



Universiteit
Leiden

The Netherlands

Using deep learning models in image processing

Xiong, Z.

Citation

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

Version: Publisher's Version

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

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

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

Bibliography

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

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

Bibliography

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

Supplementary Materials in Chapter 5

As an extra, we have added two zoomed-in samples to highlight the details of our method and other approaches. The first row of Figure A.1 shows that our method can correctly divide two glomeruli without overlap. Cascade Mask R-CNN and Mask R-CNN, however, generate multiple responses at the border of two objects. QueryInst with static queries cannot detect or segment any glomerulus. The last row of Figure A.1 demonstrates four large arteries in a line. It is complicated because two large arteries pass over most of the input image regions. Although it had multiple responses at the artery border, our method can capture long-range spatial dependencies for large objects using attention mechanisms with dynamic queries. In contrast, Cascade Mask RCNN and Mask RCNN cannot model large objects without attention modules. QueryInst uses a static attention mechanism and cannot handle complex combinations of shapes and appearances with a fixed number of queries.

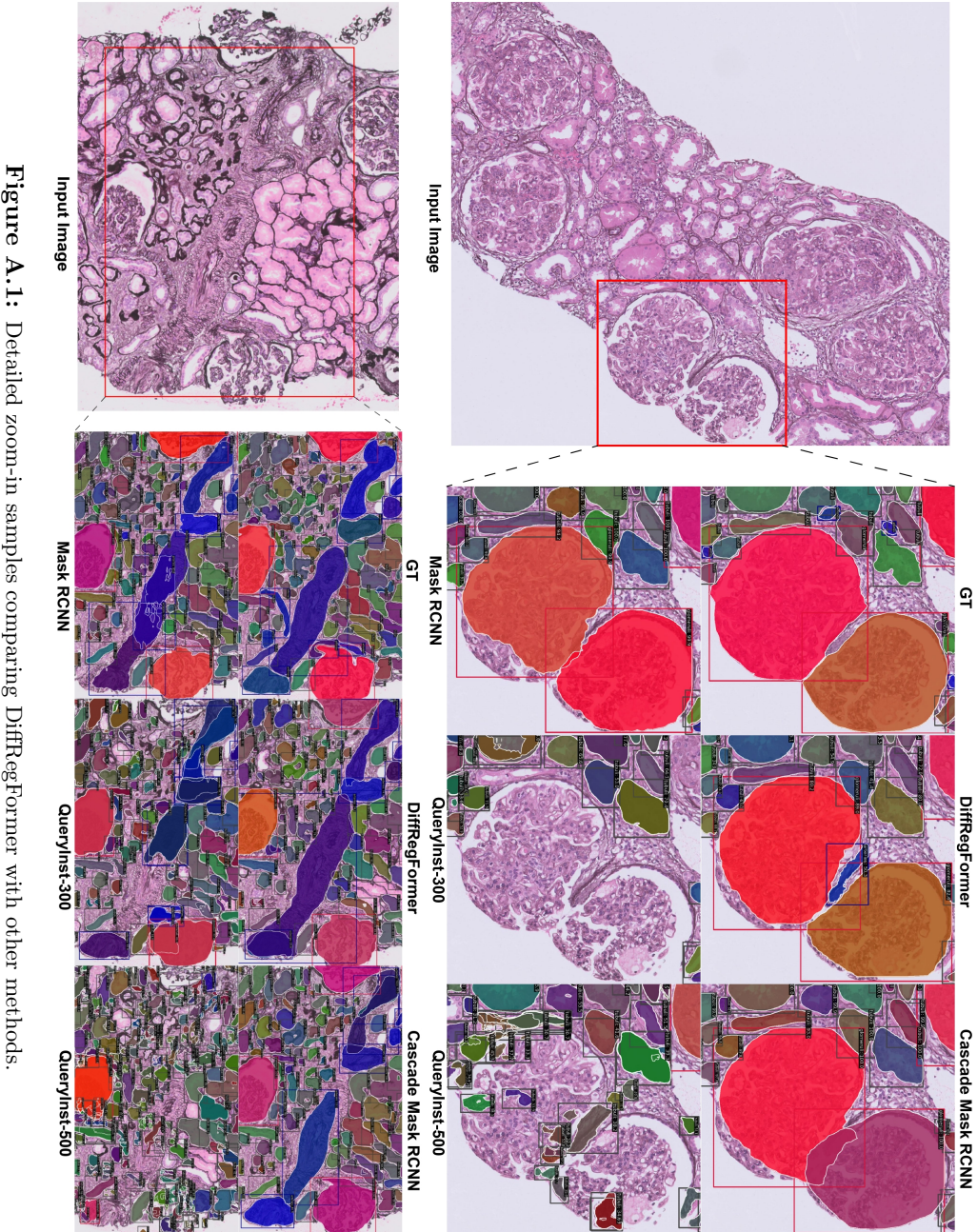


Figure A.1: Detailed zoom-in samples comparing DiffRegFormer with other methods.