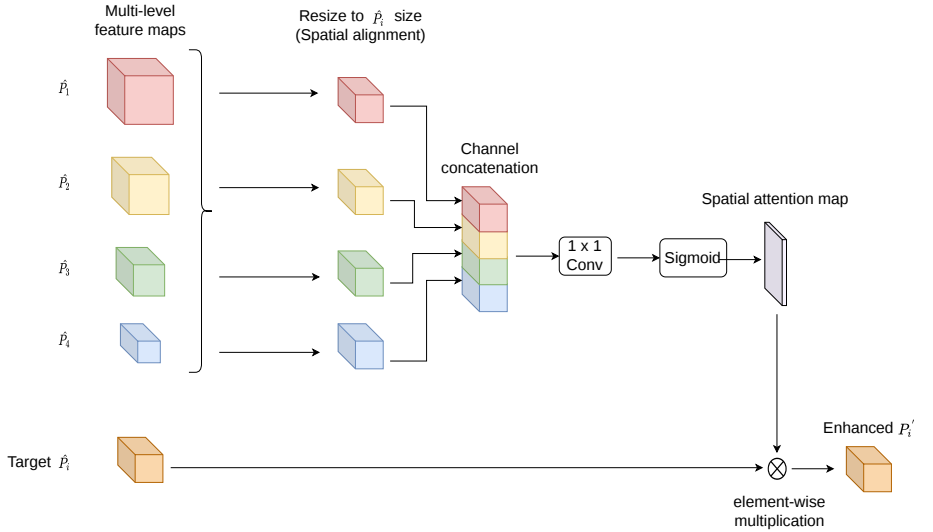


Figure 4.4: The workflow of CSAM

of SimAM and CSAM allows AFPN to retain richer high-level semantic information and simultaneously refine spatial representations crucial for few-shot recognition. Furthermore, this hierarchical attention fusion design enables the model to

generalize better across diverse visual domains, particularly under limited training samples.

In summary, our contributions are:

- We exploit a new feature pyramid architecture inherently captures multi-scale contextual information, which is crucial for recognizing objects from limited examples.
- We propose a novel dual-attention mechanism (SimAM and CSAM) enables dynamic and efficient feature refinement and interaction across scales, ensuring that semantically rich information is preserved and enhanced.
- We design a robust and efficient framework that significantly boosts few-shot classification accuracy by learning more discriminative and generalizable features from limited data, as validated by our extensive experiments.

4.2 Related Work

4.2.1 Few-Shot Learning Paradigm

Few-shot learning (FSL) has emerged as a fundamental paradigm for enabling deep models to recognize new classes with very limited annotated samples. The main challenge in FSL is how to bridge the gap between base classes and novel classes, requiring models to learn transferable representations that generalize to different tasks. Early works such as Matching Network [14] and Relation Network [24] introduced episodic training strategies to mimic the few-shot scenario during meta-learning. Later, optimization-based methods like MAML [2] and its variants [28, 191] focused on rapid adaptation through gradient-based fine-tuning. Recently, researchers have turned to embedding enhancement and regularization strategies to improve the quality of learned representations. For example, self-distillation [269], manifold alignment [117], and meta-transfer learning [198] have been used to preserve feature consistency across tasks. Furthermore, several studies explored leveraging structural priors or hierarchical semantics [276] to enhance the discriminative power of few-shot models. Despite these advances, most existing frameworks rely on single-scale or task-specific embeddings, which limits their ability to capture fine-grained details and contextual information crucial for few-shot generalization.

4.2.2 Attention Mechanisms for Feature Enrichment

Different from traditional convolutional paradigms in computer vision, attention mechanisms enable dynamic feature refinement through content-aware processing, defining a new way for neural networks to capture data dependencies. Vaswani *et al.* [277] proposed a transformer model in natural language processing, where attention

allows a network to assign dynamic importance to different features, enhancing interpretability and adaptability. This achievement was quickly extended to computer vision. Hu *et al.* [241] introduced the Squeeze-and-Excitation (SE) block, which learned to emphasize informative feature channels through global context modeling, while Woo *et al.* [278] proposed the Convolutional Block Attention Module (CBAM), which integrates both spatial and channel attention for refined feature selection. Subsequently, Wang *et al.* [279] developed Non-local Neural Networks to capture long-range dependencies beyond local convolutional receptive fields, inspiring later attention-driven architectures such as Vision Transformers (ViT) [280] and Swin Transformers [281], which demonstrated that pure attention mechanisms can outperform convolution-based models in large-scale vision tasks.

In FSL, attention mechanisms have been widely used to address intra-class variation and inter-class confusion. Hou *et al.* [22] introduced a cross-attention module to align query and support embeddings, while Ye *et al.* [239] extended this idea with self-attention for adaptive prototype learning. Other works, such as Yang *et al.* [275] and Wang *et al.* [282], explored lightweight and efficient attention variants—like SimAM and ECA-Net—that retain discriminative power with minimal computational overhead. Moreover, recent studies have investigated cross-scale and multi-head attention formulations [34, 35] to enhance spatial consistency and semantic richness across hierarchical layers, which is particularly beneficial for FSL tasks. In general, these advances demonstrate that attention mechanisms not only enhance visual discrimination but also provide a flexible way to integrate contextual and hierarchical knowledge, making them an important component for improving few-shot learning performance.

4.3 Proposed Methods

To enhance the model’s generalization ability in FSL scenarios, we propose an Attention-Guided Feature Pyramid Network (AFPV), an effective framework that integrates multi-scale representation learning with dual attention mechanisms. In FSL settings, models often fail to capture discriminative high-level semantics while maintaining fine-grained local features. To mitigate this, AFPV introduces a hierarchical feature fusion module guided by a Simple Parameter-Free Attention Module (SimAM) and a Cross-Layer Spatial Attention Module (CSAM). The proposed architecture balances local and global contextual representations, enabling the model to produce robust embeddings for unseen classes with minimal data.

AFPV mainly consists of five main components: (1) a feature extraction backbone; (2) a feature pyramid network (FPN) for top-down semantic aggregation; (3) a SimAM module that adaptively refines intra-layer neuron importance; (4) a CSAM module that models inter-layer spatial interactions; and (5) a classification head for final prediction.

4.3.1 Multi-Scale Feature Extraction

Given an input image $x \in R^{3 \times H \times W}$ (where H and W denote its height and width, respectively), AFPN employs a stack of convolutional layers to progressively extract hierarchical features from shallow to deep levels. These extracted feature maps are defined as:

$$\begin{aligned} P_1 &= f_1(x), \\ P_2 &= f_2(x), \\ P_3 &= f_3(x), \\ P_4 &= f_4(x), \end{aligned} \tag{4.1}$$

where f_i represents the convolutional transformation at the i -th layer. Each feature map $P_i \in R^{C_i \times H_i \times W_i}$ captures higher semantic information but with decreasing spatial resolution.

4.3.2 Feature Pyramid Construction

Inspired by the Feature Pyramid Network (FPN) [17], we construct a top-down fusion pathway to integrate features with different scales. The FPN module enhances low-level representations by upsampling and merging higher-level semantic information while maintaining computational efficiency:

$$\begin{aligned} \tilde{P}_4 &= P_4, \\ \tilde{P}_i &= P_i + U_p(\tilde{P}_{i+1}), \quad i = 1, 2, 3. \end{aligned} \tag{4.2}$$

where U_p denotes bilinear interpolation. The output of this module is a set of multi-scale features $\{\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\}$, which contain richer contextual information for further attention refinement.

4.3.3 Simple Parameter-Free Attention Module (SimAM)

To refine the importance of individual neurons within each feature map, AFPN employs the SimAM module, which computes neuron-level attention weights based on energy minimization without introducing additional learnable parameters. For any feature map $\tilde{P}_i \in \{\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4\}$, we use $M_i = H_i \times W_i$ to denote the number of neurons on i -th channel, SimAM computes the energy function for each neuron z as follows:

$$e_t = \frac{(t - \tilde{\mu})^2 + 2(\tilde{\sigma})^2 + 2\lambda}{4(\tilde{\sigma}^2 + \lambda)} \tag{4.3}$$

where t represents the target neuron, $\tilde{\mu} = \frac{1}{M_i} \sum_{j=1}^{M_i} z_j$ and $\tilde{\sigma}^2 = \frac{1}{M_i} \sum_{j=1}^{M_i} (z_j - \tilde{\mu})^2$, which denote the mean and variance of all neurons in the same channel excluding t , and λ is a regularization hyperparameter. The attention weights can be obtained

through a sigmoid activation:

$$A_i = \text{sigmoid}(e_i). \quad (4.4)$$

Then the refined feature map can be calculated as:

$$\hat{P}_i = A_i \odot \tilde{P}_i, \quad (4.5)$$

where $\odot$ denotes element-wise multiplication. This formulation allows the module to emphasize informative spatial locations without introducing additional parameters, making it particularly suitable for few-shot learning scenarios.

4.3.4 Cross-Layer Spatial Attention Module (CSAM)

While the SimAM module efficiently enhances salient neurons within each feature level, its capacity to model dependencies across different scales still remains limited. To address this, we design a Cross-Layer Spatial Attention Module (CSAM) that performs spatial reasoning between multi-scale features. For a target feature map P_i' , all refined feature maps are first resized to a common resolution and concatenated along the channel dimension:

$$F_i = \text{Concat} \left(\left\{ \text{Resize} \left(\hat{P}_j; H_i, W_i \right) \right\}_{j=1}^4 \right). \quad (4.6)$$

Then, we apply a 1×1 convolutional layer $g(\cdot)$ to generate a spatial attention map, followed by a sigmoid activation:

$$F_i' = \text{sigmoid}(g(F_i)). \quad (4.7)$$

Then the output of CSAM is computed as follows:

$$P_i' = \hat{P}_i \odot F_i' \quad (4.8)$$

4.3.5 Classification

For the enhanced feature maps $\{P_i'\}_{i=1}^4$ output from CSAM, we first apply a pooling layer on each layer to get a scale-specific descriptor::

$$v_i = \text{GAP}(P_i'), \quad (4.9)$$

where $\text{GAP}(\cdot)$ denotes global average pooling. Then we concatenate these descriptors to form a multi-scale embedding:

$$v = [v_1; v_2; v_3; v_4]. \quad (4.10)$$

Finally, we pass the embedding to a classifier $h(\cdot)$ to obtain the prediction:

$$y' = h(v). \quad (4.11)$$

4.4 Experiments

4.4.1 Benchmarks and Implementation Details

Benchmarks: Our experiments are carried out on three widely adopted FSL benchmarks: miniImageNet [14], tieredImageNet [62] and CIFAR-FS [112]. The miniImageNet dataset is a subset of ImageNet, including 100 classes with 600 images per class. Following the partitioning protocol in [115], we employ 64 classes for training, 16 classes for validation, and the remaining 20 classes for testing. The tieredImageNet dataset contains 608 classes organized into 34 coarse-grained categories. We select 20 categories (351 classes) for training, 6 categories (97 classes) for validation, and 8 categories (160 classes) for testing. CIFAR-FS is derived from the CIFAR100 dataset and is split into 64 training classes, 16 validation classes, and 20 testing classes, with each class containing 600 images.

Implementation Details: For miniImageNet and tieredImageNet, all images are resized to 84×84 , while images for CIFAR-FS are resized to 32×32 . We use a simple convolutional neural network consisting of four blocks (Conv4) as the backbone. Each block comprises the following components: a 64-filter 3×3 convolution layer, a batch normalization layer, a ReLU activation function and a 2×2 max-pooling layer. We use SGD with a momentum of 0.9 and a weight decay of $5e-4$. The initial learning rate is set as 0.05 and decayed with a factor of 0.1. Each epoch includes 64 samples, we train 150 epochs and decay three times for miniImageNet and tieredImageNet, 120 epochs and decay twice for CIFAR-FS. The evaluation protocol is the accuracy with the 95% confidence interval of the test set when 600 random tasks are selected, with 15 query samples each task. We train an N -way logistic regression classifier for the evaluation.

4.4.2 Experimental results

miniImageNet and tieredImageNet. Table 4.1 presents a comparison between our approach (AFPN) and the state-of-the-art methods on two popular FSL benchmarks. Across both datasets and under both 1-shot and 5-shot settings, AFPN obtains state-of-the-art performance. On miniImageNet, AFPN achieves 54.52% in the 5-way 1-shot task and 72.87% in the 5-way 5-shot task, surpassing the previously best-reported results by 1.65% and 1.85%, respectively. On tieredImageNet, AFPN achieves 57.62% and 74.69% in the 1-shot and 5-shot tasks, respectively. AFPN achieves improvements over the best-reported numbers by 2.48% under 5-way 1-shot and 0.46% under 5-way 5-shot settings.

Table 4.1: Comparison of AFPN (our approach) to prior works on miniImageNet and tieredImageNet datasets, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	miniImageNet 5-way		tieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [1]	Conv4-64	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66	53.31 $\pm$ 0.89	72.69 $\pm$ 0.74
MAML [100]	Conv4-64	48.07 $\pm$ 1.75	63.15 $\pm$ 0.91	51.67 $\pm$ 1.81	70.30 $\pm$ 1.75
RelationNet [24]	Conv4-64	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70	54.48 $\pm$ 0.93	71.31 $\pm$ 0.78
DeepEMD [37]	Conv4-64	51.72 $\pm$ 0.20	65.10 $\pm$ 0.39	51.22 $\pm$ 0.14	65.81 $\pm$ 0.68
CGRN [283]	Conv4-64	50.85 $\pm$ 0.86	64.13 $\pm$ 0.70	55.07 $\pm$ 0.20	71.34 $\pm$ 0.30
CovaMNet [151]	Conv4-64	51.19 $\pm$ 0.76	67.65 $\pm$ 0.63	54.98 $\pm$ 0.90	71.51 $\pm$ 0.75
MetaOptNet [237]	Conv4-64	52.87 $\pm$ 0.57	68.76 $\pm$ 0.48	54.71 $\pm$ 0.67	71.79 $\pm$ 0.59
DN4 [116]	Conv4-64	51.24 $\pm$ 0.74	71.02 $\pm$ 0.64	52.89 $\pm$ 0.23	73.36 $\pm$ 0.73
SSL-ProtoNet [284]	Conv4-64	52.58 $\pm$ 0.45	70.87 $\pm$ 0.36	55.14 $\pm$ 0.49	74.23 $\pm$ 0.40
AFPN	Conv4-64	54.52 $\pm$ 0.79	72.87 $\pm$ 0.60	57.62 $\pm$ 0.92	74.69 $\pm$ 0.70

CIFAR-FS. Table 4.2 illustrates similar comparisons on CIFAR-FS dataset. From Table 4.2, we can find that our method AFPN outperforms the other methods in comparison with the highest accuracy of 62.60% and 79.51% in 1-shot and 5-shot scenarios, respectively. In particular, AFPN achieves improvements over the best-reported numbers by 2.19% under 5-way 1-shot and 2.99% under 5-way 5-shot settings.

Table 4.2: Comparison of AFPN (our approach) to prior works on CIFAR-FS dataset, with mean accuracy (%) and 95% confidence interval.

Model	Backbone	CIFAR-FS 5-way	
		1-shot	5-shot
ProtoNet [1]	Conv4-64	55.50 $\pm$ 0.70	72.00 $\pm$ 0.69
MAML [100]	Conv4-64	58.90 $\pm$ 1.90	71.50 $\pm$ 1.00
CSS [249]	Conv4-64	56.49 $\pm$ 0.93	74.59 $\pm$ 0.72
SSL-ProtoNet [284]	Conv4-64	60.41 $\pm$ 0.52	76.52 $\pm$ 0.38
AFPN	Conv4-64	62.60 $\pm$ 0.91	79.51 $\pm$ 0.70

4.4.3 Ablation Study

Benefits of Feature Pyramid Module. To study the impact of feature pyramid module, we evaluate our approach with and without feature pyramid module on miniImageNet. As shown in Table 4.3, for the model trained without feature pyramid module, it achieves 54.12% and 72.45.% in 1-shot and 5-shot tasks, respectively. For the model trained with feature pyramid module, it achieves 0.4% and 0.42% improvements in 1-shot and 5-shot tasks, respectively. These results confirm the effectiveness of feature pyramid module.

Contributions of Each Layer in the Backbone. To more clearly assess the contributions of each layer in the backbone, we conduct an ablation study on mini-

4. ATTENTION-GUIDED FEATURE PYRAMID NETWORK FOR FEW-SHOT LEARNING

Table 4.3: FSL results on miniImageNet, with and without feature pyramid module.

Model	miniImageNet 5-way	
	1-shot	5-shot
AFPN (without feature pyramid module)	54.12 ± 0.79	72.45 ± 0.61
AFPN (with feature pyramid module)	54.52 ± 0.79	72.87 ± 0.60

ImageNet by training the model with different selection of layers. As reported in Table 4.4, AFPN with P_4 layer obtains accuracies of 53.91% and 72.14% for the 1-shot and 5-shot settings, respectively. AFPN gets better performance with the addition of other layers and gains 0.55% and 0.73% improvements after adding all the layers in the backbone. Overall, these results highlight the key role of each layer in the backbone.

Table 4.4: FSL results on miniImageNet, with different layers.

Selected Layer	miniImageNet 5-way	
	1-shot	5-shot
P_4	53.91 ± 0.81	72.14 ± 0.62
P_3, P_4	54.06 ± 0.79	72.22 ± 0.60
P_2, P_3, P_4	54.11 ± 0.80	72.46 ± 0.60
P_1, P_2, P_3, P_4	54.52 ± 0.79	72.87 ± 0.60

Benefits of Attention Module. To quantify the effect of attention module, we perform a comparison on miniImageNet by training the same model with different attention module. Baseline denotes our AFPN without any attention module, which only contains a backbone and a feature pyramid network module. Specifically, the baseline follows the standard AFPN pipeline, where the backbone extracts multi-level feature representations and the AFPN performs feature fusion across different scales. No additional attention mechanism is introduced in the baseline, so it can serve as a fair reference for evaluating the effectiveness of different attention modules. As shown in Table 4.5, The baseline reaches 52.73% in the 1-shot setting and 70.62% in the 5-shot setting. Incorporating SimAM module boosts the results, bringing 0.74% improvement for 1-shot and 1.02% for 5-shot compared to the baseline. AFPN further gets 1.05% and 1.23% improvements after introducing CSAM module. This consistent gain demonstrates that attention module plays a crucial role in improving performance.

4.5 Chapter Conclusions

In this chapter, we focus on a fundamental, yet underexplored challenge in few-shot image classification — how to effectively exploit multi-scale visual information

Table 4.5: FSL results on miniImageNet, with different attention module.

Model	miniImageNet 5-way	
	1-shot	5-shot
Baseline	52.73 ± 0.79	70.62 ± 0.64
AFPN with SimAM	53.47 ± 0.78	71.64 ± 0.62
AFPN with SimAM + CSAM	54.52 ± 0.79	72.87 ± 0.60

under limited supervision. To this end, we propose a novel Attention-Guided Feature Pyramid Network (AFPN), which is specifically designed to facilitate effective feature interaction across different network layers. In order to get discriminative representations, AFPN first introduces a Simple Parameter-Free Attention Module (SimAM) that efficiently highlights crucial spatial locations without adding learnable parameters, thereby preventing overfitting in data-scarce scenarios. Then, we design a Cross-Layer Spatial Attention Module (CSAM) which explicitly addresses the semantic gaps between adjacent pyramid layers through a structured fusion mechanism. Extensive experiments on standard benchmarks demonstrate that our AFPN achieves competitive performance compared to other state-of-the-art few-shot learning methods. The key insights derived from our experimental analysis can be summarized as follows:

- (1) As visualized in Table 4.5, The parameter-free attention design in SimAM proves its effectiveness in few-shot scenarios (e.g., 1-shot and 5-shot settings). The module successfully enhances model generalization while mitigating the overfitting risk in few-shot learning.
- (2) The CSAM mechanism shows complementary benefits with SimAM, as demonstrated by the performance gains observed in Table 4.5. Specifically, the integration module enables more effective propagation of high-level semantic information to lower layers while preserving fine-grained spatial details from earlier layers. This bidirectional information flow creates more discriminative representations for novel categories, with the most significant improvements observed in more challenging fine-grained classification tasks where subtle visual differences are critical for correct categorization.
- (3) AFPN keeps representational diversity, and ablation studies reveal that the combination of our proposed modules achieves better performance compared to using either module in isolation, showing their synergistic relationship. This finding emphasizes the importance of addressing both intra-layer feature refinement and inter-layer feature fusion in a coordinated framework for few-shot learning problems.

Chapter 5

Adaptive Multi-scale Feature Fusion Network for Few-shot Learning

In Chapters 3 and 4, we have proposed several methods to solve the challenges of the few-shot learning (FSL) theme. Unlike the core problem in FSL that is caused by the scarcity of labeled examples and the instability of feature representations under low-shot supervision, we move our research focus to a more general situation where models are supposed to learn with insufficient feature diversity. In this chapter, therefore, we aim to study effectively to enhance feature richness and discriminativeness under extreme data scarcity (**RQ 3**).

FSL presents critical limitations due to the scarcity of labeled training data, severely limiting the models' ability to learn discriminative feature representations. Issues such as insufficient feature diversity and scale sensitivity make image classification tasks challenging in FSL. To bridge this gap, we propose a novel adaptive multi-scale feature fusion (AMFF) network for FSL. AMFF includes two innovative modules, a feature enrichment module (FEM) and a feature fusion module (FFM). FEM increases feature richness through a multi-branch convolution structure and expands representation capacity with limited data. Then FFM reorganizes the multi-scale features with different weights, enabling the network to capture multi-scale information more effectively. Our approach systematically addresses the feature paucity problem through two complementary mechanisms: (1) expanding feature diversity via multi-scale feature extraction, and (2) enhancing feature discriminativeness through an adaptive feature fusion strategy. Extensive experiments on popular FSL benchmarks (miniImageNet, CUB) demonstrate our AMFF achieves competitive performances in few-shot classification tasks

This chapter is based on the following publication [35]:

- **He, K.**, Bakker, E. M., and Lew, M. S. "Adaptive Multi-scale Feature Fusion Network for Few-Shot Learning." International Conference on Intelligent Computing, 2025.

5.1 Introduction

Generally, deep learning approaches [285, 286, 287] achieve optimal performances based on a large amount of samples during the training process. However, in real-world scenarios, collecting and labeling large-scale datasets are time-consuming, or even infeasible. Few-shot learning (FSL) [6, 98] has emerged as a crucial part in the field of machine learning [288], solving the challenge of training models with limited labeled data. In FSL, only a few samples from each category are available for training. FSL is also denoted as an N -way K -shot task where N classes and K samples of each class are used to train the model. As a result, one of the challenges in FSL is to obtain a well-generalized model for unseen classes when trained with few samples.

In the constantly evolving field of FSL, feature fusion [289, 290] is a common way and has been widely-used in many few-shot learning models, such as adaptive feature fusion, inter-layer feature fusion and multi-layer feature fusion. By performing adaptive feature fusion, Ding et al. [291] proposed a meta-learning method which merges different features of each stage of the embedding model, and utilizes a particle swarm algorithm to optimize the feature fusion weights. Through the inter-layer fusion, Wu and Zhao [292] proposed a hierarchical metric FSL model, which constructs an intra-layer channel feature fusion method based on the network hierarchy and transfers the intra-layer fused features from the higher layer to the lower layer. Moreover, high-quality features can be extracted through this model driven by data and knowledge. In the multi-layer feature fusion model, Qin et al. [293] introduced multi-level feature representations when extracting features, and then exploit the fused multi-layer features as a subsequent metric to train the FSL model. Similarly, Gomes et al. [294] also introduced a multi-layer feature extractor and a similarity measurement approach for FSL, which effectively improves image pattern recognition performance through multi-layer feature fusion.

In this chapter, we propose a novel adaptive multiscale feature fusion (AMFF) approach for FSL. The overview of AMFF is presented in Figure 5.1. To address the insufficient feature problem, we first design a multi-branch convolution structure to extract multi-scale features, which is named as feature enrichment module (FEM). FEM enables the model to obtain more features from the very limited samples, effectively enriching the feature diversity. The model’s local perception capability can also be enhanced through these multi-scale features.

Furthermore, to better utilize the multi-scale features extracted by FEM and improve the model performance, we introduce a feature fusion module (FFM), which uses an adaptive strategy to merge these features. On one hand, FFM reorganizes the multi-scale features with different weights based on the adaptive strategy, aiming at obtaining discriminative features; on the other hand, the model’s performance on image classification tasks can be improved through these merged features.

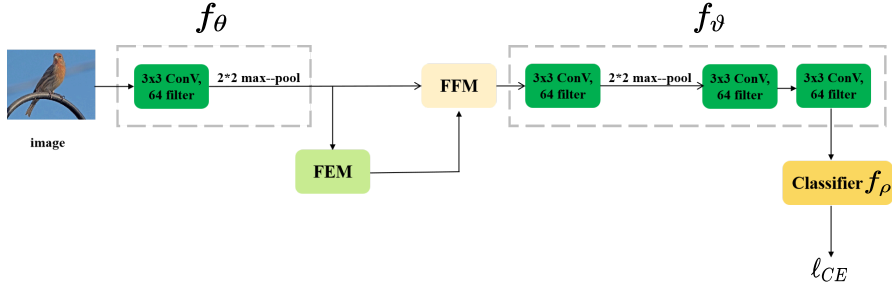


Figure 5.1: Overall training process of AMFF, where ℓ_{CE} denotes the standard cross-entropy loss, FEM is the feature enrichment module and FFM is the feature fusion module.

The main contributions of this work are summarized as follows: 1) We introduce a new multi-branch convolutional network to extract multi-scale features and to solve the insufficient features problem. 2) We propose an adaptive fusion strategy to merge multi-scale features and enable the obtained features to be more discriminative. 3) the experiments on several popular benchmark datasets demonstrate that our approach achieves new state-of-the-art performances in all the 5-way 1-shot and 5-way 5-shot tasks.

5.2 Related Work

Few-shot learning (FSL): FSL aims at classifying unseen tasks effectively through training with a few samples. Recently, a surge of FSL approaches have been proposed for image classification tasks. One popular branch of FSL is metric-based method [23, 295], which performs classification by measuring the metric (similarity or distance) between two samples. In 2015, Koch et al. [141] proposed a Siamese network to learn the similarity between images. Snell et al. [1] classified query samples based on the Euclidean distance to each class prototype, which is calculated by the mean value of samples from the same category. Prior works proved that a good embedding network is essential for FSL. Zhong et al. [296] introduced a novel generative model, which includes encoder, decoder, and discriminator, improving the model’s performance on few-shot Citrus aurantium L. diseases diagnosis. Lim et al. [284] explored knowledge transfer based on a pre-trained large scale model and then fine-tuned the model for the target few-shot task. In order to enhance the embedding representation, Lim et al. integrated FSL with self-supervised tasks.

Feature Fusion: In order to obtain a better feature representation and further improve the model’s performance, feature fusion is widely-used in the field of machine learning. By merging features from multiple different sources, feature fusion helps the model better understand data. Qiu et al. [297] proposed a traditional machine learning method based on multi-feature extraction, significantly improving

the classification accuracy. To alleviate the negative influence caused by inaccurate dependency tree information, Zhou et al. [298] applied a two-layer dynamic fusion mechanism, which adaptively adjusts the fusion weights. With the goal of enhancing the data and supplying the new category samples, Hua and Liu [299] fused multi-scale features extracted by a FPN network. Yu et al. [300] designed a deep multi-scale feature fusion module, interacting with feature information at different scales, which obtains good robustness and generalization ability.

5.3 Adaptive multi-scale feature fusion

In this section, we first make a brief introduction to common notions in FSL. Then we analyze and introduce a feature enrichment module (FEM) and a feature fusion module (FFM) in detail.

5.3.1 Problem formulation and notions

Firstly, we formulate the generalized FSL problem. We use y_{base} and y_{novel} to denote the labels of base and novel classes, respectively. Specifically, $y_{base} = \{1, \dots, C_s\}$ and $y_{novel} = \{C_s + 1, \dots, C_s + C_u\}$, where C_s and C_u are the number of base and novel classes, respectively. In FSL, samples from the base classes are used for training and those from the novel classes are used for evaluation. Besides, y_{base} and y_{novel} are disjoint, which means $y_{base} \cap y_{novel} = \emptyset$. During the evaluation process, samples from y_{novel} are split into two sets, support set and query set, where support set helps the learned model quickly adapt to new tasks and query set is used to evaluate the model's generalization ability. FSL is also named as N -way K -shot problem when the support set contains N classes and each class includes K samples.

5.3.2 Feature enrich module

The goal of feature enrich module (FEM) is to extract multi-scale features for few-shot tasks. In this work, we first introduce a feature extractor which consists of a 4-layer convolutional neural network as the backbone. We use f_θ and f_ϑ to denote the first layer and the subsequent three convolutional layers of the backbone, respectively. For any training image (x, y) from the base set D_{base} , its corresponding shallow features u_s can be obtained through f_θ , which is formulated as:

$$u_s = f_\theta(x). \quad (5.1)$$

Then, in order to obtain multi-scale features, we input the shallow features into FEM g_ϕ , where g_ϕ contains four branches with different kernel sizes. Both standard and depthwise convolutions are adopted in FEM. After that, we concat the multi-scale features which output from the four branches. The detailed diagram of FEM

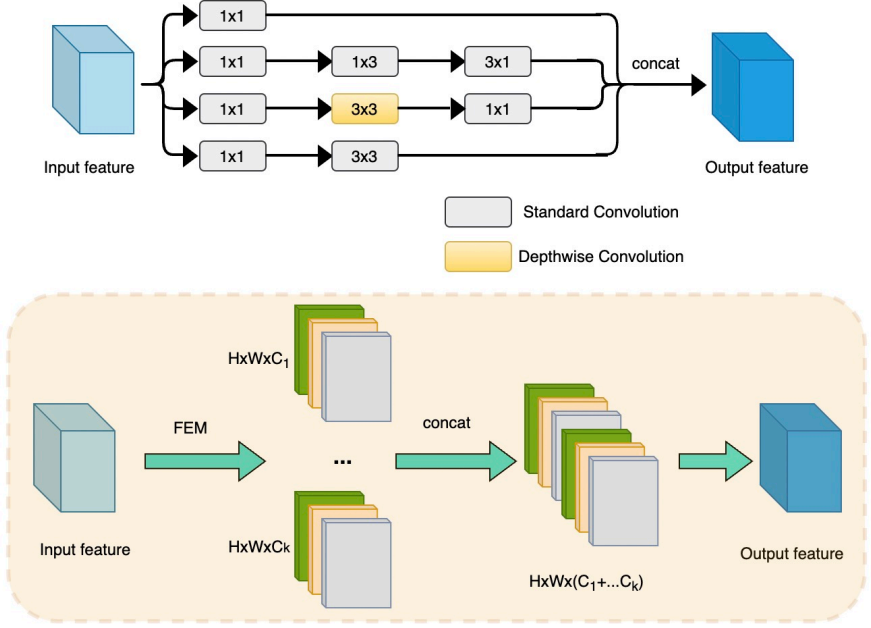


Figure 5.2: Structure of FEM.

is shown in Figure 5.2. Hence, the enriched features can be defined as:

$$u_e = g_\phi(u_s), \text{ where } u_s = f_\theta(x). \quad (5.2)$$

5.3.3 Feature fusion module

To extract discriminative features and improve the model's generalization ability for few-shot tasks, we design a novel feature fusion module (FFM) which consists of a feature interaction component (FIC) h_φ and a feature fusion component (FFC) h_ψ . Here, h_φ is implemented as a 1×1 convolution parameterized by φ , which captures dependencies among multi-scale features, while h_ψ is implemented as a 3×3 convolution parameterized by ψ , which integrates these interacted features to generate the final fused representation. The structure of FFM is shown in Figure 5.3. In order to enable the model to effectively capture multi-scale feature information, we first conduct a squeeze-and-excitation operation on the shallow features and then transmit them to FIC to improve the feature representation ability. The output features from FIC can be calculated as:

$$u_c = h_\varphi(u_s) = \text{Conv}_{1 \times 1}(u_s; \varphi). \quad (5.3)$$

After that, the updated features u_c and enriched features u_e are fused through FFC.

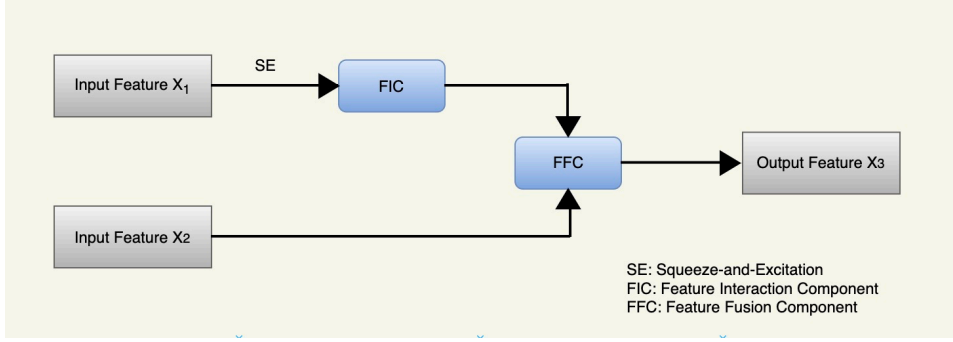


Figure 5.3: Structure of FFM.

Therefore, the fused features u_f can be defined as:

$$u_f = h_\psi(u_c, u_e) = \text{Conv}_{3 \times 1}(u_c + u_e; \psi). \quad (5.4)$$

Next, the fused features u_f are passed to the subsequent three layers, so the final deep features u_d can be calculated as:

$$u_d = f_\vartheta(u_f). \quad (5.5)$$

In the final step, we use a classifier f_ρ to perform classification tasks based on the deep features u_d , so the prediction y' can be calculated as:

$$y' = f_\rho(u_d). \quad (5.6)$$

As we formulate cross-entropy loss as the loss function, the optimization problem can be formulated as:

$$\begin{aligned} \rho, \theta, \vartheta, \phi, \varphi, \psi = \arg \min_{\rho, \theta, \vartheta, \phi, \varphi, \psi} \mathbb{E}_{(x, y) \sim D^{base}} \\ \ell_{CE}(f_\rho(f_\vartheta(h_\psi(h_\varphi(f_\theta(x)), g_\phi(f_\theta(x))))), y) \end{aligned} \quad (5.7)$$

The whole training procedure of our method AMFF is presented in Algorithm 2.

5.4 Experiments

In this section, we first describe the datasets, implementation details and evaluation protocols of our approach, and then make a comparison with state-of-the-art methods on popular benchmarks. Finally, we present an ablation study.

Algorithm 2: The training procedure of AMFF

Input: Training data D_{base} , training parameters such as epoch, batch size and learning rate.

Output: Backbone f_θ and f_ϑ , feature enrich module (FEM) g_ϕ , feature interaction component (FIC) h_φ and feature fusion component (FFM) h_ψ

for $iter = 1 \rightarrow epoch$ **do**

Sample a batch data from D_{base} ;
 Calculate the shallow features $u_s = f_\theta(x)$;
 Calculate the multi-scale features $u_e = g_\phi(u_s)$;
 Calculate the FIC output features $u_c = h_\varphi(u_s)$;
 Calculate the fused features $u_f = h_\psi(u_e, u_c)$;
 Calculate the deep features $u_d = f_\vartheta(u_f)$;
 Train the classifier f_ρ and calculate the loss ℓ_{CE} ;
 Update the parameters of f_ρ , f_ϑ , h_ψ , h_φ , g_ϕ and f_θ ;

end

Return f_ϑ , h_ψ , h_φ , g_ϕ and f_θ

5.4.1 Datasets, implementation details and evaluation protocol

Datasets. We select two benchmarks that have been widely used in FSL: mini-ImageNet [14] and CUB [113] datasets. Their details are presented in Table 5.1. MiniImageNet comes from the ImageNet dataset, containing 100 classes and each class consists of 600 images. MiniImageNet is divided into 64 training classes, 16 validation classes and 20 testing classes according to splitting protocol [115]. Images from miniImageNet are resized into $84 \times 84 \times 3$. CUB has a total of 11,788 images from 200 classes and was originally used for fine-grained bird classification tasks. We follow the split protocol proposed in [55] and divide CUB into three parts, 100 classes for training, 50 classes for validation, and 50 classes for testing. The size of the image is uniformly rescaled and set to $84 \times 84 \times 3$.

Table 5.1: Dataset description.

Dataset	miniImageNet			CUB		
	train	val	test	train	val	test
classes	64	16	20	100	50	50
samples	60000			11788		
image size	$84 \times 84 \times 3$			$84 \times 84 \times 3$		

Implementation details and evaluation protocol. To make a fair comparison with recent works, we use Conv4-64, which consists of 4 convolution layers with $64/64/64/64$ filters as our backbone. In each batch, we randomly select 64 samples from the base classes for training. The optimizer for the experiment is SGD with a

5. ADAPTIVE MULTI-SCALE FEATURE FUSION NETWORK FOR FEW-SHOT LEARNING

Table 5.2: Comparison of AMFF (our approach) to prior works on miniImageNet with mean accuracy (%) and 95% confidence interval.

Model	Backbone	miniImageNet 5-way	
		1-shot	5-shot
ProtoNet [1]	Conv4-64	49.42 $\pm$ 0.78	68.20 $\pm$ 0.66
RelationNet [24]	Conv4-64	50.44 $\pm$ 0.82	65.32 $\pm$ 0.70
MAML [100]	Conv4-64	48.07 $\pm$ 1.75	63.15 $\pm$ 0.91
IMP [301]	Conv4-64	49.60 $\pm$ 0.80	68.10 $\pm$ 0.80
Neg-Margin [126]	Conv4-64	52.84 $\pm$ 0.76	70.41 $\pm$ 0.66
MemoryNet [174]	Conv4-64	53.37 $\pm$ 0.48	66.97 $\pm$ 0.35
HFFCR[289]	Conv4-64	51.79 $\pm$ 0.85	65.74 $\pm$ 0.72
FEAT [239]	Conv4-64	55.15 $\pm$ 0.20	71.61 $\pm$ 0.16
MELR [302]	Conv4-64	55.35 $\pm$ 0.43	72.27 $\pm$ 0.35
SSL-ProtoNet [284]	Conv4-64	52.58 $\pm$ 0.45	70.87 $\pm$ 0.36
Match-sum [162]	SF4-64	55.74 $\pm$ 0.65	72.18 $\pm$ 0.70
AMFF	Conv4-64	56.00 $\pm$ 0.80	73.25 $\pm$ 0.60

momentum of 0.9 and a weight decay of $5e-4$. The initial learning rate is 0.05 and decayed with a factor of 0.1.

We train the model 100 and 150 epochs for miniImageNet and CUB datasets, respectively. Both of them are decayed twice during the training. In addition, random cropping, color jittering, and random horizontal flip are used as data augmentation. In this chapter, we take the common 5-way 1-shot and 5-way 5-shot settings to evaluate the performance of our method. In each task, 5 test classes (5-way) are randomly selected, and each class contains 1 or 5 support samples (K -shot) along with 15 query samples. To evaluate the classification performance, 600 unseen tasks are randomly constructed from the novel classes. Then, an N -way logistic regression classifier is trained for evaluation. Finally, we present the average 5-way K -shot classification accuracies with 95% confidence intervals.

5.4.2 Experimental results and ablative analysis

Experimental results. In this section, we carry out 5-way 1-shot and 5-way 5-shot experiments on miniImageNet and CUB datasets. For a fair comparison, we adopt Conv4-64 as the backbone. The experimental results of our AMFF and other SOTA approaches are reported in Table 5.2 and Table 5.3. Table 5.2 presents the performance comparison on the miniImageNet dataset. It is notable that our AMFF obtains 56.00% and 73.25% accuracy on 5-way 1-shot and 5-way 5-shot tasks, respectively. Compared to a deeper backbone SF4-64, our method AMFF outperforms with 0.26% and 1.07% improvements in 5-way 1-shot and 5-way 5-shot scenarios, respectively. Table 5.3 illustrates similar comparisons on CUB datasets. From Table 5.3, we can find that our method AMFF outperforms other methods

Table 5.3: Comparison of AMFF (our approach) to prior works on CUB dataset, with mean accuracy (%) and 95% confidence.

Model	Backbone	CUB 5-way	
		1-shot	5-shot
MatchingNet [14]	Conv4-64	61.16 $\pm$ 0.89	72.86 $\pm$ 0.70
ProtoNet [1]	Conv4-64	64.42 $\pm$ 0.48	81.82 $\pm$ 0.35
Hyperbolic [303]	Conv4-64	64.02 $\pm$ 0.24	82.53 $\pm$ 0.14
MAML [100]	Conv4-64	55.92 $\pm$ 0.95	72.09 $\pm$ 0.76
RelationNet[24]	Conv4-64	62.45 $\pm$ 0.98	76.11 $\pm$ 0.69
FEAT [239]	Conv4-64	68.87 $\pm$ 0.22	82.90 $\pm$ 0.15
DeepEMD [37]	Conv4-64	64.00	81.10
Match-sum [162]	SF4-64	67.35 $\pm$ 0.93	83.82 $\pm$ 0.61
AMFF	Conv4-64	69.75 $\pm$ 0.86	84.50 $\pm$ 0.57

in comparison with the highest accuracy of 69.75% and 84.50% in 1-shot and 5-shot scenarios, respectively. In particular, AMFF achieves improvements over the best-reported numbers by 0.88% under 5-way 1-shot and 0.68% under 5-way 5-shot settings.

Ablative analysis. To study the impact of FEM and FFM, we evaluate our approach with and without these modules on miniImageNet dataset. We use the original Conv4-64 as the baseline. The ablation results on miniImageNet are presented in Table 5.4. From the results, we can infer the importance of each module in our approach. From Table 5.4, we can find that FEM obtains a gain of 2.31% and 2.56% improvements compared to the baseline on miniImageNet 5-way 1-shot and 5-shot tasks, respectively. The significant improvements effectively demonstrate the contribution of FEM. FFM also helps achieve further 1.3% and 0.55% gains in 5-way 1-shot and 5-way 5-shot tasks, respectively. These results confirm the effectiveness of FFM.

Table 5.4: FSL results on miniImageNet with different models.

Model	miniImageNet 5-way	
	1-shot	5-shot
Baseline	52.39 $\pm$ 0.80	70.14 $\pm$ 0.64
Baseline+FEM	54.70 $\pm$ 0.78	72.70 $\pm$ 0.63
Baseline+FEM+FFM	56.00 $\pm$ 0.80	73.25 $\pm$ 0.60

5.5 Chapter Conclusions

In this work, we propose a novel adaptive multi-scale feature fusion network, which leverages multi-scale features and feature fusion to enhance the performance in few-shot image classification problems. First, to solve the insufficient features caused by

5. ADAPTIVE MULTI-SCALE FEATURE FUSION NETWORK FOR FEW-SHOT LEARNING

limited samples during the training process, we introduce a feature enrichment module, which consists of 4 branches with different kernel size. We exploit the feature enrichment module to generate multi-scale features and enrich feature representations. Then, to further obtain discriminative features and improve the model’s generalization ability, we design a feature fusion module to merge these multi-scale features. The feature fusion module adaptively reorganizes these multi-scale features and effectively improves the model’s robustness. Extensive experimental results demonstrate the importance of the two modules in our approach and we obtain better performance than other SOTA approaches on popular few-shot learning benchmark datasets. Since the proposed method is only used in image classification tasks in this chapter, we will pay attention to other fields in the future, such as object detection and semantic segmentation tasks.

Chapter 6

Dual Selective Knowledge Transfer for Few-Shot Classification

In the previous Chapters 3, 4 and 5, we have explored various aspects of feature representation and metric optimization for few-shot learning (FSL) tasks. Starting with this chapter, we research on a more fundamental challenge in knowledge transfer for FSL - the effective utilization of pre-trained knowledge while maintaining adaptability to novel tasks. Therefore, in this chapter, we focus on understanding how transferable knowledge can be selectively leveraged to build more generalizable feature representations for few-shot tasks (**RQ 4**).

Few-shot learning aims at recognizing novel visual categories from very few labelled examples. Different from the existing few-shot classification methods that are mainly based on metric learning or meta-learning, in this work we focus on improving the representation capacity of feature extractors. For this purpose, we propose a new two-stage dual selective knowledge transfer (DSKT) framework, to guide models towards better optimization. Specifically, we first exploit an improved multi-task learning approach to train a feature extractor with robust representation capability as a teacher model. Then, we design an effective dual selective knowledge distillation method, which enables the student model to selectively learn knowledge from the teacher model and current samples, thereby improving the student model’s ability to generalize on unseen classes. Extensive experimental results show that our DSKT achieves competitive performances on four well-known few-shot classification benchmarks.

This chapter is based on the following publication [36]:

- **He, K.**, Pu, N., Lao, M., Bakker, E. M., and Lew, M.S. “Dual selective knowledge transfer for few-shot classification.” *Applied Intelligence*, 2023.

6.1 Introduction

Deep learning models have achieved breakthroughs in many visual understanding tasks, such as image classification [304] and object detection [305]. However, their performances degrade significantly when few labelled examples are available and these large-capacity models usually have difficulties in transferring the learnt knowledge to unseen classes. As such, there has been increasing interest in few-shot learning (FSL) [1, 24, 101], which aims to develop well-generalized models to recognize new categories using only limited annotated samples.

Recently, meta-learning [2, 14, 115], *a.k.a.* learning-to-learn, has made significant advances in FSL. The primary idea of meta-learning is to exploit episodic training to transfer the knowledge learnt from a massive known meta-training set to novel classes. For example, Sun et al. [198] proposed Meta-Transfer Learning, which introduced the scaling and translation parameters to adjust the weight parameters, meeting the needs of new tasks. However, the success of transfer ability depends on the similarity between tasks. It's difficult to transfer knowledge effectively when new tasks are significantly different from old ones. Another practical solution to FSL are metric-based methods [23, 116, 306], which aim to learn good representations through deep networks and generate final predictions based on the similarity or distance between support and query samples, such as RelationNet [24] and ProtoNet [1]. Though metric learning has become the prominent methods for FSL, it's still challenging to select an appropriate distance metric for various tasks and datasets, as different metrics can yield different results.

In this chapter, we propose a new dual selective knowledge transfer (DSKT) framework for FSL tasks, following the transfer learning strategy. The framework and overall training process are illustrated in Figure 6.1. In the first stage of the DSKT framework (DSKT-1), we introduce a multi-task loss to mine useful signals or information from the limited labelled examples. This enables the model to reduce the risk of overfitting and to ensure heterogeneity in the prediction space, simultaneously.

Furthermore, to implement selective knowledge transfer, we propose an effective dual selective knowledge distillation method in the second stage (DSKT-2). Specifically, we first employ the learnt model in DSKT-1 as a teacher network and train a student model based on the teacher's outputs. Then, different from the classical knowledge distillation (KD) [307] loss that considers different classes equally, we reformulate the KD loss into two independent components according to the class labels of samples, inspired by [308]. By separately fine-tuning the KD strength of each component, models are endowed with the ability to selectively transfer knowledge. Meanwhile, we introduce PolyLoss [309] to adjust the strengths of learning on current samples, which complements the knowledge transferred from the teacher model. After training, we freeze the student model up to the penultimate layer and

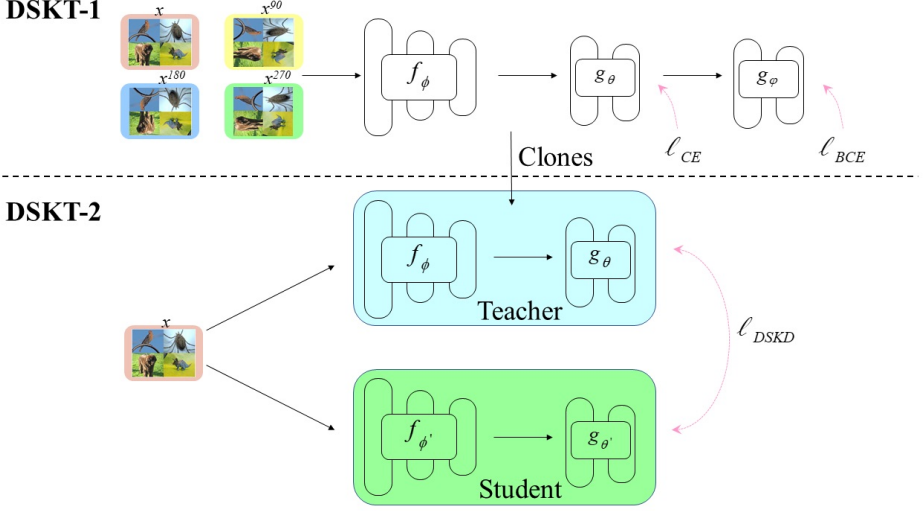


Figure 6.1: Overall training process of DSKT, where f_ϕ is a feature extractor, g_θ and g_φ are two different classifiers. DSKT-1 denotes the first process which includes a binary cross entropy (BCE) loss and a cross entropy (CE) loss, with the goal to learn a good feature extractor. DSKT-2 represents the stage of dual selective knowledge distillation (DSKD), which selectively transfers knowledge to ensure the student model’s generalization ability.

apply it as a feature extractor. During the evaluation, we fit a logistic regression classifier based on the frozen feature extractor without fine-tune, implemented as in [110].

Compared to previous work [269], we take similar action in the first stage to train a discriminative feature extractor. However, we design a distinct DSKD framework for knowledge distillation. The difference and superiority can be simply concluded as follows: 1) We only use the original samples during the knowledge distillation process, which reduces computational costs to some extent. 2) We propose a new DSKD framework for knowledge distillation, which integrates decoupled knowledge distillation (DKD) and PolyLoss, enabling the student model could effectively learn knowledge from both the teacher model and current samples, enhancing the student model’s capacity to generalize on new tasks.

The main contributions of this work are as follows: 1) We incorporate KD and PolyLoss as a knowledge transfer framework for FSL tasks. 2) We exploit DKD to improve the conventional KD, which enables the student model to selectively learn knowledge from the teacher. 3) the experiments on four popular benchmark datasets show that our approach achieves new state-of-the-art performances in all the 5-way 1-shot and 5-way 5-shot tasks.

6.2 Related Work

Few-shot learning (FSL): As a surprising research area in deep learning, FSL focus on learning patterns from a set of data (base dataset) and then adapting to a disjoint set (novel dataset) with few labelled samples. In order to evaluate the model’s performance, the novel dataset is split into a support set and a query set. FSL is also named N -way K -shot classification when the support set contains N classes and K samples per class.

Two main training methods have been explored in few-shot learning, namely meta-learning and metric learning. On the one hand, meta-learning consists of two phases, namely meta-training and meta-testing, where each phase involves a family of tasks. Meta-learning aims to perform well on new tasks based on the knowledge learnt from previous tasks. For example, Finn et al. [2] proposed a model-agnostic meta-learning (MAML) approach, which can adapt to new tasks quickly with a better initialization weight. With the goal to simplify MAML, Reptile [310] removes re-initialization for each task. MetaOptNet [237] employs SVM to replace its original linear classifier and solves the convex optimization problem by using quadratic programming (QP). Sun et al. [198] proposed a meta-transfer learning (MTL) method, which learns to adapt a deep NN for few-shot learning tasks. Bansal et al. [311] proposed a meta-learning model, which combines supervised learning and self-supervised learning for natural language classification tasks.

On the other hand, metric learning pays attention to finding an existing or learned metric space, where the support set can be simply matched with the query set. In metric learning, additional parameters are not required to learn once the metric is acquired. Snell et al. [1] proposed ProtoNet, which learns a metric space where classification can be performed by computing distances to prototype representations of each class. RelationNet [24] applies a network to learn the relation (similarity) between support and query images. Based on memory module, MatchingNet [14] uses cosine distance as a metric to classify query samples. To achieve better classification, CAN [22] is designed to highlight the proper region of interest by learning an attention module. Zhang et al. [37] proposed a DeepEMD algorithm, which adopts the Earth Mover’s Distance (EMD) as a distance metric to calculate the similarity. Different from these approaches, we propose a new two-stage knowledge transfer framework, further improving the model’s generalization ability for FSL tasks.

Knowledge distillation (KD): With the goal to deploy demanding deep models on devices with limited resources, KD is developed for model compression and acceleration. KD works by effectively transferring knowledge from a large and complex model, called teacher model, to a smaller and simple one, called student model. During the process of KD, the student’s learning performance is influenced by multiple factors, such as knowledge types, distillation strategies and teacher-student architectures.

Hinton et al. [307] first generalized and brought this idea into deep learning. By minimizing the loss between the teacher model’s and student model’s outputs, informative knowledge can be transferred. Tasks [312, 313] show KD’s benefits for knowledge transfer and optimization. Besides, sequential distillation [314] is introduced to improve the performance of teacher models. Mobahi et al. [315] presented a theoretical analysis of self-distillation. Lim et al. [316] proposed an Efficient-PrototypicalNet, which involves both transfer learning and knowledge distillation for few-shot learning tasks. Liu et al. [317] proposed learning a model through online self-distillation, which combines supervised training with knowledge distillation via a continuously updated teacher. A novel Supervised Masked Knowledge Distillation model (SMKD) [318] is also designed for few-shot Transformers, which incorporates label information into self-distillation frameworks. Unlike commonly-used KD which only transfers knowledge from the teacher model, we use a dual selective knowledge distillation method to encourage the student model to selectively learn knowledge from the teacher model and examples simultaneously.

6.3 Dual Selective Knowledge Transfer

To address the challenges caused by FSL, we propose a new dual selective knowledge transfer (DSKT) framework, which consists of two stages: multi-task learning and dual selective knowledge distillation (DSKD). In the first stage, we introduce multi-task learning and enforce the learnt representations equivariant to image transformations, which is beneficial for extracting low-level features. Furthermore, we propose an effective DSKD, which enables the student model to selectively learn knowledge from both the teacher model and the current samples.

6.3.1 Problem Formulation

In this work, we use a large-scale labelled base dataset for training a feature extractor. The base dataset is defined as $D^{base} = \{x_t^{base}, y_t^{base}\}_{t=1}^{N^{base}}$, with label $y_t^{base} \in C^{base}$. In order to learn a good feature extractor, we hold the assumption that both the amount of classes ($|C^{base}|$) and the number of examples (N^{base}) are large. For the novel dataset used for evaluation, is denoted as $D^{novel} = \{x_t^{novel}, y_t^{novel}\}_{t=1}^{N^{novel}}$, with label $y_t^{novel} \in C^{novel}$. Notice that base classes and novel classes are disjoint, which means $C^{base} \cap C^{novel} = \emptyset$. The testing of the learnt model is organized in episodes, in which each episode i contains a support set $D_i^{support} = \{x_{i,t}^{support}, y_{i,t}^{support}\}_{t=1}^{CK}$ and a query set $D_i^{query} = \{x_{i,t}^{query}, y_{i,t}^{query}\}_{t=1}^{CK'}$, which contains C classes, with K and K' examples per class from the novel dataset D^{novel} , respectively.

6.3.2 Multi-task Learning

In the first stage of our DSKT framework, we introduce a multi-task learning method, which consists of a category task and a rotation task to train a discriminative feature extractor with good representations on the base dataset. Following a similar training strategy [110], we adopt a neural network that contains a feature extractor f_ϕ , a category classifier g_θ and an additional rotation classifier g_φ .

First, we randomly sample a minibatch $B = \{x_i, y_i\}_{i=1}^m$ from the base dataset D^{base} , where m stands for the batch size. We impose the rotate transformation with 90, 180 and 270 degrees on the images x , to generate augmented copies x^{90} , x^{180} and x^{270} , respectively. Then, we stack the original images and their transformed versions together, resulting in a single tensor $\hat{x} = \{x, x^{90}, x^{180}, x^{270}\} \in \mathbb{R}^{(h \times w) \times (4 \times m)}$ with corresponding labels $\hat{y} \in \mathbb{R}^{4 \times m}$. According to the rotation direction, we also create one-hot encoded labels $\hat{r} = \{r_i \in \mathbb{R}^s\}_{4 \times m}$, where $s = 4$ indicates that four rotation directions are applied in our method.

Next, by using the feature extractor f_ϕ , the stacked tensor $\hat{x}$ is mapped to feature vectors $\hat{v} = f_\phi(\hat{x}) \in \mathbb{R}^{d \times (4 \times m)}$, where d denotes the feature map size. Then, we pass the feature vectors $\hat{v}$ through category classifier g_θ to obtain its corresponding logits $\hat{p} = g_\theta(\hat{v}) \in \mathbb{R}^{c \times (4 \times m)}$. Finally, we apply rotation classifier g_φ to map the logits $\hat{p}$ to the rotation logits $\hat{q} = g_\varphi(\hat{p}) \in \mathbb{R}^{s \times (4 \times m)}$.

Afterwards, to train all network modules (i.e., f_ϕ , g_θ and g_φ), we use two loss functions to jointly optimize three modules. One is a commonly-used cross entropy loss [110] for classifying categories of samples, which is denoted by ℓ_{CE} . The other is a binary cross entropy loss for rotation prediction, which serves as the rotation loss ℓ_{BCE} . The optimization problem is formulated as:

$$\begin{aligned} \phi, \theta, \varphi = \arg \min_{\phi, \theta, \varphi} \mathbb{E}_{(x, y) \sim D^{\text{base}}} & \left[\ell_{CE} \left(g_\theta \left(f_\phi \left(\hat{x} \right) \right), \hat{y} \right) \right. \\ & \left. + \alpha \ell_{BCE} \left(g_\varphi \left(g_\theta \left(f_\phi \left(\hat{x} \right) \right) \right), \hat{r} \right) \right]. \end{aligned} \quad (6.1)$$

where α is a weighting coefficient to control the strength of rotation loss.

Learning a good feature extractor is challenging in FSL as limited labelled samples are available during this process. In this work, we introduce multi-task learning and it enables the feature extractor to effectively learn the low-level features [319]. In addition, the risk of overfitting can be reduced through multi-task loss [319]. Experimental results in Table 6.4 demonstrate the superiority of our method.

6.3.3 Dual Selective Knowledge Distillation

To improve the model’s generalization ability by knowledge transfer, we develop an effective dual selective knowledge distillation (DSKD) to promote models in learn-

ing informative knowledge. Once the first stage is finished, we take one clone of the trained model as a teacher model, where weights are frozen and only used for inference. The knowledge distillation (KD) [307] loss is reformulated into two components, by separately adjusting the strength of each component, informative knowledge can be selectively transferred to a new student model, which only contains a feature extractor and a classifier. In addition, PolyLoss [309] is introduced to ensure the student model selectively learns knowledge from current training samples. By considering such dual learning strategies, our DSKD enables the student model to simultaneously learn knowledge from the teacher model and current samples.

To make the different components of the transferred knowledge controllable, we first reformulate the KD loss into target class knowledge distillation (TCKD) loss and non-target class knowledge distillation (NCKD) loss, inspired by [308]. For the i -th input image, $p_i \in \mathbb{R}^c$ denotes its output logits, and $p_{i,t}$ represents the logit of the t -th class. Hence, the probability that the i -th sample belongs to the t -th class $z_{i,t}$ and the probability that it belongs to any of the other classes $z_{i,\setminus t}$ can be formulated as:

$$z_{i,t} = \frac{e^{p_{i,t}}}{\sum_j e^{p_{i,j}}}, \quad z_{i,\setminus t} = \frac{\sum_{k=1, k \neq t}^c e^{p_{i,k}}}{\sum_j e^{p_{i,j}}} = 1 - z_{i,t}. \quad (6.2)$$

Meanwhile, the vector $\hat{z}_i$ denotes the probability among non-target classes (i.e., the t -th class is not considered),

$$\hat{z}_i = \left[\hat{z}_{i,1}, \dots, \hat{z}_{i,t-1}, \hat{z}_{i,t+1}, \dots, \hat{z}_{i,c} \right], \quad (6.3)$$

where $\hat{z}_{i,k} = \frac{e^{p_{i,k}}}{\sum_{j=1, j \neq t}^c e^{p_{i,j}}}.$

Thereby, TCKD loss and NCKD loss can be defined as:

$$\ell_{TCKD} = \text{KL}(b_i^T || b_i^S), \quad \ell_{NCKD} = \text{KL}\left(\hat{z}_i^T || \hat{z}_i^S\right), \quad (6.4)$$

where $b_i = [z_{i,t}, z_{i,\setminus t}]^T \in \mathbb{R}^{2 \times 1},$

where KL denotes Kullback-Leibler divergence, T and S represent the teacher and student model, respectively. By adjusting the strength of ℓ_{TCKD} and ℓ_{NCKD} , the student model can selectively learn the knowledge from well-predicted samples.

Apart from learning from the teacher model, we further explore adjustable learning from current samples by introducing PolyLoss [309]. PolyLoss provides a unified view on common loss functions for classification problems. It defines loss function as a linear combination of polynomial functions. Under polynomial expansion, focal loss is a horizontal shift of the polynomial coefficients compared to cross entropy loss. In this chapter, PolyLoss is used to calculate the loss between predictions and ground-truth labels so as to guide the student model's training toward good

optimization. The function of PolyLoss is defined as follows:

$$\ell_{Poly-N} = -\log(P_t) + \sum_{k=1}^N \beta_k (1 - P_t)^k, \quad (6.5)$$

where P_t denotes the model’s prediction probability of the target ground-truth class, N represents the number of leading coefficients. With this flexible form, PolyLoss can easily adjust the importance of different polynomial bases according to the targeting datasets and tasks, consistently improving the performance. In this work, we set $N = 1$ and the experimental results show it achieves significant improvement.

Finally, the optimization problem of DSKD can be written as,

$$\begin{aligned} \phi', \theta' = \arg \min_{\phi', \theta'} & \mathbb{E}_{(x,y) \sim D^{base}} \\ & [\eta_1 \ell_{Poly-1}(g_{\theta'}(f_{\phi'}(x)), y) \\ & + \eta_2 [\lambda_1 \ell_{TCKD}(g_{\theta'}(f_{\phi'}(x)), g_{\theta}(f_{\phi}(x)), y) \\ & + \lambda_2 \ell_{NCKD}(g_{\theta'}(f_{\phi'}(x)), g_{\theta}(f_{\phi}(x)), y)] \end{aligned} \quad (6.6)$$

where $\eta_2 = 1 - \eta_1$, λ_1 and λ_2 are the weights of ℓ_{TCKD} and ℓ_{NCKD} , respectively. Our proposed DSKD method enables the student model to selectively learn knowledge from the teacher model and current samples.

6.4 Experiments

In this section, we elaborate on the experimental configuration and evaluation. The description of the datasets and implementation details are first presented, followed by experimental results of our approach on popular benchmark datasets. Finally, an ablative analysis is presented.

6.4.1 Datasets and Implementation details

Datasets: We conduct extensive experiments on four popular FSL benchmark datasets, miniImageNet [14], tieredImageNet [62], Fewshot-CIFAR100 (FC100) [23] and Caltech-UCSD Birds-200-2011 (CUB) [113]. For miniImageNet, as a subset of ImageNet, it includes 100 classes and each class contains 600 images. We follow the splitting protocol proposed in [115], with 64 classes for training, 16 classes for validation and 20 classes for testing. The tieredImageNet contains 608 classes, which can be grouped into 34 high-level categories. We use 20 categories (351 classes), 6 categories (97 classes), and 8 categories (160 classes) for training, validation and testing, respectively. FC100 is derived from CIFAR100 dataset and employs a split similar to tieredImageNet, which can be divided into 60 training classes, 20 validation classes and 20 testing classes. Each class includes 600 images. The CUB was originally used for fine-grained bird classification, which has 11,788 images from 200

Table 6.1: Comparison of DSKT (our approach) to prior works on miniImageNet and tieredImageNet datasets, with mean accuracy (%) and 95% confidence interval. ‡ indicates a deeper backbone. DSKT-1 and DSKT-2 represent the first and second stage, respectively.

Model	Backbone	miniImageNet 5-way		tieredImageNet 5-way	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [1] (NIPS’17)	ResNet-12	60.37 $\pm$ 0.83	78.02 $\pm$ 0.57	65.65 $\pm$ 0.92	83.40 $\pm$ 0.65
TADAM [23] (NIPS’18)	ResNet-12	58.50 $\pm$ 0.30	76.70 $\pm$ 0.30	-	-
MTL [198] (CVPR’19)	ResNet-12	61.20 $\pm$ 1.80	75.50 $\pm$ 0.80	65.62 $\pm$ 1.80	80.61 $\pm$ 0.90
MetaOptNet [237] (CVPR’19)	ResNet-12	62.64 $\pm$ 0.61	78.63 $\pm$ 0.46	65.99 $\pm$ 0.72	81.56 $\pm$ 0.53
TapNet [107] (ICML’19)	ResNet-12	61.65 $\pm$ 0.15	76.36 $\pm$ 0.10	63.08 $\pm$ 0.15	80.26 $\pm$ 0.12
Shot-Free [320] (ICCV’19)	ResNet-12	59.04 $\pm$ 0.43	77.64 $\pm$ 0.39	66.87 $\pm$ 0.43	82.64 $\pm$ 0.43
DeepEMD [37] (CVPR’20)	ResNet-12	65.91 $\pm$ 0.82	82.41 $\pm$ 0.56	71.16 $\pm$ 0.87	86.03 $\pm$ 0.58
DSN-MR [106] (CVPR’20)	ResNet-12	64.60 $\pm$ 0.72	79.51 $\pm$ 0.50	67.39 $\pm$ 0.83	82.85 $\pm$ 0.56
FEAT [239] (CVPR’20)	ResNet-12	66.78 $\pm$ 0.20	82.05 $\pm$ 0.14	70.80 $\pm$ 0.23	84.79 $\pm$ 0.16
Neg-Cosine [126] (ECCV’20)	ResNet-12	63.85 $\pm$ 0.81	81.57 $\pm$ 0.56	-	-
AssoAlign [54] (ECCV’20)	ResNet-18‡	59.88 $\pm$ 0.67	80.35 $\pm$ 0.73	69.29 $\pm$ 0.56	85.97 $\pm$ 0.49
RFS-distill [110] (ECCV’20)	ResNet-12	64.82 $\pm$ 0.82	82.41 $\pm$ 0.43	71.52 $\pm$ 0.69	86.03 $\pm$ 0.49
P-Transfer [67] (AAAI’21)	ResNet-12	64.21 $\pm$ 0.77	80.38 $\pm$ 0.59	-	-
ALFA+MeTA [196] (ICCV’21)	ResNet-12	66.61 $\pm$ 0.28	81.43 $\pm$ 0.25	70.29 $\pm$ 0.40	86.17 $\pm$ 0.35
MixtFSL [161] (ICCV’21)	ResNet-12	63.98 $\pm$ 0.79	82.04 $\pm$ 0.49	70.97 $\pm$ 1.03	86.16 $\pm$ 0.67
BML [256] (ICCV’21)	ResNet-12	67.04 $\pm$ 0.63	83.63 $\pm$ 0.29	68.99 $\pm$ 0.50	85.49 $\pm$ 0.34
FRN [122] (CVPR’21)	ResNet-12	66.45 $\pm$ 0.19	82.83 $\pm$ 0.13	71.16 $\pm$ 0.22	86.01 $\pm$ 0.15
SKD-GEN1 [269] (BMVC’21)	ResNet-12	67.04 $\pm$ 0.85	83.54 $\pm$ 0.54	72.03 $\pm$ 0.91	86.50 $\pm$ 0.58
APP2S [321] (AAAI’22)	ResNet-12	66.25 $\pm$ 0.20	83.42 $\pm$ 0.15	72.00 $\pm$ 0.22	86.23 $\pm$ 0.15
DCAP [322] (ACM’22)	ResNet-12	65.20 $\pm$ 0.67	80.93 $\pm$ 0.53	70.15 $\pm$ 0.74	85.33 $\pm$ 0.55
MDM-Net [323] (IJMLC’22)	ResNet-12	59.88 $\pm$ 0.42	76.60 $\pm$ 0.24	-	-
DSKT-1	ResNet-12	65.88 $\pm$ 0.81	83.09 $\pm$ 0.54	71.60 $\pm$ 0.90	86.61 $\pm$ 0.60
DSKT-2	ResNet-12	67.33 $\pm$ 0.82	84.19 $\pm$ 0.50	72.11 $\pm$ 0.89	86.69 $\pm$ 0.59

classes. We follow the split division in [55] that 100, 50, and 50 classes are grouped for training, validation and testing, respectively.

Implementation details: To make a fair comparison with recent works [110], we use ResNet12 as our backbone. We follow [237] to apply a Dropblock as a regularizer and adjust the amount of filters from (64, 128, 256, 512) to (64, 160, 320, 640). Besides, a 4-neuron fully-connected layer is applied after the final classification layer. Each batch contains 64 samples. We use SGD with a momentum of 0.9 and a weight decay of $5e-4$. The initial learning rate is set to 0.05 and decayed with a factor of 0.1. We train 100 epochs and decay twice for miniImageNet and CUB, we train 60 epochs and decay three times for tieredImageNet, and finally we train 65 epochs and decay once for FC100. The same learning schedule is applied during distillation. Furthermore, we use random cropping, color jittering and random horizontal flip for data augmentation during the whole process. For the hyper-parameters, we set the temperature coefficient to 4.0 and $\eta_1 = \eta_2 = 0.5$, where α , β_1 , λ_1 and λ_2 are tuned on the validation dataset. For evaluation, we train a N -way logistic regression classifier.

6.4.2 Evaluations

miniImageNet and tieredImageNet. Table 6.1 presents a comparison between our approach and the state-of-the-art methods in FSL tasks on the two ImageNet-

6. DUAL SELECTIVE KNOWLEDGE TRANSFER FOR FEW-SHOT CLASSIFICATION

Table 6.2: Comparison of DSKT (our approach) to prior works on FC100 dataset, with mean accuracy (%) and 95% confidence interval. ‡ indicates a deeper backbone. DSKT-1 and DSKT-2 represent the first and second stage, respectively.

Model	Backbone	FC100 5-way	
		1-shot	5-shot
ProtoNet [1] (NIPS’17)	ResNet-12	37.5 $\pm$ 0.6	52.5 $\pm$ 0.6
TADAM [23] (NIPS’18)	ResNet-12	40.1 $\pm$ 0.4	56.1 $\pm$ 0.4
MTL [198] (CVPR’19)	ResNet-12	45.1 $\pm$ 1.8	57.6 $\pm$ 0.9
MetaOptNet [237] (CVPR’19)	ResNet-12	41.1 $\pm$ 0.6	55.5 $\pm$ 0.6
DeepEMD [37] (CVPR’20)	ResNet-12	46.5 $\pm$ 0.8	63.2 $\pm$ 0.7
AssoAlign [54] (ECCV’20)	ResNet-18‡	45.8 $\pm$ 0.5	59.7 $\pm$ 0.6
RFS-distill [110] (ECCV’20)	ResNet-12	44.6 $\pm$ 0.7	60.9 $\pm$ 0.6
MixtFSL [161] (ICCV’21)	ResNet-12	44.9 $\pm$ 0.6	60.7 $\pm$ 0.7
Meta-Navigator [324] (ICCV’21)	ResNet-12	45.6 $\pm$ 0.8	59.9 $\pm$ 0.8
SKD-GEN1 [269] (BMVC’21)	ResNet-12	46.5 $\pm$ 0.8	63.1 $\pm$ 0.7
CORL [259] (WACV’22)	ResNet-12	44.8 $\pm$ 0.7	61.3 $\pm$ 0.5
MDM-Net [323] (IJMLC’22)	ResNet-12	43.6 $\pm$ 0.4	57.4 $\pm$ 0.3
DSKT-1	ResNet-12	45.4 $\pm$ 0.8	62.7 $\pm$ 0.7
DSKT-2	ResNet-12	46.6 $\pm$ 0.8	63.7 $\pm$ 0.7

based benchmarks. Our method is denoted as DSKT, where DSKT-1 and DSKT-2 represent the first and second stage, respectively. On both datasets and in both 1-shot and 5-shot scenarios, our approach yields state-of-the-art results. On mini-ImageNet, DSKT-1 achieves 65.88% and 83.09% in 5-way 1-shot and 5-way 5-shot tasks, respectively. This shows a gain of 1.06% and 0.68% over RFS-distill. The improvement becomes more substantial after distillation, with DSKT-2 producing a gain of 2.51% and 1.78% over RFS-distill under 5-way 1-shot and 5-way 5-shot settings, respectively. On tieredImageNet, DSKT-1 achieves improvements over RFS-distill by 0.08% and 0.58% in the 1-shot and 5-shot tasks, respectively. The improvements are 0.59% and 0.66% with DSKT-2.

FC100. Table 6.2 illustrates similar comparisons, this time on FC100. Here, DSKT provides accuracy improvements in all cases. For DSKT-1, the improvements over RFS-distill for 1-shot and 5-shot are 0.8% and 1.8%, respectively. Furthermore, the addition of distillation (DSKT-2) shows an exclusive improvement of 1.2% under 5-way 1-shot and 1.0% under 5-way 5-shot settings.

CUB. Table 6.3 compares our approach DSKT, against the state-of-the-art on CUB for fine-grained classification. Here, our method outperforms previous works even if they are implemented with deeper backbones. In particular, DSKT-2 achieves improvements over the best-reported numbers by 0.68% and 1.04% in 5-way 1-shot and 5-way 5-shot scenarios, respectively.

Table 6.3: Comparison of DSKT (our approach) to prior works on CUB dataset, with mean accuracy (%) and 95% confidence interval. ‡ indicates a deeper backbone. DSKT-1 and DSKT-2 represent the first and second stage, respectively.

Model	Backbone	CUB 5-way	
		1-shot	5-shot
MatchNet [14] (NIPS’16)	ResNet-12	71.87 ± 0.9	85.08 ± 0.6
ProtoNet [1] (NIPS’17)	ResNet-18‡	71.88 ± 0.9	86.64 ± 0.5
MAML [2] (ICML’17)	ResNet-18‡	68.42 ± 1.0	83.47 ± 0.6
RelationNet [24] (CVPR’18)	ResNet-18‡	67.59 ± 1.0	82.75 ± 0.6
DEML [325] (Arxiv’18)	ResNet-50‡	66.95 ± 1.0	77.11 ± 0.8
Robust-20 [326] (ICCV’19)	ResNet-18‡	58.67 ± 0.7	75.62 ± 0.5
DeepEMD [37] (CVPR’20)	ResNet-12	75.65 ± 0.8	88.69 ± 0.5
Neg-Margin [126] (ECCV’20)	ResNet-18‡	72.66 ± 0.9	89.40 ± 0.4
MixtFSL [161] (ICCV’21)	ResNet-18‡	73.94 ± 1.1	86.01 ± 0.5
APP2S [321] (AAAI’22)	ResNet-12	77.64 ± 0.1	90.43 ± 0.1
DSKT-1	ResNet-12	76.26 ± 0.8	90.76 ± 0.4
DSKT-2	ResNet-12	78.32 ± 0.8	91.47 ± 0.4

6.4.3 Ablative Analysis

Benefits of multi-task learning. To study the impact of multi-task learning, we evaluate our approach with and without the rotation loss on CUB. As shown in Table 6.4, we first simply train the DSKT-1 with cross entropy loss, which is similar to RFS-simple [110], the model achieves 71.74% and 87.23% in 5-way 1-shot and 5-way 5-shot tasks, respectively. Then, we train the DSKT-1 with additional rotation loss, the model performance improves to 76.26% and 90.76%, which presents an absolute gain of 4.52% and 3.53%. From the results, we can infer the importance of multi-task learning during the training process.

Table 6.4: FSL results on CUB, with different combinations of loss functions. For DSKT-2, the loss functions on the left side of the arrow are employed to train the DSKT-1 model.

Model	Loss function	CUB 5-way	
		1-shot	5-shot
DSKT-1	ℓ_{CE}	71.74	87.23
	$\ell_{CE} + \alpha \ell_{BCE}$	76.26	90.76
DSKT-2	$\ell_{CE} \rightarrow \ell_{KD}$	73.59	88.47
	$\ell_{CE} \rightarrow \lambda_1 \ell_{TCKD} + \lambda_2 \ell_{NCKD}$	74.05	88.67
	$\ell_{CE} \rightarrow \eta_1 \ell_{Poly-1} + \eta_2 (\lambda_1 \ell_{TCKD} + \lambda_2 \ell_{NCKD})$	74.60	89.20
	$\ell_{CE} + \alpha \ell_{BCE} \rightarrow \ell_{KD}$	77.87	91.08
	$\ell_{CE} + \alpha \ell_{BCE} \rightarrow \lambda_1 \ell_{TCKD} + \lambda_2 \ell_{NCKD}$	78.16	91.23
	$\ell_{CE} + \alpha \ell_{BCE} \rightarrow \eta_1 \ell_{Poly-1} + \eta_2 (\lambda_1 \ell_{TCKD} + \lambda_2 \ell_{NCKD})$	78.32	91.47

6. DUAL SELECTIVE KNOWLEDGE TRANSFER FOR FEW-SHOT CLASSIFICATION

Table 6.5: FSL results on CUB, with different processing methods in the first stage.

Processing methods	CUB 5-way	
	1-shot	5-shot
Puzzle	75.75 ± 0.8	89.50 ± 0.5
Fusion	73.93 ± 0.8	89.02 ± 0.5
Rotation	76.26 ± 0.8	90.76 ± 0.4

Table 6.6: FSL results on CUB, with different classifiers during the evaluation process.

Classifiers	CUB 5-way	
	1-shot	5-shot
Nearest classifier	77.92 ± 0.7	88.02 ± 0.5
Cosine classifier	77.67 ± 0.7	88.04 ± 0.5
Logistic Regression	78.32 ± 0.8	91.47 ± 0.4

Benefits of DSKD. To better evaluate the contribution of DSKD, we train the model DSKT-2 with different combinations of loss functions on CUB. From Table 6.4, we can find that, for models trained only with cross entropy loss in the first stage, DSKD gives 1.01% and 0.73% gains compared with KD (KD is defined as $\ell_{KD} = KL(p_i^T p_i^S)$, where p_i denotes the output logits of the i -th sample, and T and S represent the teacher and student model, respectively.) in 1-shot and 5-shot tasks, respectively. For models trained with both cross entropy and binary cross loss functions, DSKD still achieves 0.45% and 0.39% improvements than KD. These results confirm the effectiveness of DSKD.

Benefits of rotation and logistic regression classifier. To study the impact of rotation and logistic regression classifiers, we evaluate our approach with other processing methods and classifiers on CUB. Table 6.5 presents the results on CUB with different processing methods (i.e., puzzle and fusion) in the first stage (DSKT-1). From Table 6.5 we can find that rotation performs much better than fusion (Three channels of RGB images are fused and the coefficient of each channel is set to 0.5 in this work). When compared to puzzle (split the original image into pieces, randomly sort and then reassemble. The puzzle size is set to 7 on CUB 5-way tasks.), rotation still obtains 0.51% and 1.26% improvements on CUB 5-way 1-shot and 5-shot tasks, respectively. Table 6.6 shows the FSL results with different classifiers on CUB during the evaluation process. From Table 6.6, we can find that logistic regression classifier achieves the best performance when compared to Nearest classifier (make predictions based on the euclidean distances between query and support samples) and Cosine classifier (make predictions based on the cosine similarities between query and support samples) on both CUB 5-way 1-shot and 5-shot tasks during the evaluation process.

Table 6.7: FSL results on CUB, with different inputs for the rotation classifier in the first stage.

Input of the rotation classifier	CUB 5-way	
	1-shot	5-shot
The feature extractor’s output	75.37 ± 0.8	90.10 ± 0.4
The category classifier’s output	76.26 ± 0.8	90.76 ± 0.4

Table 6.8: Results of different degrees of the polynomial used in PolyLoss on CUB.

N	CUB 5-way	
	1-shot	5-shot
1	78.32 ± 0.8	91.47 ± 0.4
2	77.57 ± 0.8	91.10 ± 0.4
3	77.42 ± 0.8	91.11 ± 0.4

Table 6.9: Results of different applications of rotation loss on CUB.

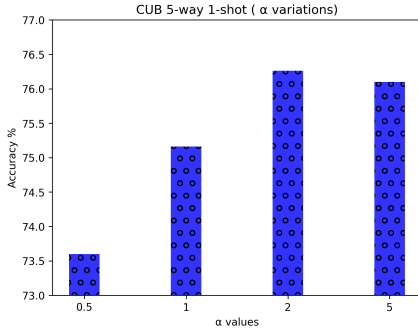
Application of the rotation loss	CUB 5-way	
	1-shot	5-shot
Applied on the features	75.73 ± 0.8	89.84 ± 0.4
Applied on the category logits	76.26 ± 0.8	90.76 ± 0.4

Benefits of using the output of the category classifier as input for the rotation classifier. In this section, we exploit the output of category classifier as input for the rotation classifier, rather than that of the feature extractor. This can help improve the model’s performance on complex data. Table 6.7 present the comparison results on CUB. From Table 6.7, we can find that applying the output of the category classifier gains 0.89% and 0.66% improvements in CUB 5-way 1-shot and 5-shot tasks, respectively.

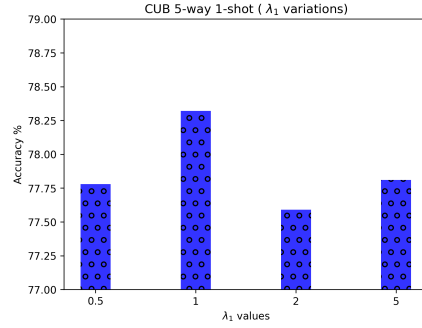
Different degrees of polynomial in PolyLoss. The degree of the polynomial used in PolyLoss N is set to 1 in this work. To study the impact of the degree, we assign the degree with different values in CUB 5-way tasks. From Table 6.8, we can find that when $N = 1$, our method achieves around 0.8% and 0.3% improvements compared to $N = 2$ or $N = 3$ in CUB 5-way 1-shot and 5-shot tasks, respectively.

Application of the rotation loss. In this work, the rotation loss is directly applied to the predicted logits of the category classifier. To learn its contribution, we make a comparison with applying rotation loss on the features in CUB 5-way tasks. Table 6.9 shows that compared with applying multi-task loss on the features, applying on the category logits achieves 0.53% and 0.92% improvements on CUB 5-way 1-shot and 5-shot tasks, respectively.

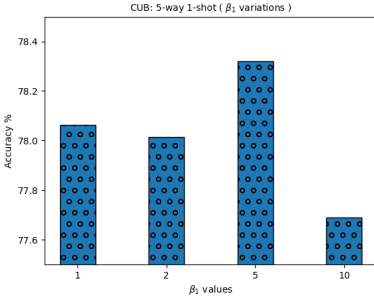
6. DUAL SELECTIVE KNOWLEDGE TRANSFER FOR FEW-SHOT CLASSIFICATION



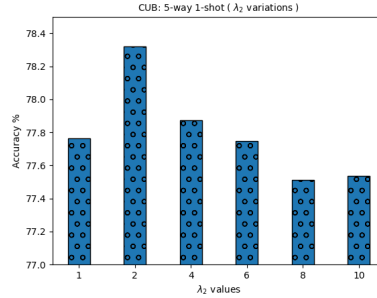
(a) Ablation study on the sensitivity of hyper-parameter α on CUB 5-way 1-shot tasks.



(b) Ablation study on the sensitivity of hyper-parameter λ_1 on CUB 5-way 1-shot tasks.



(c) Ablation study on the sensitivity of hyper-parameter β_1 on CUB 5-way 1-shot tasks.



(d) Ablation study on the sensitivity of hyper-parameter λ_2 on CUB 5-way 1-shot tasks.

Figure 6.2: Ablative Analysis.

Variations of Hyper-parameters: There are totally seven hyper-parameters in this work, t , α , β_1 , λ_1 , λ_2 , η_1 and η_2 . t denotes the temperature coefficient, α controls the contribution of rotation loss during DSKT-1, β_1 is used to adjust the contribution of the first polynomial base, λ_1 , λ_2 , η_1 and η_2 are the weights of different loss functions. We mainly investigate the variants of hyper-parameters α , β_1 , λ_1 and λ_2 , where the default values of t , η_1 and η_2 are 4.0, 0.5 and 0.5, respectively. Figure 6.2a shows the DSKT-1 performance on CUB 5-way 1-shot tasks by changing α . It's obviously that our model performance increases from 0.5 to 2, and then decreases when $\alpha = 5$, which indicates the importance of multi-task learning. Figure 6.2c shows the DSKT-2 performance on CUB 5-way 1-shot tasks by changing β_1 . We observe that DSKT-2 achieves over 78% when β_1 increases from 1 till 5, while $\beta_1 = 10$ performs the lowest. Note that the performance drops only by about 0.6%, which indicates that it is not sensitive to the value of β_1 . Figure 6.2b presents the DSKT-2 performance on CUB 5-way 1-shot tasks by changing λ_1 . When λ_1 ranges from 0.5 till 5, DSKT-2 can always obtain over 77% accuracies and achieves the highest

accuracy when $\lambda_1 = 1.0$, indicating it is not sensitive to the value of λ_1 . Figure 6.2d presents the DSKT-2 performance on the same tasks by changing λ_2 . We find that the performance increases from 1 to 2, then decreases with larger values of λ_2 . These results indicate that it is not a good idea to blindly increase the weights of ℓ_{NCKD} .

6.5 Chapter Conclusions

In this work, we aim to raise awareness of the importance of training a well-generalized feature extractor for FSL tasks by proposing a new two-stage dual selective knowledge transfer framework. First, we use multi-task learning to enforce the feature extractor to learn robust low-level features. Then, we propose an effective dual knowledge distillation method, which enables the student model to selectively learn knowledge from the teacher model and current examples, further improving the model’s generalization ability. Extensive experimental results demonstrate the importance of strong feature extractors for FSL and our approach outperforms the state-of-the-art on four popular FSL benchmark datasets.

Future Work. Despite the promising results in our study, there are some areas that are worth further investigation. We only apply our method on classification tasks in this chapter, object detection tasks will be our future research direction. The correlation between the distillation performance and polynomial function coefficients β_i is not fully investigated, we will expand upon our work in future researches.

Chapter 7

Conclusions

In this thesis, we have devoted previous four research chapters to addressing four research questions regarding the three themes: metric learning, feature enrichment and knowledge transfer. In this chapter, we formulate the main findings from our approaches and results. In addition, we discuss the limitations of our methods and possible solutions to address them. Lastly, we point out several trends for future work.

7.1 Main Findings

In each research chapter, we have proposed a new approach to answer the corresponding research question. Next, we conclude these approaches and present the main findings inspired by empirical and theoretical analyses.

(1) We start the research part in Chapter 3 with proposing a new yet practical prototype similarity network (ProtoSimNet) for few-shot learning (FSL). A key difficulty in FSL is that each new class is represented by only a few labeled examples; as a result, the estimated class prototypes can be noisy and easily perturbed, so even small changes in the support set may noticeably shift prototypes in the embedding space and blur the decision boundary. This makes the choice of metric critical: a metric that is overly sensitive to such shifts can cause unstable predictions across tasks. To alleviate these issues, we propose a refined metric formulation, named cosine sharpening transformation, which focuses comparison on angular (directional) similarity and sharpens the similarity distribution to increase separation between classes in the decision space. Intuitively, this makes classification depend less on accidental variations in feature magnitude and more on consistent directional structure, which helps preserve clearer margins when prototypes are estimated from very limited data. On top of this, we introduce an elastic regularization term that constrains the deformation of the embedding space, preventing unstable shifts in feature geometry while still keeping sufficient flexibility for adapting to new categories.

(2) After investigating the metric learning theme, we now move our focus to address the feature enrichment theme in Chapter 4. In few-shot settings, the model must generalize from very limited supervision, so relying on a single feature level can be fragile. To more effectively capture the structural diversity required for robust few-shot recognition, we introduce an attention-based feature pyramid network (AFPN), which enriches the model’s representational capability by integrating multi-level features extracted from different layers of the backbone. AFPN uses an adaptive top-down attention mechanism to select and fuse informative spatial scales in a task-aware manner, allowing the model to combine local details with higher-level semantics. Experimental results across standard few-shot benchmarks demonstrate that the proposed AFPN significantly improves the discriminability and stability of features under limited supervision.

(3) In Chapter 5 we further explore the efficient strategies for enriching visual representations in FSL scenarios. Traditional meta-learning pipelines often aggregate features in a fixed way, but under few-shot supervision a fixed fusion strategy can be suboptimal because different tasks may require different spatial resolutions. Beyond traditional meta-learning approaches that often rely on a fixed feature aggregation strategy, we propose a novel adaptive multi-level feature fusion (AMFF) framework to strengthen representation diversity. AMFF dynamically integrates and recalibrates feature representations across multiple spatial resolutions, enriching the feature distribution while preserving the flexibility needed for adapting to new categories with only a few labeled samples. We believe that the AMFF can be extended to a wide range of representation-oriented tasks, such as fine-grained recognition, cross-domain FSL, and lightweight deep learning.

(4) After investigating the themes of metric learning and feature enrichment, we then turn to address the questions about the knowledge transfer theme for FSL. A central challenge in few-shot learning is that novel classes provide too little data to learn a strong representation from scratch; therefore, performance often depends on how effectively prior knowledge can be transferred. However, transferring everything indiscriminately can be harmful when the transferred information is not equally relevant to the new task. In Chapter 6, we introduce a dual selective knowledge transfer (DSKT) framework that transfers rich, efficient knowledge to guide the learning of novel categories. By emphasizing task-relevant knowledge while suppressing less useful ones, DSKT produces more effective distilled representations, yielding significant improvements over existing approaches and establishing new competitive baselines for FSL.

7.2 Limitations

Our methods in this thesis have addressed four research questions and achieved promising results in terms of three important research themes. However, they still

have some limitations which can be discussed from the following three perspectives.

7.2.1 Algorithmic perspective

In Chapter 3, ProtoSimNet integrates cosine sharpening transformation and elastic regularization to effectively strengthen inter-class prototype distinction. However, its effectiveness depends on the stable estimation of the prototype variance, which is inherently fragile in FSL settings. In addition, the fixed sharpen factor may suppress the model’s adaptability across diverse few-shot scenarios. A potential direction is to develop an adaptive sharpening mechanism that dynamically adjusts the transformation parameters based on episode difficulty.

In Chapter 4, AFPN demonstrates its improvements in recognition performance by integrating multi-level features with attention-guided fusion. Yet the top-down attention propagation increases computational cost as the pyramid depth and feature resolution grow. One possible optimization is to implement hierarchical attention mechanisms that progressively refine the features fused with reduced computational cost.

In Chapter 5, AMFF improves feature expressiveness by enriching representations through parallel convolutional branches. However, its multi-branch convolutions inevitably increase the parameter budget. Potential improvements include investigating dynamic neural architectures that adaptively activate different branches based on input complexity, or developing knowledge distillation techniques to compress the multi-branch knowledge into more compact representations.

In Chapter 6, DSKT facilitates knowledge transfer by selectively extracting efficient feature representations while simultaneously filtering out task-specific noise that may influence generalization. However, this selectivity requires the student to compare its representations with those of the teacher at multiple points, which increases the overall computational cost. A potential solution for future research is to develop in-place selective distillation, enabling the student to obtain relevant information without reconstructing every component of the teacher model.

7.2.2 Practical perspective

In Chapter 3, ProtoSimNet demonstrates its effectiveness in tightening class boundaries, yet it is sensitive to highly distorted feature distributions —as commonly occurs in long-tailed or domain-shifted datasets. A promising direction is to incorporate uncertainty-aware sharpening, where cosine sharpening transformation is modulated by confidence estimates or Bayesian feature variance to avoid excessive margin tightening.

In Chapter 4, AFPN enhances feature representation through multi-scale feature pyramids, but storing all scale outputs introduces a high demand for memory space. This makes it difficult to deploy AFPN on real-time applications where computational and memory resources are limited. A potential solution is exploring a more lightweight and efficient pyramid designs to maintain multi-scale benefits while fitting to practical constraints.

In Chapter 5, AMFF enriches feature diversity through multi-branch convolutions, but the additional branches introduce noticeable extra computational cost when applied to deeper backbones. One practical alternative is to reduce redundant branches or employing dynamic fusion mechanisms.

In Chapter 6, DSKT transfers selective teacher knowledge to improve generalization, but comparing student and teacher representations still requires excessive computational cost. A potential option is to distill compressed teacher features—such as low-rank projections or compact codes—allowing the student to rely on lightweight embedding descriptors rather than full feature matching.

7.3 Future Research Directions

In the previous four chapters, we have introduced several methods addressing the key research questions within the three major themes. These explorations also provide a range of promising directions for further advancement. In this section, we briefly discuss future research directions regarding each theme.

7.3.1 Metric Learning

Metric learning is a practical solution for FSL and has obtained noticeable performance in past years. However, most existing metrics are fixed once trained, making it difficult for the model to adapt to new tasks across episodes. New works, such as adaptive prototype networks [327] and prototype-rectification strategies [328], show that episode-dependent or distribution-aware metrics can significantly improve robustness. Therefore, a practical direction for future research is to design adaptive metric formulations that adjust distance functions based on support-set structure, or cross-domain statistics, rather than using a single global metric.

7.3.2 Feature Enrichment

Recent work has shown that feature enrichment approaches can significantly improve FSL performance by mitigating the lack of feature diversity. However, most existing enrichment methods require increased memory usage. A realistic challenge is to design enrichment modules that can enhance representation quality without requiring additional backbone stages or heavy multi-branch operations. A potential solution is to explore implicit feature enrichment, where fine-grained representations

are strengthened without explicitly expanding network width or depth. Recent studies demonstrate that lightweight modulation layers can effectively reshape feature distributions for few-shot generalization [329], cross-layer aggregation enhances hierarchical consistency across backbone stages [294], and self-augmenting features generated directly in latent space provide a parameter-efficient way to boost discriminative capability [330]. These insights indicate that implicit enrichment can obtain significant improvements in representation quality while keeping lower computational costs and memory usage.

7.3.3 Knowledge Transfer

Knowledge transfer for FSL has recently evolved from simple pretraining and fine-tuning to more sophisticated distillation-based frameworks, including dual selective distillation and multi-teacher knowledge transfer [331]. There is also growing interest in combining FSL with knowledge distillation in NLP and multi-modal settings, for instance, distilling few-shot capabilities from large language models into smaller learners [332]. Yet, a critical challenge arises in resource-constrained environments, where storing the full teacher model is impractical. Thus, it will be promising to explore compressed teacher representations or episodic subspace transfer, enabling stable knowledge transfer without excessive memory or computation.

Bibliography

- [1] Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. (2017) 4077–4087
- [2] Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. Volume 70. (2017) 1126–1135
- [3] Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. *Science* **350** (2015) 1332–1338
- [4] Shyam, P., Gupta, S., Dukkipati, A.: Attentive recurrent comparators. In: International Conference on Machine Learning. Volume 70. (2017) 3173–3181
- [5] Dumoulin, V., Hounsby, N., Evci, U., Zhai, X., Goroshin, R., Gelly, S., Larochelle, H.: A unified few-shot classification benchmark to compare transfer and meta learning approaches. In: Advances in Neural Information Processing Systems. (2021)
- [6] Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* **55** (2023) 271:1–271:40
- [7] Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.* **11** (2010) 1109–1135
- [8] Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Advances in Neural Information Processing Systems. (2016) 1849–1857
- [9] Song, H.O., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2016) 4004–4012
- [10] Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2017) 3420–3429
- [11] Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017)
- [12] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. (2009)
- [13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115** (2015) 211–252
- [14] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems. (2016) 3630–3638
- [15] Ochal, M., Patacchiola, M., Vazquez, J., Storkey, A., Wang, S.: Few-shot learning with class imbalance. *IEEE Trans. Artif. Intell.* **4** (2023) 1348–1358
- [16] Polak, P.B., Prusa, J.D., Khoshgftaar, T.M.: Low-shot learning and class imbalance: a survey. *J. Big Data* **11** (2024) 1
- [17] Zhang, H., Chen, S., Luo, L., Yang, J.: Few-shot learning with long-tailed labels. *Pattern*

- Recognit. **156** (2024) 110806
- [18] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. Volume 119. (2020) 1597–1607
 - [19] Lu, Y., Wen, L., Liu, J., Liu, Y., Tian, X.: Self-supervision can be a good few-shot learner. In: European Conference on Computer Vision. Volume 13679. (2022) 740–758
 - [20] Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations. (2018)
 - [21] Tseng, H., Lee, H., Huang, J., Yang, M.: Cross-domain few-shot classification via learned feature-wise transformation. In: International Conference on Learning Representations. (2020)
 - [22] Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Cross attention network for few-shot classification. In: Advances in Neural Information Processing Systems. (2019) 4005–4016
 - [23] Oreshkin, B.N., López, P.R., Lacoste, A.: TADAM: task dependent adaptive metric for improved few-shot learning. In: Advances in Neural Information Processing Systems. (2018) 719–729
 - [24] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 1199–1208
 - [25] Antoniou, A., Storkey, A.J., Edwards, H.: Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2017)
 - [26] Li, K., Zhang, Y., Li, K., Fu, Y.: Adversarial feature hallucination networks for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 13467–13476
 - [27] Wang, H., Tian, S., Fu, Y., Zhou, J., Liu, J., Chen, D.: Feature augmentation based on information fusion rectification for few-shot image classification. Scientific Reports **13** (2023) 3607
 - [28] Antoniou, A., Edwards, H., Storkey, A.J.: How to train your MAML. In: International Conference on Learning Representations. (2019)
 - [29] Jamal, M.A., Qi, G.: Task agnostic meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 11719–11727
 - [30] Paeddeh, N., Pratama, M., Ma’sum, M.A., Mayer, W., Cao, Z., Kowalczyk, R.: Cross-domain few-shot learning via adaptive transformer networks. Knowl. Based Syst. **288** (2024) 111458
 - [31] Cheraghian, A., Rahman, S., Fang, P., Roy, S.K., Petersson, L., Harandi, M.: Semantic-aware knowledge distillation for few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2534–2543
 - [32] Chen, R., Chen, T., Hui, X., Wu, H., Li, G., Lin, L.: Knowledge graph transfer network for few-shot recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2020) 10575–10582
 - [33] He, K., Pu, N., Lao, M., Lew, M.S.: Few-shot and meta-learning methods for image understanding: a survey. Int. J. Multim. Inf. Retr. **12** (2023) 14
 - [34] He, K., Bakker, E., Lew, M.S.: Attention-guided feature pyramid network for few-shot learning. In: Benelux Conference on Artificial Intelligence and Machine Learning. (2024)
 - [35] He, K., Bakker, E.M., Lew, M.S.: Adaptive multi-scale feature fusion network for few-shot learning. In: International Conference on Intelligent Computing. (2025) 472–482
 - [36] He, K., Pu, N., Lao, M., Bakker, E.M., Lew, M.S.: Dual selective knowledge transfer for few-shot classification. Appl. Intell. **53** (2023) 27779–27789
 - [37] Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers. In: Proceedings of the IEEE/CVF

- Conference on Computer Vision and Pattern Recognition. (2020) 12200–12210
- [38] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60** (2017) 84–90
 - [39] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*. (2015)
 - [40] Akata, Z., Geiger, A., Sattler, T.: Computer vision and pattern recognition 2020. *Int. J. Comput. Vis.* **129** (2021) 3169–3170
 - [41] Voulodimos, A., Doulamis, N., Doulamis, A.D., Protopapadakis, E.: Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018** (2018) 7068349:1–7068349:13
 - [42] Mahesh, B.: Machine learning algorithms-a review. *International Journal of Science and Research* **9** (2020) 381–386
 - [43] Zhang, Z., Sejdic, E.: Radiological images and machine learning: Trends, perspectives, and prospects. *Comput. Biol. Medicine* **108** (2019) 354–370
 - [44] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A.Q., Duan, Y., Al-Shamma, O., Santamaria, J., Fadhel, M.A., Al-Amidie, M., Farhan, L.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **8** (2021) 53
 - [45] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. *Neurocomputing* **234** (2017) 11–26
 - [46] Sarker, I.H.: Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput. Sci.* **2** (2021) 420
 - [47] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. *Pattern Recognit.* **77** (2018) 354–377
 - [48] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2015) 1–9
 - [49] Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2009) 248–255
 - [50] Shahroudy, A., Liu, J., Ng, T., Wang, G.: NTU RGB+D: A large scale dataset for 3d human activity analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2016) 1010–1019
 - [51] Liu, B., Guo, W., Chen, X., Gao, K., Zuo, X., Wang, R., Yu, A.: Morphological attribute profile cube and deep random forest for small sample classification of hyperspectral image. *IEEE Access* **8** (2020) 117096–117108
 - [52] Shu, J., Xu, Z., Meng, D.: Small sample learning in big data era. *arXiv preprint arXiv:1808.04572* (2018)
 - [53] Zhu, F., Ma, Z., Li, X., Chen, G., Chien, J., Xue, J., Guo, J.: Image-text dual neural network with decision strategy for small-sample image classification. *Neurocomputing* **328** (2019) 182–188
 - [54] Afrasiyabi, A., Lalonde, J., Gagné, C.: Associative alignment for few-shot image classification. In: *European Conference on Computer Vision*. Volume 12350. (2020) 18–35
 - [55] Chen, W., Liu, Y., Kira, Z., Wang, Y.F., Huang, J.: A closer look at few-shot classification. In: *International Conference on Learning Representations*. (2019)
 - [56] Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2018) 7229–7238
 - [57] Dong, J., Wang, Y., Lai, J., Xie, X.: Improving adversarially robust few-shot image classification with generalizable representations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 9015–9024

- [58] Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 9367–9376
- [59] Wang, R., Zhu, F., Zhang, X., Liu, C.: Training with scaled logits to alleviate class-level over-fitting in few-shot learning. *Neurocomputing* **522** (2023) 142–151
- [60] Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: *Advances in Neural Information Processing Systems*. (2013) 46–54
- [61] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. *Proc. IEEE* **109** (2021) 43–76
- [62] Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: *International Conference on Learning Representations*. (2018)
- [63] Antoniou, A., Storkey, A.J.: Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation. *arXiv preprint arXiv:1902.09884* (2019)
- [64] Chao, X., Zhang, L.: Few-shot imbalanced classification based on data augmentation. *Multim. Syst.* **29** (2023) 2843–2851
- [65] Hu, T., Tang, T., Lin, R., Chen, M., Han, S., Wu, J.: A simple data augmentation algorithm and a self-adaptive convolutional architecture for few-shot fault diagnosis under different working conditions. *Measurement* **156** (2020) 107539
- [66] Cai, J., Shen, S.M.: Cross-domain few-shot learning with meta fine-tuning. *arXiv preprint arXiv:2005.10544* (2020)
- [67] Shen, Z., Liu, Z., Qin, J., Savvides, M., Cheng, K.: Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. (2021) 9594–9602
- [68] Zhang, J., Bui, T., Yoon, S., Chen, X., Liu, Z., Xia, C., Tran, Q.H., Chang, W., Yu, P.S.: Few-shot intent detection via contrastive pre-training and fine-tuning. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. (2021) 1906–1912
- [69] Elsken, T., Staffler, B., Metzen, J.H., Hutter, F.: Meta-learning of neural architectures for few-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020) 12362–12372
- [70] Goldblum, M., Fowl, L., Goldstein, T.: Adversarially robust few-shot learning: A meta-learning approach. In: *Advances in Neural Information Processing Systems*. (2020)
- [71] Wang, D., Cheng, Y., Yu, M., Guo, X., Zhang, T.: A hybrid approach with optimization-based and metric-based meta-learner for few-shot learning. *Neurocomputing* **349** (2019) 202–211
- [72] Widhianingsih, T.D.A., Kang, D.: Augmented domain agreement for adaptable meta-learner on few-shot classification. *Appl. Intell.* **52** (2022) 7037–7053
- [73] Qiao, L., Shi, Y., Li, J., Tian, Y., Huang, T., Wang, Y.: Transductive episodic-wise adaptive metric for few-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 3602–3611
- [74] Wang, K., Liu, X., Bagdanov, A., Herranz, L., Jui, S., van de Weijer, J.: Incremental meta-learning via episodic replay distillation for few-shot image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 3728–3738
- [75] Zhu, P., Zhu, Z., Wang, Y., Zhang, J., Zhao, S.: Multi-granularity episodic contrastive learning for few-shot learning. *Pattern Recognit.* **131** (2022) 108820
- [76] Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. In: *Advances in Neural Information Processing Systems*. (2015) 2539–2547
- [77] Ratner, A.J., Ehrenberg, H.R., Hussain, Z., Dunnmon, J., Ré, C.: Learning to compose domain-specific transformations for data augmentation. In: *Advances in Neural Information Processing Systems*. (2017) 3236–3246

- [78] Wang, J., Perez, L., et al.: The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit* **11** (2017) 1–8
- [79] Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F.: Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610* (2022)
- [80] dos Santos, F.P., Thumé, G.S., Ponti, M.A.: Data augmentation guidelines for cross-dataset transfer learning and pseudo labeling. In: *34th SIBGRAPI Conference on Graphics, Patterns and Images*. (2021) 207–214
- [81] Do, J., Yoo, M., Kim, S.: A semi-supervised sar image classification with data augmentation and pseudo labeling. In: *2022 IEEE International Conference on Consumer Electronics-Asia*. (2022) 1–4
- [82] Yu, J., Zhang, L., Du, S., Chang, H., Lu, K., Zhang, Z., Yu, Y., Wang, L., Ling, Q.: Pseudo-label generation and various data augmentation for semi-supervised hyperspectral object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 304–311
- [83] Khodadadeh, S., Bölöni, L., Shah, M.: Unsupervised meta-learning for few-shot image classification. In: *Advances in Neural Information Processing Systems*. (2019) 10132–10142
- [84] Singh, R., Bharti, V., Purohit, V., Kumar, A., Singh, A.K., Singh, S.K.: Metamed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognit.* **120** (2021) 108111
- [85] Gaikwad, M., Doke, A.: Survey on meta learning algorithms for few shot learning. In: *2022 6th International Conference on Intelligent Computing and Control Systems*. (2022) 1876–1879
- [86] Huang, W., He, M., Wang, Y.: A survey on meta-learning based few-shot classification. In: *Machine Learning and Intelligent Communications*. Volume 438. (2021) 243–253
- [87] Luo, S., Li, Y., Gao, P., Wang, Y., Serikawa, S.: Meta-seg: A survey of meta-learning for image segmentation. *Pattern Recognit.* **126** (2022) 108586
- [88] Wang, S., Wang, D., Kong, D., Wang, J., Li, W., Zhou, S.: Few-shot rolling bearing fault diagnosis with metric-based meta learning. *Sensors* **20** (2020) 6437
- [89] Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: *International Conference on Machine Learning*. Volume 37. (2015) 2152–2161
- [90] Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2018) 5542–5551
- [91] Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning - the good, the bad and the ugly. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2017) 3077–3086
- [92] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2015) 2425–2433
- [93] Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: *Advances in Neural Information Processing Systems*. (2016) 289–297
- [94] Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2016) 4613–4621
- [95] Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 8419–8428
- [96] Pérez-Rúa, J., Zhu, X., Hospedales, T.M., Xiang, T.: Incremental few-shot object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020) 13843–13852
- [97] Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: FSCE: few-shot object detection via contrastive

- proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 7352–7362
- [98] Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **53** (2021) 63:1–63:34
- [99] Li, X., Sun, Z., Xue, J., Ma, Z.: A concise review of recent few-shot meta-learning methods. *Neurocomputing* **456** (2021) 463–468
- [100] Finn, C., Xu, K., Levine, S.: Probabilistic model-agnostic meta-learning. In: Advances in Neural Information Processing Systems. (2018) 9537–9548
- [101] Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 4367–4375
- [102] Gidaris, S., Komodakis, N.: Generating classification weights with GNN denoising autoencoders for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 21–30
- [103] Tokmakov, P., Wang, Y., Hebert, M.: Learning compositional representations for few-shot recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 6371–6380
- [104] Li, X., Yang, X., Ma, Z., Xue, J.: Deep metric learning for few-shot image classification: A review of recent developments. *Pattern Recognit.* **138** (2023) 109381
- [105] Nguyen, V.N., Løkse, S., Wickstrøm, K., Kampffmeyer, M., Roverso, D., Jenssen, R.: SEN: A novel feature normalization dissimilarity measure for prototypical few-shot learning networks. In: European Conference on Computer Vision. Volume 12368. (2020) 118–134
- [106] Simon, C., Koniusz, P., Nock, R., Harandi, M.: Adaptive subspaces for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 4135–4144
- [107] Yoon, S.W., Seo, J., Moon, J.: Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In: International Conference on Machine Learning. Volume 97. (2019) 7115–7123
- [108] Deng, S., Liao, D., Gao, X., Zhao, J., Ye, K.: Improving few-shot image classification with self-supervised learning. In: Cloud Computing. Volume 13731. (2022) 54–68
- [109] Hou, M., Sato, I.: A closer look at prototype classifier for few-shot image classification. *Advances in Neural Information Processing Systems* **35** (2022) 25767–25778
- [110] Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P.: Rethinking few-shot image classification: A good embedding is all you need? In: European Conference on Computer Vision. Volume 12359. (2020) 266–282
- [111] Lake, B.M., Salakhutdinov, R., Gross, J., Tenenbaum, J.B.: One shot learning of simple visual concepts. In: Proceedings of the 33th Annual Meeting of the Cognitive Science Society. (2011)
- [112] Bertinetto, L., Henriques, J.F., Torr, P.H.S., Vedaldi, A.: Meta-learning with differentiable closed-form solvers. In: International Conference on Learning Representations. (2019)
- [113] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. (2011)
- [114] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200. (2010)
- [115] Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations. (2017)
- [116] Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 7260–7268
- [117] Mangla, P., Singh, M., Sinha, A., Kumari, N., Balasubramanian, V.N., Krishnamurthy,

- B.: Charting the right manifold: Manifold mixup for few-shot learning. In: IEEE Winter Conference on Applications of Computer Vision. (2020) 2207–2216
- [118] Wang, H., Deng, Z.: Cross-domain few-shot classification via adversarial task augmentation. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. (2021) 1075–1081
- [119] Yue, Z., Zhang, H., Sun, Q., Hua, X.S.: Interventional few-shot learning. *Advances in Neural Information Processing Systems* **33** (2020) 2734–2746
- [120] Dumoulin, V., Hounsby, N., Evci, U., Zhai, X., Goroshin, R., Gelly, S., Larochelle, H.: Comparing transfer and meta learning approaches on a unified few-shot classification benchmark. *arXiv preprint arXiv:2104.02638* (2021)
- [121] Sun, X., Xv, H., Dong, J., Zhou, H., Chen, C., Li, Q.: Few-shot learning for domain-specific fine-grained image classification. *IEEE Trans. Ind. Electron.* **68** (2021) 3588–3598
- [122] Wertheimer, D., Tang, L., Hariharan, B.: Few-shot classification with feature map reconstruction networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 8012–8021
- [123] Zheng, W., Tian, X., Yang, B., Liu, S., Ding, Y., Tian, J., Yin, L.: A few shot classification methods based on multiscale relational networks. *Applied Sciences* **12** (2022) 4059
- [124] Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. In: International Conference on Learning Representations. (2020)
- [125] Kang, D., Kwon, H., Min, J., Cho, M.: Relational embedding for few-shot classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 8802–8813
- [126] Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., Hu, H.: Negative margin matters: Understanding margin in few-shot classification. In: European Conference on Computer Vision. Volume 12349. (2020) 438–455
- [127] Ma, J., Xie, H., Han, G., Chang, S., Galstyan, A., Abd-Almageed, W.: Partner-assisted learning for few-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 10553–10562
- [128] Parnami, A., Lee, M.: Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291* (2022)
- [129] Guo, Y., Codella, N., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: European Conference on Computer Vision. Volume 12372. (2020) 124–141
- [130] Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. In: International Conference on Learning Representations. (2021)
- [131] Ziko, I.M., Dolz, J., Granger, E., Ayed, I.B.: Laplacian regularized few-shot learning. In: International Conference on Machine Learning. Volume 119. (2020) 11660–11670
- [132] Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: Exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 9042–9051
- [133] Najdenkoska, I., Zhen, X., Worring, M.: Meta learning to bridge vision and language models for multimodal few-shot learning. In: International Conference on Learning Representations. (2023)
- [134] Yang, P., Ren, S., Zhao, Y., Li, P.: Calibrating cnns for few-shot meta learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision. (2022) 408–417
- [135] Guo, N., Di, K., Liu, H., Wang, Y., Qiao, J.: A metric-based meta-learning approach combined attention mechanism and ensemble learning for few-shot learning. *Displays* **70** (2021) 102065
- [136] Li, X., Yu, L., Fu, C., Fang, M., Heng, P.: Revisiting metric learning for few-shot image classification. *Neurocomputing* **406** (2020) 49–58

- [137] Li, P., Zhao, G., Xu, X.: Coarse-to-fine few-shot classification with deep metric learning. *Inf. Sci.* **610** (2022) 592–604
- [138] Zhang, P., Bai, Y., Wang, D., Bai, B., Li, Y.: Few-shot classification of aerial scene images via meta-learning. *Remote. Sens.* **13** (2021) 108
- [139] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a siamese time delay neural network. In: *Advances in Neural Information Processing Systems*. (1993) 737–744
- [140] Baldi, P., Chauvin, Y.: Neural networks for fingerprint recognition. *Neural Comput.* **5** (1993) 402–418
- [141] Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: *ICML deep learning workshop*. Volume 2. (2015) 1–30
- [142] Wang, J., Zhu, Z., Li, J., Li, J.: Attention based siamese networks for few-shot learning. In: *2018 IEEE 9th International Conference on Software Engineering and Service Science*. (2018) 551–554
- [143] Chen, Z., Fu, Y., Wang, Y., Ma, L., Liu, W., Hebert, M.: Image deformation meta-networks for one-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2019) 8680–8689
- [144] Eloff, R., Engelbrecht, H.A., Kamper, H.: Multimodal one-shot learning of speech and images. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. (2019) 8623–8627
- [145] Tsutsui, S., Fu, Y., Crandall, D.J.: Meta-reinforced synthetic data for one-shot fine-grained visual recognition. In: *Advances in Neural Information Processing Systems*. (2019) 3057–3066
- [146] Lungu, I., Hu, Y., Liu, S.: Multi-resolution siamese networks for one-shot learning. In: *2nd IEEE International Conference on Artificial Intelligence Circuits and Systems*. (2020) 183–187
- [147] Zhang, S., Zheng, D., Hu, X., Yang, M.: Bidirectional long short-term memory networks for relation classification. In: *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. (2015)
- [148] Brauwuers, G., Frasincar, F.: A general survey on attention mechanisms in deep learning. *IEEE Trans. Knowl. Data Eng.* **35** (2023) 3279–3298
- [149] Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. *Neuro-computing* **452** (2021) 48–62
- [150] Mai, S., Hu, H., Xu, J.: Attentive matching network for few-shot learning. *Comput. Vis. Image Underst.* **187** (2019) 102781
- [151] Li, W., Xu, J., Huo, J., Wang, L., Gao, Y., Luo, J.: Distribution consistency based covariance metric networks for few-shot learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. (2019) 8642–8649
- [152] Wang, J., Zhai, Y.: Prototypical siamese networks for few-shot learning. In: *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, IEEE (2020) 178–181
- [153] Fan, M., Bai, Y., Sun, M., Li, P.: Large margin prototypical network for few-shot relation classification with fine-grained features. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. (2019) 2353–2356
- [154] Huang, H., Zhang, J., Zhang, J., Wu, Q., Xu, J.: Compare more nuanced: Pairwise alignment bilinear network for few-shot fine-grained learning. In: *IEEE International Conference on Multimedia and Expo*. (2019) 91–96
- [155] Zhu, Y., Liu, C., Jiang, S.: Multi-attention meta learning for few-shot fine-grained image recognition. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. (2020) 1090–1096
- [156] Li, X., Wu, J., Sun, Z., Ma, Z., Cao, J., Xue, J.: Bsnet: Bi-similarity network for few-shot

- fine-grained image classification. *IEEE Trans. Image Process.* **30** (2021) 1318–1331
- [157] Pele, O., Werman, M.: Fast and robust earth mover’s distances. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* (2009) 460–467
 - [158] Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover’s distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40** (2000) 99–121
 - [159] Yu, Z., Herman, G.: On the earth mover’s distance as a histogram similarity metric for image retrieval. In: *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo.* (2005) 686–689
 - [160] Xie, J., Long, F., Lv, J., Wang, Q., Li, P.: Joint distribution matters: Deep brownian distance covariance for few-shot classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* (2022) 7962–7971
 - [161] Afrasiyabi, A., Lalonde, J., Gagné, C.: Mixture-based feature space learning for few-shot image classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* (2021) 9021–9031
 - [162] Afrasiyabi, A., Larochelle, H., Lalonde, J., Gagné, C.: Matching feature sets for few-shot image classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* (2022) 9004–9014
 - [163] Karunaratne, G., Schmuck, M., Le Gallo, M., Cherubini, G., Benini, L., Sebastian, A., Rahimi, A.: Robust high-dimensional memory-augmented neural networks. *Nature communications* **12** (2021) 2468
 - [164] Nichol, A., Schulman, J.: Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* **2** (2018) 4
 - [165] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86** (1998) 2278–2324
 - [166] Rumelhart, D.E., McClelland, J.L.: On learning the past tenses of english verbs. Technical report (1985)
 - [167] Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45** (1997) 2673–2681
 - [168] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9** (1997) 1735–1780
 - [169] Santoro, A., Bartunov, S., Botvinick, M.M., Wierstra, D., Lillicrap, T.P.: Meta-learning with memory-augmented neural networks. In: *International Conference on Machine Learning.* Volume 48. (2016) 1842–1850
 - [170] Collier, M., Beel, J.: Implementing neural turing machines. In: *Artificial Neural Networks and Machine Learning.* Volume 11141. (2018) 94–104
 - [171] Faradonbe, S.M., Safi-Esfahani, F., Karimian-kelishadrokh, M.: A review on neural turing machine (NTM). *SN Comput. Sci.* **1** (2020) 333
 - [172] Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. *arXiv preprint arXiv:1410.5401* (2014)
 - [173] Tran, K., Sato, H., Kubo, M.: Memory augmented matching networks for few-shot learnings. *International Journal of Machine Learning and Computing* **9** (2019)
 - [174] Cai, Q., Pan, Y., Yao, T., Yan, C., Mei, T.: Memory matching networks for one-shot image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* (2018) 4080–4088
 - [175] Miller, A.H., Fisch, A., Dodge, J., Karimi, A., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* (2016) 1400–1409
 - [176] Munkhdalai, T., Yu, H.: Meta networks. In: *International Conference on Machine Learning.* Volume 70. (2017) 2554–2563
 - [177] Munkhdalai, T., Yuan, X., Mehri, S., Trischler, A.: Rapid adaptation with conditionally

- shifted neurons. In: International Conference on Machine Learning. Volume 80. (2018) 3661–3670
- [178] Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: Meta-learning with temporal convolutions. arXiv preprint arXiv:1707.03141 (2017)
- [179] Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: A simple neural attentive meta-learner. In: International Conference on Learning Representations. (2018)
- [180] Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y.W., Rezende, D.J., Eslami, S.M.A.: Conditional neural processes. In: International Conference on Machine Learning. Volume 80. (2018) 1690–1699
- [181] Bian, W., Chen, Y., Ye, X., Zhang, Q.: An optimization-based meta-learning model for MRI reconstruction with diverse dataset. *J. Imaging* **7** (2021) 231
- [182] Cho, H., Cho, Y., Yu, J., Kim, J.: Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 11149–11158
- [183] Feurer, M., Springenberg, J.T., Hutter, F.: Initializing bayesian hyperparameter optimization via meta-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2015) 1128–1135
- [184] Gao, K., Sener, O.: Modeling and optimization trade-off in meta-learning. In: Advances in Neural Information Processing Systems. Volume 33. (2020) 11154–11165
- [185] Reif, M., Shafait, F., Dengel, A.: Meta-learning for evolutionary parameter optimization of classifiers. *Mach. Learn.* **87** (2012) 357–380
- [186] Satrya, W.F., Yun, J.: Combining model-agnostic meta-learning and transfer learning for regression. *Sensors* **23** (2023) 583
- [187] Sendera, M., Tabor, J., Nowak, A., Bedychaj, A., Patacchiola, M., Trzcinski, T., Spurek, P., Zieba, M.: Non-gaussian gaussian processes for few-shot regression. In: Advances in Neural Information Processing Systems. (2021) 10285–10298
- [188] Zhao, C., Chen, F.: Unfairness discovery and prevention for few-shot regression. In: 2020 IEEE International Conference on Knowledge Graph. (2020) 137–144
- [189] Fallah, A., Mokhtari, A., Ozdaglar, A.E.: Provably convergent policy gradient methods for model-agnostic meta-reinforcement learning. arXiv preprint arXiv:2002.05135 (2020)
- [190] Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., Levine, S.: Meta-reinforcement learning of structured exploration strategies. In: Advances in Neural Information Processing Systems. (2018) 5307–5316
- [191] Liu, H., Socher, R., Xiong, C.: Taming MAML: efficient unbiased meta-reinforcement learning. In: International Conference on Machine Learning. Volume 97. (2019) 4061–4071
- [192] Ding, L., Liu, P., Shen, W., Lu, W., Chen, S.: Gradient-based meta-learning using uncertainty to weigh loss for few-shot learning. arXiv preprint arXiv:2208.08135 (2022)
- [193] Yap, P.C., Ritter, H., Barber, D.: Addressing catastrophic forgetting in few-shot problems. In: International Conference on Machine Learning. Volume 139. (2021) 11909–11919
- [194] Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P., Larochelle, H.: Meta-dataset: A dataset of datasets for learning to learn from few examples. In: International Conference on Learning Representations. (2020)
- [195] Zintgraf, L.M., Shiarlis, K., Kurin, V., Hofmann, K., Whiteson, S.: Fast context adaptation via meta-learning. In: International Conference on Machine Learning. Volume 97. (2019) 7693–7702
- [196] Baik, S., Choi, J., Kim, H., Cho, D., Min, J., Lee, K.M.: Meta-learning with task-adaptive loss function for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 9445–9454
- [197] Wang, X., Yu, F., Wang, R., Darrell, T., Gonzalez, J.E.: Tafe-net: Task-aware feature em-

- beddings for low shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1831–1840
- [198] Sun, Q., Liu, Y., Chua, T., Schiele, B.: Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 403–412
- [199] Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. In: International Conference on Learning Representations. (2019)
- [200] Gupta, A., Thadani, K., O’Hare, N.: Effective few-shot classification with transfer learning. In: Proceedings of the 28th International Conference on Computational Linguistics. (2020) 1061–1066
- [201] Rostami, M., Kolouri, S., Eaton, E., Kim, K.: Deep transfer learning for few-shot SAR image classification. *Remote. Sens.* **11** (2019) 1374
- [202] Rostami, M., Kolouri, S., Eaton, E., Kim, K.: SAR image classification using few-shot cross-domain transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 907–915
- [203] Yazdanpanah, M., Rahman, A.A., Chaudhary, M., Desrosiers, C., Havaei, M., Belilovsky, E., Kahou, S.E.: Revisiting learnable affines for batch norm in few-shot transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 9099–9108
- [204] Yu, Z., Chen, L., Cheng, Z., Luo, J.: Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 12853–12861
- [205] Qu, Y., Baghbaderani, R.K., Qi, H.: Few-shot hyperspectral image classification through multitask transfer learning. In: 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing. (2019) 1–5
- [206] Tai, Y., Tan, Y., Xiong, S., Sun, Z., Tian, J.: Few-shot transfer learning for SAR image classification without extra SAR samples. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **15** (2022) 2240–2253
- [207] Sun, N., Yang, P.: T2L: trans-transfer learning for few-shot fine-grained visual categorization with extended adaptation. *Knowl. Based Syst.* **264** (2023) 110329
- [208] Liu, W., Chang, X., Yan, Y., Yang, Y., Hauptmann, A.G.: Few-shot text and image classification via analogical transfer learning. *ACM Trans. Intell. Syst. Technol.* **9** (2018) 71:1–71:20
- [209] Osahor, U., Nasrabadi, N.M.: Ortho-shot: Low displacement rank regularization with data augmentation for few-shot learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision. (2022) 2040–2049
- [210] Qin, T., Li, W., Shi, Y., Gao, Y.: Diversity helps: Unsupervised few-shot learning via distribution shift-based data augmentation. *arXiv preprint arXiv:2004.05805* (2020)
- [211] Wei, J., Huang, C., Vosoughi, S., Cheng, Y., Xu, S.: Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,. (2021) 5493–5500
- [212] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. (2014) 2672–2680
- [213] Mehrotra, A., Dukkipati, A.: Generative adversarial residual pairwise networks for one shot learning. *arXiv preprint arXiv:1703.08033* (2017)
- [214] Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., Song, Y.: Metagan: An adversarial approach to few-shot learning. In: Advances in Neural Information Processing Systems. (2018) 2371–2380

- [215] Liu, Y., Zhang, H., Zhang, W., Lu, G., Tian, Q., Ling, N.: Few-shot image classification: Current status and research trends. *Electronics* **11** (2022) 1752
- [216] Yang, J., Guo, X., Li, Y., Marinello, F., Ercisli, S., Zhang, Z.: A survey of few-shot learning in smart agriculture: developments, applications, and challenges. *Plant Methods* **18** (2022) 28
- [217] Ebrahimi, S., Petryk, S., Gokul, A., Gan, W., Gonzalez, J.E., Rohrbach, M., Darrell, T.: Remembering for the right reasons: Explanations reduce catastrophic forgetting. In: *International Conference on Learning Representations*. (2021)
- [218] Shi, G., Chen, J., Zhang, W., Zhan, L., Wu, X.: Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In: *Advances in Neural Information Processing Systems*. (2021) 6747–6761
- [219] Wen, J., Cao, Y., Huang, R.: Few-shot self reminder to overcome catastrophic forgetting. *arXiv preprint arXiv:1812.00543* (2018)
- [220] Li, A., Luo, T., Xiang, T., Huang, W., Wang, L.: Few-shot learning with global class representations. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 9714–9723
- [221] Ding, G., Han, X., Wang, S., Wu, S., Jin, X., Tu, D., Huang, Q.: Attribute group editing for reliable few-shot image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 11184–11193
- [222] Köksal, A., Schick, T., Schütze, H.: MEAL: stable and active learning for few-shot prompting. In: *Findings of the Association for Computational Linguistics*. (2023) 506–517
- [223] Li, W., Liu, X., Bilen, H.: Cross-domain few-shot learning with task-specific adapters. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 7151–7160
- [224] Sun, J., Lapuschkin, S., Samek, W., Zhao, Y., Cheung, N., Binder, A.: Explanation-guided training for cross-domain few-shot classification. In: *25th International Conference on Pattern Recognition*. (2020) 7609–7616
- [225] Han, G., Ma, J., Huang, S., Chen, L., Chellappa, R., Chang, S.: Multimodal few-shot object detection with meta-learning based cross-modal prompting. *arXiv preprint arXiv:2204.07841* (2022)
- [226] Moon, J., Le, N.A., Minaya, N.H., Choi, S.I.: Multimodal few-shot learning for gait recognition. *Applied Sciences* **10** (2020) 7619
- [227] Pahde, F., Puscas, M.M., Klein, T., Nabi, M.: Multimodal prototypical networks for few-shot learning. In: *IEEE Winter Conference on Applications of Computer Vision*. (2021) 2643–2652
- [228] Peng, Z., Li, Z., Zhang, J., Li, Y., Qi, G., Tang, J.: Few-shot image recognition with knowledge transfer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 441–449
- [229] Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: *CVPR*. (2023) 19187–19197
- [230] Cheng, G., Lai, P., Gao, D., Han, J.: Class attention network for image recognition. *Sci. China Inf. Sci.* **66** (2023)
- [231] Wang, C., Bochkovskiy, A., Liao, H.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *CVPR*. (2023) 7464–7475
- [232] Kaur, R., Singh, S.: A comprehensive review of object detection with deep learning. *Digit. Signal Process.* **132** (2022) 103812
- [233] Chen, J., Lu, J., Zhu, X., Zhang, L.: Generative semantic segmentation. In: *CVPR*. (2023) 7111–7120
- [234] Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: *CVPR*. (2023) 2945–2954

- [235] Ye, H., Zhan, D., Jiang, Y., Zhou, Z.: Heterogeneous few-shot model rectification with semantic mapping. *IEEE Trans. Pattern Anal. Mach. Intell.* **43** (2021) 3878–3891
- [236] Abbas, M., Xiao, Q., Chen, L., Chen, P., Chen, T.: Sharp-maml: Sharpness-aware model-agnostic meta learning. In: *ICML*. Volume 162. (2022) 10–32
- [237] Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: *CVPR*. (2019) 10657–10665
- [238] Ji, Z., Chai, X., Yu, Y., Pang, Y., Zhang, Z.: Improved prototypical networks for few-shot learning. *Pattern Recognit. Lett.* **140** (2020) 81–87
- [239] Ye, H., Hu, H., Zhan, D., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: *CVPR*. (2020) 8805–8814
- [240] Huang, H., Wu, Z., Li, W., Huo, J., Gao, Y.: Local descriptor-based multi-prototype network for few-shot learning. *Pattern Recognit.* **116** (2021) 107935
- [241] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *CVPR*. (2018) 7132–7141
- [242] Liu, Y., Zheng, T., Song, J., Cai, D., He, X.: DMN4: few-shot learning via discriminative mutual nearest neighbor neural network. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. (2022) 1828–1836
- [243] Dong, C., Li, W., Huo, J., Gu, Z., Gao, Y.: Learning task-aware local representations for few-shot learning. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. (2020) 716–722
- [244] Qiao, Q., Xie, Y., Zeng, Z., Li, F.: Tallds-net: Task-aware adaptive local descriptors selection for few-shot image classification. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. (2024) 3750–3754
- [245] Liu, Y., Zhang, W., Xiang, C., Zheng, T., Cai, D., He, X.: Learning to affiliate: Mutual centralized learning for few-shot classification. In: *CVPR*. (2022) 14391–14400
- [246] Zhang, M., Huang, S., Li, W., Wang, D.: Tree structure-aware few-shot image classification via hierarchical aggregation. In: *ECCV*. Volume 13680. (2022) 453–470
- [247] Yan, L., Li, F., Zheng, X., Zhang, L.: Few-shot learning via task-aware discriminant local descriptors network. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. (2023) 2887–2894
- [248] Fang, J., Jiang, M., Fan, J., Wang, B., Li, F.: Complementarily learning decoupled category-region-aware prototype for few-shot classification. *ACM Trans. Multim. Comput. Commun. Appl.* **21** (2025) 206:1–206:22
- [249] An, Y., Xue, H., Zhao, X., Zhang, L.: Conditional self-supervised learning for few-shot classification. In Zhou, Z., ed.: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. (2021) 2140–2146
- [250] Li, H., Yang, L., Gao, F.: More attentional local descriptors for few-shot learning. In Farkas, I., Masulli, P., Wermter, S., eds.: *Artificial Neural Networks and Machine Learning*. Volume 12396. (2020) 419–430
- [251] Zheng, Z., Feng, X., Yu, H., Li, X., Gao, M.: BDLA: bi-directional local alignment for few-shot learning. *Appl. Intell.* **53** (2023) 769–785
- [252] Song, Q., Zhou, S., Chen, D.: Learning more discriminative local descriptors with parameter-free weighted attention for few-shot learning. *Mach. Vis. Appl.* **35** (2024) 71
- [253] Sun, Z., Zheng, W., Guo, P.: Klsanet: Key local semantic alignment network for few-shot image classification. *Neural Networks* **178** (2024) 106456
- [254] Yan, B.: Feature aligning few shot learning method using local descriptors weighted rules. *arXiv preprint arXiv:2408.14192* (2024)
- [255] Gao, R., Su, H., Prasad, S., Tang, P.: Few-shot classification with multisemantic information fusion network. *Image Vis. Comput.* **141** (2024) 104869
- [256] Zhou, Z., Qiu, X., Xie, J., Wu, J., Zhang, C.: Binocular mutual learning for improving few-shot classification. In: *Proceedings of the IEEE/CVF International Conference on Computer*

- Vision. (2021) 8382–8391
- [257] Wu, J., Zhang, T., Zhang, Y., Wu, F.: Task-aware part mining network for few-shot learning. In: ICCV. (2021) 8413–8422
- [258] Wang, Z., Lu, Y., Qiu, Q.: Meta-ole: Meta-learned orthogonal low-rank embedding. In: IEEE/CVF Winter Conference on Applications of Computer Vision. (2023) 5294–5303
- [259] He, J., Kortylewski, A., Yuille, A.L.: CORL: compositional representation learning for few-shot classification. In: IEEE/CVF Winter Conference on Applications of Computer Vision. (2023) 3879–3888
- [260] Ye, H., Zhou, D., Hong, L., Li, Z., Wei, X., Zhan, D.: Contextualizing meta-learning via learning to decompose. *IEEE Trans. Pattern Anal. Mach. Intell.* **46** (2024) 117–133
- [261] Qiao, Q., Xie, Y., Huang, S., Li, F.: A simple task-aware contrastive local descriptor selection strategy for few-shot learning between inter class and intra class. In: International Conference on Artificial Neural Networks. Volume 15016. (2024) 75–88
- [262] Wang, M., Yan, B.: Ldca: Local descriptors with contextual augmentation for few-shot learning. *arXiv preprint arXiv:2401.13499* (2024)
- [263] Cai, A., Hu, W., Zheng, J.: Few-shot learning for medical image classification. In: Artificial Neural Networks and Machine Learning. Volume 12396. (2020) 441–452
- [264] Achmamad, A., Ghazouani, F., Ruan, S.: Few-shot learning for brain tumor segmentation from mri images. In: 2022 16th IEEE International Conference on signal processing (ICSP). Volume 1. (2022) 489–494
- [265] Hong, B., Zhou, Y., Qin, H., Wei, Z., Liu, H., Yang, Y.: Few-shot object detection using multimodal sensor systems of unmanned surface vehicles. *Sensors* **22** (2022) 1511
- [266] Zhuang, Y., Liu, P., Yang, H., Zhang, K., Wang, Y., Pu, Z.: Few-shot learning for novel object detection in autonomous driving. *Communications in Transportation Research* **5** (2025) 100194
- [267] Alajaji, D.A., Alhichri, H.: Few shot scene classification in remote sensing using meta-agnostic machine. In: 2020 6th conference on data science and machine learning applications. (2020) 77–80
- [268] Chen, X., Zhu, G., Liu, M., Chen, Z.: Few-shot remote sensing image scene classification based on multiscale covariance metric network (mcmnet). *Neural Networks* **163** (2023) 132–145
- [269] Rajasegaran, J., Khan, S., Hayat, M., Khan, F.S., Shah, M.: Self-supervised knowledge distillation for few-shot learning. In: The British Machine Vision Conference. (2021) 179
- [270] Chen, D., Chen, Y., Li, Y., Mao, F., He, Y., Xue, H.: Self-supervised learning for few-shot image classification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2021) 1745–1749
- [271] Dong, S., Hong, X., Tao, X., Chang, X., Wei, X., Gong, Y.: Few-shot class-incremental learning via relation knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2021) 1255–1263
- [272] Huang, X., Choi, S.H.: Sapenet: Self-attention based prototype enhancement network for few-shot learning. *Pattern Recognit.* **135** (2023) 109170
- [273] Yang, X., Yao, H., Wei, Y.: One meta-tuned transformer is what you need for few-shot learning. In: Proceedings of the 41st International Conference on Machine Learning. (2024)
- [274] Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2017) 936–944
- [275] Yang, L., Zhang, R., Li, L., Xie, X.: Simam: A simple, parameter-free attention module for convolutional neural networks. In: International Conference on Machine Learning. Volume 139. (2021) 11863–11874
- [276] Li, A., Luo, T., Lu, Z., Xiang, T., Wang, L.: Large-scale few-shot learning: Knowledge

- transfer with class hierarchy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 7212–7220
- [277] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. (2017) 5998–6008
- [278] Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: European Conference on Computer Vision. (2018) 3–19
- [279] Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 7794–7803
- [280] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. (2021)
- [281] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 9992–10002
- [282] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 11531–11539
- [283] Jia, X., Su, Y., Zhao, H.: Few-shot learning via relation network based on coarse-grained granulation. *Appl. Intell.* **53** (2023) 996–1008
- [284] Lim, J.Y., Lim, K., Lee, C., Tan, Y.X.: Ssl-protonet: Self-supervised learning prototypical networks for few-shot learning. *Expert Syst. Appl.* **238** (2024) 122173
- [285] Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2024) 4818–4829
- [286] Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F.N., Krishnamoorthi, R., Chandra, V.: EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2024) 16111–16121
- [287] Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., Ganslandt, T.: Transfer learning for medical image classification: a literature review. *BMC Medical Imaging* **22** (2022) 69
- [288] Blaes, S., Burwick, T.: Few-shot learning in deep networks through global prototyping. *Neural Networks* **94** (2017) 159–172
- [289] Jia, X., Mao, Y., Pan, Z., Wang, Z., Ping, P.: Few-shot learning based on hierarchical feature fusion via relation networks. *Int. J. Approx. Reason.* **170** (2024) 109186
- [290] Lim, J.Y., Lim, K., Lee, C., Tan, Y.X.: SFT: few-shot learning via self-supervised feature fusion with transformer. *IEEE Access* **12** (2024) 86690–86703
- [291] Ding, E., Chu, X., Liu, Z., Zhang, K., Yu, Q.: A novel hierarchical adaptive feature fusion method for meta-learning. *Applied Sciences* **12** (2022) 5458
- [292] Wu, Z., Zhao, H.: Hierarchical few-shot learning with feature fusion driven by data and knowledge. *Inf. Sci.* **639** (2023) 119012
- [293] Qin, W., Zhao, W., Li, M.: Multi-level feature representation and multi-layered fusion contrast for few-shot classification. *IAENG International Journal of Computer Science* **49** (2022)
- [294] Gomes, J.C., de A. B. Borges, L., Borges, D.L.: A multi-layer feature fusion method for few-shot image classification. *Sensors* **23** (2023) 6880
- [295] Jiang, W., Huang, K., Geng, J., Deng, X.: Multi-scale metric learning for few-shot learning. *IEEE Trans. Circuits Syst. Video Technol.* **31** (2021) 1091–1102

- [296] Zhong, F., Chen, Z., Zhang, Y., Xia, F.: Zero- and few-shot learning for diseases recognition of citrus aurantium l. using conditional adversarial autoencoders. *Comput. Electron. Agric.* **179** (2020) 105828
- [297] Qiu, L., Chen, Y., Jia, H., Zhang, Z.: Query intent recognition based on multi-class features. *IEEE Access* **6** (2018) 52195–52204
- [298] Zhou, X., Zhang, T., Cheng, C., Song, S.: Dynamic multichannel fusion mechanism based on a graph attention network and BERT for aspect-based sentiment classification. *Appl. Intell.* **53** (2023) 6800–6813
- [299] Hua, J., Liu, J., Zhao, Y.: Few shot object detection based on feature fusion. *Comput Sci* **50** (2023) 209–213
- [300] Yu, L., Tian-tian, Y., Dong-sheng, Z., Xiao-peng, W.: Natural scene text detection based on attention mechanism and deep multi-scale feature fusion. *Journal of Graphics* **44** (2023) 473
- [301] Allen, K.R., Shelhamer, E., Shin, H., Tenenbaum, J.B.: Infinite mixture prototypes for few-shot learning. In: *International Conference on Machine Learning*. Volume 97. (2019) 232–241
- [302] Fei, N., Lu, Z., Xiang, T., Huang, S.: MELR: meta-learning via modeling episode-level relationships for few-shot learning. In: *International Conference on Learning Representations*. (2021)
- [303] Khruklov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I.V., Lempitsky, V.S.: Hyperbolic image embeddings. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2020) 6417–6427
- [304] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2016) 770–778
- [305] Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., Qu, R.: A survey of deep learning-based object detection. *IEEE Access* **7** (2019) 128837–128868
- [306] Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: Learning embedding adaptation for few-shot learning. *arXiv preprint arXiv:1812.03664* **7** (2018)
- [307] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015)
- [308] Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2022) 11943–11952
- [309] Leng, Z., Tan, M., Liu, C., Cubuk, E.D., Shi, J., Cheng, S., Anguelov, D.: Polyloss: A polynomial expansion perspective of classification loss functions. In: *International Conference on Learning Representations*. (2022)
- [310] Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018)
- [311] Bansal, T., Jha, R., Munkhdalai, T., McCallum, A.: Self-supervised meta-learning for few-shot natural language classification tasks. *arXiv preprint arXiv:2009.08445* (2020)
- [312] Gan, C., Gong, B., Liu, K., Su, H., Guibas, L.J.: Geometry guided convolutional neural networks for self-supervised video representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2018) 5589–5597
- [313] Gan, C., Zhao, H., Chen, P., Cox, D.D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 7052–7061
- [314] Furlanello, T., Lipton, Z.C., Tschannen, M., Itti, L., Anandkumar, A.: Born-again neural networks. In: *International Conference on Machine Learning*. Volume 80. (2018) 1602–1611
- [315] Mobahi, H., Farajtabar, M., Bartlett, P.L.: Self-distillation amplifies regularization in hilbert space. In: *Advances in Neural Information Processing Systems*. (2020)

- [316] Lim, J.Y., Lim, K., Ooi, S.Y., Lee, C.: Efficient-prototypicalnet with self knowledge distillation for few-shot learning. *Neurocomputing* **459** (2021) 327–337
- [317] Liu, S., Wang, Y.: Few-shot learning with online self-distillation. In: *IEEE/CVF International Conference on Computer Vision Workshops*. (2021) 1067–1070
- [318] Lin, H., Han, G., Ma, J., Huang, S., Lin, X., Chang, S.: Supervised masked knowledge distillation for few-shot transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2023) 19649–19659
- [319] Asano, Y.M., Rupprecht, C., Vedaldi, A.: A critical analysis of self-supervision, or what we can learn from a single image. In: *International Conference on Learning Representations*. (2020)
- [320] Ravichandran, A., Bhotika, R., Soatto, S.: Few-shot learning with embedded class models and shot-free meta training. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 331–339
- [321] Ma, R., Fang, P., Drummond, T., Harandi, M.: Adaptive poincaré point to set distance for few-shot classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. (2022) 1926–1934
- [322] He, J., Hong, R., Liu, X., Xu, M., Sun, Q.: Revisiting local descriptor for improved few-shot classification. *ACM Trans. Multim. Comput. Commun. Appl.* **18** (2022) 127:1–127:23
- [323] Gao, F., Cai, L., Yang, Z., Song, S., Wu, C.: Multi-distance metric network for few-shot learning. *Int. J. Mach. Learn. Cybern.* **13** (2022) 2495–2506
- [324] Zhang, C., Ding, H., Lin, G., Li, R., Wang, C., Shen, C.: Meta navigator: Search for a good adaptation policy for few-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2021) 9415–9424
- [325] Zhou, F., Wu, B., Li, Z.: Deep meta-learning: Learning to learn in the concept space. *arXiv preprint arXiv:1802.03596* (2018)
- [326] Dvornik, N., Mairal, J., Schmid, C.: Diversity with cooperation: Ensemble methods for few-shot classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019) 3722–3730
- [327] Xiao, Y., Jin, Y., Hao, K.: Adaptive prototypical networks with label words and joint representation learning for few-shot relation classification. *IEEE Trans. Neural Networks Learn. Syst.* **34** (2023) 1406–1417
- [328] Liu, J., Song, L., Qin, Y.: Prototype rectification for few-shot learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J., eds.: *European Conference on Computer Vision*. Volume 12346. (2020) 741–756
- [329] Sun, W., Du, Y., Zhen, X., Wang, F., Wang, L., Snoek, C.G.M.: Metamodulation: Learning variational feature hierarchies for few-shot learning with fewer tasks. In: *International Conference on Machine Learning*. Volume 202. (2023) 32847–32858
- [330] Du, Y., Xiao, Z., Liao, S., Snoek, C.: Protodiff: Learning to learn prototypical networks by task-guided diffusion. In: *Advances in Neural Information Processing Systems*. (2023)
- [331] Ji, Z., Ni, J., Liu, X., Pang, Y.: Teachers cooperation: team-knowledge distillation for multiple cross-domain few-shot learning. *Frontiers Comput. Sci.* **17** (2022) 172312
- [332] Hsieh, C., Li, C., Yeh, C., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: *Findings of the Association for Computational Linguistics*. (2023) 8003–8017

List of Abbreviations

Abbreviation	Full Name / Short Definition
DCNNs	Deep Convolutional Neural Networks / A regularized version of multilayer perceptrons based on convolution kernels
ReLU	Rectified Linear Unit / An activation function returns 0 if it receives any negative input
FC	Fully-Connected (layer)
KNN	K-Nearest Neighbors
FSL	Few-Shot Learning / A computer vision task
ProtoSimNet	Prototype Similarity Network / A metric-based framework for enhanced class separability in few-shot embedding space.
AFPN	Attention-based Feature Pyramid Network / a scale-aware hierarchical feature representation with attention-guided cross-level aggregation for few-shot learning.
AMFF	Adaptive Multi-scale Feature Fusion / a multi-scale feature enrichment mechanism for improving representation diversity under few-shot supervision.
DSKT	Dual Selective Knowledge Transfer / a selective knowledge transfer paradigm for robust generalization to novel few-shot tasks.
MSE	Mean Squared Error / A metric to measure the distance between two features
BCE	Binary Cross Entropy / A metric that tracks incorrect labeling of the data class by a model
SSL	Self-supervised Learning / A learning process where the model trains itself to learn one part of the input from another part of the input
KL-divergence	Kullback–Leibler divergence / A metric to measure the distance between two distributions
OOD	out-of distribution / it is used to describe a dataset where its train and test splits are in different label distributions
IID	identically distributed (data)

English Summary

We are living in an era of artificial intelligence where the demand for rapid adaptation to new visual concepts grows exponentially. It is crucial to develop intelligent visual recognition systems that can learn efficiently from limited examples. A particularly important capability, given the expanding need for customized visual applications, is few-shot learning (FSL). A FSL system aims to recognize novel categories from very few labeled instances, enabling fast deployment in specialized domains where data collection is expensive or time-consuming, providing a paradigm for fast, flexible, and scalable recognition in real-world environments.

To enable FSL systems to meet the so-called open-world challenges, we explore three themes that are challenging yet practical in real-world application scenarios.

The first theme of this thesis is to address metric learning challenges in FSL. While metric-based approaches have become predominant, conventional similarity measurements often fail to maintain discriminative power when dealing with fine-grained categories or significant domain shifts. To migrate these limitations, we propose a novel prototype similarity network (ProtoSimNet), which integrates cosine sharpening transformation (CST) and elastic regularization. CST introduces a parametric sharpening function that nonlinearly expands angular decision margins in the cosine similarity space, while the elastic regularization component maintains embedding stability across diverse learning episodes. This combined approach effectively enlarges inter-class differences while preventing overfitting to limited support examples. ProtoSimNet preserves relative similarity rankings while significantly improving model’s discriminative ability, particularly in fine-grained classification scenarios. Extensive experiments demonstrate that our ProtoSimNet establishes new state-of-the-art performance across multiple few-shot benchmarks.

The second theme focuses on feature enrichment, emphasizing the need for richer and more structured visual representations under limited supervision. Traditional feature extractors suffer from insufficient diversity and poor scale adaptability when trained with limited samples. To address these issues, we develop two frameworks: the attention-guided feature pyramid network (AFPN) and the adaptive multi-scale feature fusion (AMFF). AFPN enhances hierarchical feature integration through parameter-free attention mechanisms and cross-layer spatial attention, enabling dynamic emphasis on the most relevant semantic levels for each few-shot task. AMFF

systematically enriches feature representation through structured multi-branch enrichment and adaptive fusion, effectively simulating the feature diversity typically unavailable in few-shot settings. These frameworks demonstrate remarkable effectiveness in handling scale variations, domain shifts, and fine-grained categorization, significantly outperforming conventional feature extraction methods.

The third theme explores knowledge transfer to leverage information learned from large base categories to guide fast learning on novel classes. Standard distillation methods lack selectivity and often transfer irrelevant or misleading information when categories differ significantly. To address this issue, we propose a dual-selective knowledge transfer (DSKT) framework that dynamically selects the most beneficial knowledge components from both teacher models and current task features. By selectively distilling both instance-level and task-level supervisory information, DSKT preserves transferable structure, suppresses noise, and enhances generalization to unseen categories. It is shown that our proposed methods achieve the new state-of-the-art performance in FSL tasks.

We conducted thorough experiments to verify the effectiveness of our proposed methods across these three themes. The results demonstrate significant improvements over various baselines and state-of-the-art methods. Thus this thesis provides important novel contributions, insights, and findings for the research community and future applications in the field of visual understanding and FSL systems.

Nederlandse Samenvatting

We leven in een tijdperk van kunstmatige intelligentie waarin de vraag naar snelle aanpassing aan nieuwe visuele concepten exponentieel groeit. Het is cruciaal om intelligente visuele herkenningssystemen te ontwikkelen die efficiënt kunnen leren van een beperkt aantal voorbeelden. Een bijzonder belangrijke eigenschap, gezien de toenemende behoefte aan gepersonaliseerde visuele toepassingen, is *few-shot learning* (FSL). Een FSL-systeem is erop gericht nieuwe categorieën te herkennen op basis van een zeer klein aantal gelabelde voorbeelden, waardoor snelle implementatie mogelijk is in gespecialiseerde domeinen waar dataverzameling duur of tijdrovend is. Dit biedt een paradigma voor snelle, flexibele en schaalbare herkenning in realistische omgevingen.

Om FSL-systemen in staat te stellen de zogenaamde open-werelduitdagingen aan te gaan, onderzoeken we drie thema's die uitdagend maar tegelijkertijd praktisch zijn in toepassingsscenario's in de echte wereld.

Het eerste thema van deze scriptie is het aanpakken van de uitdagingen van metrische leertechnieken in FSL. Hoewel op metrische methoden gebaseerde benaderingen dominant zijn geworden, slagen conventionele gelijkenismetingen er vaak niet in om hun onderscheidend vermogen te behouden bij fijnmazige categorieën of significante domeinverschuivingen. Om deze beperkingen te ondervangen, stellen we een nieuw prototype gelijkenisnetwerk voor (ProtoSimNet), dat het cosinusverscherpingstransformatie (CST) en elastische regularisatie integreert. CST introduceert een parametrische verscherpingsfunctie die de hoekige beslissingsmarges in de cosinusgelijkenisruimte niet-lineair vergroot, terwijl de elastische regularisatiecomponent de stabiliteit van de embedding over verschillende leerepisodes behoudt. Deze gecombineerde aanpak vergroot effectief de verschillen tussen klassen en voorkomt tegelijkertijd overfitting aan voorbeelden met beperkte ondersteuning. ProtoSimNet behoudt relatieve gelijkenisrangschikkingen en verbetert tegelijkertijd het onderscheidend vermogen van het model aanzienlijk, met name in fijnmazige classificatiescenario's. Uitgebreide experimenten tonen aan dat ons ProtoSimNet nieuwe state-of-the-art prestaties behaalt op meerdere *few-shot* benchmarks.

Het tweede thema richt zich op het verrijken van kenmerken, waarbij de nadruk ligt op de behoefte aan rijkere en meer gestructureerde visuele representaties onder

beperkte supervisie. Traditionele kenmerkextractors hebben te lijden onder onvoldoende diversiteit en een slechte schaalaanpassing wanneer ze getraind worden met een beperkt aantal voorbeelden. Om deze problemen aan te pakken, ontwikkelen we twee frameworks: het attention-guided feature pyramid network (AFPN) en het adaptive multi-scale feature fusion (AMFF). AFPN verbetert de hiërarchische integratie van kenmerken door middel van parameterloze aandachtmechanismen en ruimtelijke aandacht over meerdere lagen, waardoor dynamische nadruk kan worden gelegd op de meest relevante semantische niveaus voor elke few-shot taak. AMFF verrijkt systematisch de kenmerkrepresentatie door middel van gestructureerde multi-branch verrijking en adaptieve fusie, waardoor de kenmerkdiversiteit die doorgaans niet beschikbaar is in few-shot situaties effectief wordt gesimuleerd. Deze frameworks tonen een opmerkelijke effectiviteit in het omgaan met schaalvariëaties, domeinverschuivingen en fijnmazige categorisatie, en presteren significant beter dan conventionele methoden voor kenmerkextractie.

Het derde thema onderzoekt kennisoverdracht om informatie die is opgedaan uit grote basiscategorieën te benutten voor snel leren van nieuwe klassen. Standaard distillatiemethoden missen selectiviteit en dragen vaak irrelevante of misleidende informatie over wanneer categorieën significant verschillen. Om dit probleem aan te pakken, stellen we een dual-selective knowledge transfer (DSKT)-framework voor dat dynamisch de meest nuttige kenniscomponenten selecteert uit zowel de leermodellen als de huidige taakkenmerken. Door selectief zowel supervisie-informatie op instantieniveau als op taakniveau te distilleren, behoudt DSKT de overdraagbare structuur, onderdrukt ruis en verbetert de generalisatie naar onbekende categorieën. Er wordt aangetoond dat onze voorgestelde methoden de nieuwe state-of-the-art prestaties behalen in FSL-taken.

We hebben grondige experimenten uitgevoerd om de effectiviteit van onze voorgestelde methoden te verifiëren aan de hand van deze drie thema's. De resultaten tonen significante verbeteringen ten opzichte van diverse basismodellen en state-of-the-art methoden. Deze dissertatie levert daarmee belangrijke nieuwe bijdragen, inzichten en bevindingen voor de onderzoeksgemeenschap en toekomstige toepassingen op het gebied van visuele interpretatie en FSL-systemen.

Acknowledgements

This thesis represents an important stage of my academic journey. During this period, I experienced uncertainty, pressure, and many moments of self-doubt, but I also gained valuable knowledge, confidence, and personal growth. The completion of this work would not have been possible without the people who supported me, guided me, and encouraged me along the way.

I would like to express my sincere gratitude to my supervisors, Prof.dr. Michael Lew and Dr. Erwin Bakker, for their patient guidance and continuous support throughout this research. Their suggestions helped me improve not only the technical quality of my work, but also the way I think about research problems. Whenever I encountered difficulties in experiments, writing, or research design, their advice encouraged me to reconsider the problem carefully and move forward with greater confidence. They reviewed my thesis carefully and gave me insightful comments and suggestions. I am also grateful to my colleagues, Dr. Nan Pu and Dr. Mingrui Lao, and other friends who offered helpful discussions and suggestions during my study. Their comments helped me view my work from different perspectives and improve the clarity of this thesis. The time spent discussing research, solving problems, and sharing experiences with them made this journey less lonely and more meaningful.

I would like to give special appreciation to my wife. Thank you for your unwavering love, patience, and companionship throughout this journey. Your encouragement gave me strength during difficult times, and your understanding helped me continue pursuing my academic goals with confidence. Because of your support, the challenges along the way became easier to face, and the ordinary days became warmer and more meaningful. I am truly grateful to have you by my side. Let us continue to grow together and welcome the better days ahead.

Finally, I would like to thank everyone who has accompanied me during this journey. This thesis is not only a record of my research work, but also a reminder of the support, kindness, and encouragement I received throughout this important period of my life.

Kai He

22 September 2026

Leiden, the Netherlands

Curriculum Vitae

Kai He was born in Yangzhou, China on April 2, 1995. In 2013, he started his bachelor's study at Soochow University in Soochow, Jiangsu, China, and received his bachelor's degree in 2017. After that, he started his master's study at Shanxi Normal University in Xi'an, China, and obtained his master's degree in 2021 under the supervision of Prof.dr. Yali Peng.

In December 2021, he started his PhD research supported by the China Scholarship Council and worked at the MediaLab in the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, the Netherlands, under the supervision of Prof.dr. Michael Lew and Dr. Erwin Bakker. During his PhD studies, he took courses in Computer Vision by Learning and scientific conduct, among others. Kai He's research interests include computer vision, multimedia retrieval and deep learning. He is focusing on few-shot learning. He has published papers in international journals and conferences, including Applied Intelligence, International Journal of Multimedia Information Retrieval, International Conference on Intelligent Computing *etc.* He also serves as a reviewer for some conferences and journals, such as IEEE Access, IEEE Transactions on Cognitive Communications and Networking

Publication List

Main publications for the thesis

- **He, K.**, Pu, N., Lao, M., and Lew, M.S. “Few-shot and meta-learning methods for image understanding: a survey.” International Journal of Multimedia Information Retrieval, 2023.
- **He, K.**, Pu, N., Bakker, E. M., and Lew, M. S. “ProtoSimNet: Exploiting Prototype Similarity for Few-shot Learning.” accepted by the Neurocomputing journal, 2026.
- **He, K.**, Bakker, E. M., and Lew, M. S. “Attention-guided Feature Pyramid Network for few-shot learning.” Benelux Conference on Artificial Intelligence and Machine Learning, 2024.
- **He, K.**, Bakker, E. M., and Lew, M. S. “Adaptive Multi-scale Feature Fusion Network for Few-Shot Learning.” International Conference on Intelligent Computing, 2025.
- **He, K.**, Pu, N., Lao, M., Bakker, E. M., and Lew, M.S. “Dual selective knowledge transfer for few-shot classification.” Applied Intelligence, 2023.

Other publications

- Lao, M., Pu, N., Liu, Y., **He K.**, Bakker, E. M., and Lew, M. S. “COCA: Collaborative CAusal Regularization for Audio-Visual Question Answering.” AAAI Conference on Artificial Intelligence, 2023.