Bibliography

[1] Accessed: 2025-06-07. 2023. URL: https://github.com/Alpha-VLLM/LLaMA2-Accessory/tree/f7fe19834b23e38f333403b91bb0330afe19f79e/Large-DiT-ImageNet.

[2] Henrik Aanaes et al. "Large-Scale Data for Multiple-View Stereopsis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[3] Amir Ahmadyan et al. "Objectron: A Large Scale Dataset of Object-Centric Videos in the Wild with Pose Annotations". In: *CVPR*. 2021.

[4] Yuval Alaluf et al. *Third Time's the Charm? Image and Video Editing with StyleGAN3*. 2022. arXiv: 2201.13433 [cs.CV]. URL: https://arxiv.org/abs/2201.13433.

[5] Jay Alammar. *The Illustrated Transformer*. https://jalammar.github.io/illustrated-transformer/. 2018.

[6] Jean-Baptiste Alayrac et al. "Flamingo: A Visual Language Model for Few-Shot Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[7] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. "Measuring the Objectness of Image Windows". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2189–2202. DOI: 10.1109/TPAMI.2012.28.

[8] Nicholas Vieau Alger. *Data-scalable Hessian preconditioning for distributed parameter PDE-constrained inverse problems*. The University of Texas at Austin, 2019.

[9] Kara-Ali Aliev et al. *Neural Point-Based Graphics*. 2020. arXiv: 1906.08240 [cs.CV]. URL: https://arxiv.org/abs/1906.08240.

[10] Xiang An et al. *LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training*. 2025. arXiv: 2509.23661 [cs.CV]. URL: https://arxiv.org/abs/2509.23661.

[11] Andrej Karpathy. *convnetjs — ConvNet (CNN) in JavaScript*. https://cs.stanford.edu/people/karpathy/convnetjs/. [Online; accessed 9-June-2025]. 2015.

[12] Appen. *Computer Vision vs Machine Vision: What's the Difference?* URL: https://www.appen.com/blog/computer-vision-vs-machine-vision-2.

[13] Relja Arandjelović and Andrew Zisserman. "Look, Listen and Learn". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 609–617. DOI: 10.1109/ICCV.2017.73.

[14] Martin Arjovsky, Soumith Chintala, and L'eon Bottou. "Wasserstein GAN". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017. URL: https://arxiv.org/abs/1701.07875.

[15] Maryam Armandpour et al. "Partitioning Gated Mechanisms for Graph Generation". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. PMLR, 2021, pp. 382–393. URL: https://par.nsf.gov/servlets/purl/10273833.

[16] Anurag Arnab et al. "ViViT: A Video Vision Transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 6836–6846. arXiv: 2103.15691. URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Arnab_ViViT_A_Video_Vision_Transformer_ICCV_2021_paper.pdf.

[17] Ali Athar et al. "BURST: A Benchmark for Unifying Object Recognition, Segmentation and Tracking in Video". In: *WACV*. 2023.

[18] Zhihu Author. *Understanding Classifier-Free Guidance in Stable Diffusion*. https://zhuanlan.zhihu.com/p/660518657. Accessed: 2025-05-18. 2023.

[19] Anas Awadalla et al. *OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models*. 2023. arXiv: 2308.01390 [cs.CV]. URL: https://arxiv.org/abs/2308.01390.

[20] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. *Multiple Object Recognition with Visual Attention*. 2015. arXiv: 1412.7755 [cs.CV].

[21] Moussa Baccouche et al. "Sequential Deep Learning for Human Action Recognition". In: *International Conference on Human Behavior Understanding*. 2011, pp. 29–39. DOI: 10.1007/978-3-642-25446-8_4.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL]. URL: https://arxiv.org/abs/1409.0473.

[23] Jinze Bai et al. "Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities". In: *arXiv preprint arXiv:2308.12966* (2023). URL: https://arxiv.org/abs/2308.12966.

[24] Jinze Bai et al. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. 2023. arXiv: 2308.12966 [cs.CV]. URL: https://arxiv.org/abs/2308.12966.

[25] Jun Bai et al. "BasicTAD: An Empirical Strong Baseline for Temporal Action Detection". In: *ICCV*. 2023.

[26] Allison H. Baker, Alexander Pinard, and Dorit M. Hammerling. "On a Structural Similarity Index Approach for Floating-Point Data". In: (2022). arXiv: 2202.02616 [cs.CV]. URL: https://arxiv.org/abs/2202.02616.

[27] Nicolas Ballas et al. “Delving Deeper into Convolutional Networks for Learning Video Representations”. In: *International Conference on Learning Representations (ICLR)*. 2016.

[28] Adrien Bardes, Jean Ponce, and Yann LeCun. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *ICLR* (2022).

[29] Jonathan T. Barron et al. *Mip-NeRF 360: Unbounded Anti-Aliased in-the-Wild Scene Rendering*. 2022. arXiv: 2111.12077 [cs.CV]. URL: https://arxiv.org/abs/2111.12077.

[30] Jonathan T. Barron et al. *Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields*. 2021. arXiv: 2103.13415 [cs.CV]. URL: https://arxiv.org/abs/2103.13415.

[31] Jonathan T. Barron et al. *Zip-NeRF: Anti-Aliased, Compositional Neural Representation for Editable 3D Scenes*. 2023. arXiv: 2304.06706 [cs.CV]. URL: https://arxiv.org/abs/2304.06706.

[32] David Bau et al. *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. 2017. arXiv: 1704.05796 [cs.CV]. URL: https://arxiv.org/abs/1704.05796.

[33] Maksym Bekuzarov. *Losses Explained: Contrastive Loss*. https://medium.com/@maksym.bekuzarov/losses-explained-contrastive-loss-f8f57fe32246. Accessed: 2025-06-14. 2022.

[34] Maksym Bekuzarov et al. “XMem++: Production-level Video Segmentation From Few Annotated Frames”. In: *ICCV*. 2023.

[35] Irwan Bello et al. *Revisiting ResNets: Improved Training and Scaling Strategies*. 2021. arXiv: 2103.07579 [cs.CV]. URL: https://arxiv.org/abs/2103.07579.

[36] Y. Bengio, P. Simard, and P. Frasconi. “Learning long-term dependencies with gradient descent is difficult”. In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166. DOI: 10.1109/72.279181.

[37] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828.

[38] Benjamin Paepper. *Interactive Visualization of Stable Diffusion Image Embeddings*. [Online; accessed 11-June-2025]. 2023. URL: https://www.paepper.com/blog/posts/interactive-visualization-of-stable-diffusion-image-embeddings/.

[39] James Bergstra and Yoshua Bengio. “Random Search for Hyper-Parameter Optimization”. In: *Journal of Machine Learning Research* 13.10 (2012), pp. 281–305. URL: http://jmlr.org/papers/v13/bergstra12a.html.

[40] James Bergstra and Yoshua Bengio. “Random Search for Hyper-Parameter Optimization”. In: *Journal of Machine Learning Research* 13.10 (2012), pp. 281–305. URL: http://jmlr.org/papers/v13/bergstra12a.html.

[41] Maxim Berman et al. “MultiGrain: a Unified Image Embedding for Classes and Instances”. In: *International Conference on Learning Representations (ICLR)*. 2019. arXiv: 1902.05509 [cs.CV].

[42] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. “Is Space-Time Attention All You Need for Video Understanding?” In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021. arXiv: 2102.05095. URL: https://proceedings.mlr.press/v139/bertasius21a/bertasius21a.pdf.

[43] Lucas Beyer et al. *FlexiViT: One Model for All Patch Sizes*. 2023. arXiv: 2212.08013 [cs.CV]. URL: https://arxiv.org/abs/2212.08013.

[44] Jia-Wang Bian et al. “NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior”. In: *arXiv:2307.05401* (2023). URL: https://arxiv.org/abs/2307.05401.

[45] Mikolaj Binkowski et al. “Demystifying MMD GANs”. In: *International Conference on Learning Representations (ICLR)* (2018). URL: https://openreview.net/pdf?id=H1QRgziT-.

[46] Andreas Blattmann et al. “Stable Video Diffusion: Scaling Latent Video Diffusion Models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023. URL: https://arxiv.org/abs/2304.08818.

[47] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “Yolov4: Optimal speed and accuracy of object detection”. In: *arXiv preprint arXiv:2004.10934* (2020).

[48] Daniel Bolya et al. *Perception Encoder: The best visual embeddings are not at the output of the network*. 2025. arXiv: 2504.13181 [cs.CV]. URL: https://arxiv.org/abs/2504.13181.

[49] Daniel Bolya et al. *Token Merging: Your ViT But Faster*. 2023. arXiv: 2210.09461 [cs.CV]. URL: https://arxiv.org/abs/2210.09461.

[50] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. “Food-101 – Mining Discriminative Components with Random Forests”. In: *European Conference on Computer Vision (ECCV)*. Vol. 8694. Lecture Notes in Computer Science. Dataset also available from ETH Zurich CVL. Springer, 2014, pp. 446–461. DOI: 10.1007/978-3-319-10599-4_29. URL: https://www.vision.ee.ethz.ch/datasets_extra/food-101/.

[51] Joshua B. Tenenbaum Brenden M. Lake Ruslan Salakhutdinov. “Human-Level Concept Learning Through Probabilistic Program Induction”. In: *Science* 350.6266 (2015), pp. 1332–1338.

[52] Andrew Brock, Jeff Donahue, and Karen Simonyan. “Large Scale GAN Training for High Fidelity Natural Image Synthesis”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[53] Andrew Brock et al. *High-Performance Large-Scale Image Recognition Without Normalization*. 2021. arXiv: 2102.06171 [cs.CV]. URL: https://arxiv.org/abs/2102.06171.

[54] Andrew Brock et al. *High-Performance Large-Scale Image Recognition Without Normalization*. 2021. arXiv: 2102.06171 [cs.CV]. URL: https://arxiv.org/abs/2102.06171.

[55] Andrew Brock et al. “Neural Photo Editing with Introspective Adversarial Networks”. In: *International Conference on Learning Representations (ICLR)*. 2017.

[56] Rodney A Brooks and Thomas O Binford. “Model-based three-dimensional interpretations of two-dimensional images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.3 (1979), pp. 224–241. DOI: 10.1109/TPAMI.1979.4766909.

[57] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. *InstructPix2Pix: Learning to Follow Image Editing Instructions*. 2023. arXiv: 2211.09800 [cs.CV]. URL: https://arxiv.org/abs/2211.09800.

[58] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[59] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[60] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2018, pp. 77–91.

[61] Han Cai, Ligeng Zhu, and Song Han. *ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware*. 2019. arXiv: 1812.00332 [cs.LG]. URL: https://arxiv.org/abs/1812.00332.

[62] Zhi Cai et al. *Align-DETR: Enhancing End-to-end Object Detection with Aligned Loss*. 2024. arXiv: 2304.07527 [cs.CV]. URL: https://arxiv.org/abs/2304.07527.

[63] John Canny. "A computational approach to edge detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (1986), pp. 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[64] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV]. URL: https://arxiv.org/abs/2005.12872.

[65] Nicolas Carion et al. "SAM 3: Segment Anything with Concepts". In: *AI at Meta* (2025). URL: https://ai.meta.com/research/publications/sam-3-segment-anything-with-concepts/.

[66] Nicholas Carlini and David Wagner. "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods". In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*. 2017, pp. 3–14. DOI: 10.1145/3128572.3140444.

[67] Nicholas Carlini and David Wagner. "Towards Evaluating the Robustness of Neural Networks". In: (2017). arXiv: 1608.04644 [cs.CR].

[68] Nicholas Carlini et al. "Universal and Transferable Adversarial Attacks on Aligned Language Models". In: *arXiv preprint arXiv:2307.15043* (2023). URL: https://arxiv.org/abs/2307.15043.

[69] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *ICCV*. 2021.

[70] Mathilde Caron et al. "Deep Clustering for Unsupervised Learning of Visual Features". In: *European Conference on Computer Vision (ECCV)* (2018).

[71] Mathilde Caron et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2021.

[72] Mathilde Caron et al. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[73] Mathilde Caron et al. *Unsupervised Pre-Training of Image Features on Non-Curated Data*. 2019. arXiv: 1905.01278 [cs.CV]. URL: https://arxiv.org/abs/1905.01278.

[74] João Carreira and Andrew Zisserman. “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4724–4733. DOI: 10.1109/CVPR.2017.502.

[75] William Chan et al. “Listen, Attend and Spell”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016.

[76] Angel X. Chang et al. *ShapeNet: An Information-Rich 3D Model Repository*. 2015. arXiv: 1512.03012 [cs.GR]. URL: https://arxiv.org/abs/1512.03012.

[77] Huiwen Chang et al. *MaskGIT: Masked Generative Image Transformer*. 2022. arXiv: 2202.04200 [cs.CV]. URL: https://arxiv.org/abs/2202.04200.

[78] Yu-Wei Chao et al. “Rethinking the Faster R-CNN Architecture for Temporal Action Localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 1130–1139. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Chao_Rethinking_the_Faster_CVPR_2018_paper.html.

[79] Anpei Chen et al. *MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo*. 2021. arXiv: 2103.15595 [cs.CV]. URL: https://arxiv.org/abs/2103.15595.

[80] Anpei Chen et al. *TensoRF: Tensorial Radiance Fields*. 2022. arXiv: 2203.09517 [cs.CV]. URL: https://arxiv.org/abs/2203.09517.

[81] Jianbo Chen et al. “HopSkipJumpAttack: A Query-Efficient Decision-Based Attack”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 4671–4680. DOI: 10.1109/CVPR42600.2020.00474.

[82] Liang-Chieh Chen et al. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Vol. 40. 4. IEEE, 2017, pp. 834–848.

[83] Mark Chen et al. “Generative Pretraining From Pixels”. In: 2020.

[84] Mark Chen et al. “Generative Pretraining from Pixels”. In: *ICML*. 2020.

[85] Minghua Chen et al. “Fantasia3D: Disentangling Geometry and Appearance for High-Quality Text-to-3D Content Creation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[86] Ricky T. Q. Chen and Yaron Lipman. “Flow Matching on General Geometries”. In: *arXiv preprint arXiv:2302.03660* (2023). URL: https://arxiv.org/abs/2302.03660.

[87] Ricky T. Q. Chen et al. *Neural Ordinary Differential Equations*. 2019. arXiv: 1806.07366 [cs.LG]. URL: https://arxiv.org/abs/1806.07366.

[88] Ting Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG]. URL: https://arxiv.org/abs/2002.05709.

[89] Ting Chen et al. “SimCLRv2: A Stronger Baseline for Contrastive Learning”. In: (2020). arXiv: 2006.09882 [cs.CV].

[90] X. Chen et al. “Adaptive Image Transformer for One-Shot Object Detection”. In: *arXiv preprint arXiv:2103.?????* Please update with the exact venue and identifier if available. 2021.

[91] Xinlei Chen et al. “Rethinking Video Representation Learning via Per-Clip Supervision”. In: *NeurIPS*. 2021.

[92] Xinlei Chen and Kaiming He. “Exploring Simple Siamese Representation Learning”. In: *CVPR* (2021).

[93] Xinlei Chen, Saining Xie, and Kaiming He. *An Empirical Study of Training Self-Supervised Vision Transformers*. 2021. arXiv: 2104.02057 [cs.CV]. URL: https://arxiv.org/abs/2104.02057.

[94] Xinlei Chen, Saining Xie, and Kaiming He. “An Empirical Study of Training Self-Supervised Vision Transformers”. In: *arXiv preprint arXiv:2104.02057* (2021). arXiv: 2104.02057 [cs.CV]. URL: https://arxiv.org/abs/2104.02057.

[95] Xinlei Chen et al. *Improved Baselines with Momentum Contrastive Learning*. 2020. arXiv: 2003.04297 [cs.CV]. URL: https://arxiv.org/abs/2003.04297.

[96] Yen-Chun Chen et al. “UNITER: Learning Universal Image-Text Representations”. In: *ECCV*. 2020. URL: https://arxiv.org/abs/1909.11740.

[97] Yu Chen and Gim Hee Lee. “Deep Bundle-Adjusting Generalizable Neural Radiance Fields”. In: *arXiv:2306.02279* (2023). URL: https://arxiv.org/abs/2306.02279.

[98] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. *Per-Pixel Classification is Not All You Need for Semantic Segmentation*. 2021. arXiv: 2107.06278 [cs.CV]. URL: https://arxiv.org/abs/2107.06278.

[99] Bowen Cheng et al. “Masked-Attention Mask Transformer for Universal Image Segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 1290–1299. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Cheng_Masked-Attention_Mask_Transformer_for_Universal_Image_Segmentation_CVPR_2022_paper.html.

[100] Bowen Cheng et al. “YOLO-World: Real-Time Open-Vocabulary Object Detection”. In: *arXiv preprint arXiv:2401.17270* (2024). URL: https://arxiv.org/abs/2401.17270.

[101] Ho Kei Cheng et al. “Putting the Object Back into Video Object Segmentation”. In: *CVPR*. 2024.

[102] Ming-Ming Cheng et al. “BING: Binarized normed gradients for objectness estimation at 300fps”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 3286–3293.

[103] Zesen Cheng et al. “VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs”. In: *arXiv preprint arXiv:2406.07476* (2024). URL: https://arxiv.org/abs/2406.07476.

[104] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1251–1258. DOI: 10.1109/CVPR.2017.195.

[105] Junyoung Chung et al. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. 2014. arXiv: 1412.3555 [cs.NE]. URL: https://arxiv.org/abs/1412.3555.

[106] Özgün Çiçek et al. *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation*. 2016. arXiv: 1606.06650 [cs.CV].

[107] Aidan Clark, Jeff Donahue, and Karen Simonyan. *Adversarial Video Generation on Complex Datasets*. 2019. arXiv: 1907.06571 [cs.CV]. URL: https://arxiv.org/abs/1907.06571.

[108] Djork-Arné Clevert. "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv preprint arXiv:1511.07289* (2015).

[109] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. 2016. arXiv: 1511.07289 [cs.LG]. URL: https://arxiv.org/abs/1511.07289.

[110] Joseph Paul Cohen, Michael Luck, and Sina Honari. "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2018.

[111] Ekin D Cubuk et al. "AutoAugment: Learning Augmentation Policies from Data". In: *CVPR*. 2019.

[112] Ekin D. Cubuk et al. "RandAugment: Practical Data Augmentation with No Separate Search". In: *CVPR Workshops*. 2020.

[113] Angela Dai et al. "ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes". In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[114] Wenliang Dai et al. *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning*. 2023. arXiv: 2305.06500 [cs.CV]. URL: https://arxiv.org/abs/2305.06500.

[115] Google DeepMind. "Gemini: A Family of Highly Capable Multimodal Models". In: *arXiv preprint arXiv:2312.11805* (2023).

[116] Mostafa Dehghani et al. *Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution*. 2023. arXiv: 2307.06304 [cs.CV]. URL: https://arxiv.org/abs/2307.06304.

[117] Mitchell Deitke et al. *Objaverse: A Universe of Annotated 3D Objects*. https://objaverse.allenai.org/. 2023.

[118] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[119] Jacob Devlin, Bharath Gupta, and Ross Girshick. "Exploring Nearest Neighbor Approaches for Image Captioning". In: *arXiv preprint arXiv:1505.02044* (2015).

[120] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[121] Terrance DeVries and Graham W. Taylor. *Improved Regularization of Convolutional Neural Networks with Cutout*. 2017. arXiv: 1708.04552 [cs.CV]. URL: https://arxiv.org/abs/1708.04552.

[122] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG]. URL: https://arxiv.org/abs/2105.05233.

[123] Prafulla Dhariwal and Alexander Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *NeurIPS*. 2021. arXiv: 2105.05233 [cs.LG].

[124] Henghui Ding et al. "MOSE: A New Dataset for Video Object Segmentation in Complex Scenes". In: *arXiv preprint arXiv:2302.01872* (2023).

[125] Shuangrui Ding et al. "SAM2Long: Enhancing SAM 2 for Long Video Segmentation with a Training-Free Memory Tree". In: *arXiv preprint arXiv:2410.16268* (2024). URL: https://arxiv.org/abs/2410.16268.

[126] Xiaohan Ding et al. "ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetry". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1911–1920.

[127] Xiaohan Ding et al. *RepVGG: Making VGG-style ConvNets Great Again*. 2021. arXiv: 2101.03697 [cs.CV]. URL: https://arxiv.org/abs/2101.03697.

[128] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. *Density estimation using Real NVP*. 2017. arXiv: 1605.08803 [cs.LG]. URL: https://arxiv.org/abs/1605.08803.

[129] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction". In: *International Conference on Computer Vision (ICCV)*. 2015.

[130] Jeff Donahue and Karen Simonyan. "Large Scale Adversarial Representation Learning". In: *NeurIPS*. 2019. arXiv: 1907.02544 [cs.LG].

[131] Jeff Donahue et al. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 2625–2634. DOI: 10.1109/CVPR.2015.7298878.

[132] et al. Doron Levy. "Deep Learning in Medical Imaging: Revolutionizing Radiology". In: *Nature Medicine* (2016).

[133] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[134] Danny Driess et al. "PaLM-E: An Embodied Multimodal Language Model". In: *arXiv preprint arXiv:2303.03378* (2023).

[135] Quang Vu Duc et al. "Self-Knowledge Distillation: An Efficient Approach for Falling Detection". In: *Artificial Intelligence in Data and Big Data Processing*. Ed. by Ngoc Hoang Thanh Dang et al. https://github.com/vdquang1991/self_KD_falling_detection. Cham: Springer International Publishing, 2022, pp. 369–380. DOI: 10.1007/978-3-030-97610-1_29.

[136] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. "A learned representation for artistic style". In: *International Conference on Learning Representations (ICLR)*. 2017.

[137] Emilien Dupont et al. "Equivariant Neural Rendering". In: *International Conference on Machine Learning (ICML)*. 2020.

[138] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations". In: *ICCV*. 2021.

[139] Dylan Ebert. *Introduction to 3D Gaussian Splatting*. Hugging Face Blog; published September 18, 2023. URL: https://huggingface.co/blog/gaussian-splatting.

[140] Alexei A. Efros and Thomas K. Leung. "Texture Synthesis by Non-Parametric Sampling". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1999, pp. 1033–1038.

[141] David Eigen, Christian Puhrsch, and Rob Fergus. *Depth Map Prediction from a Single Image using a Multi-Scale Deep Network*. 2014. arXiv: 1406.2283 [cs.CV]. URL: https://arxiv.org/abs/1406.2283.

[142] Ron Eldan and Ohad Shamir. "The power of depth for feedforward neural networks". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)* (2016), pp. 907–916.

[143] Kemal Erdem. "Understanding Region of Interest - Part 2 (RoI Align)". In: *https://erdem.pl* (2020). URL: https://erdem.pl/2020/02/understanding-region-of-interest-part-2-ro-i-align.

[144] Linus Ericsson, Henry Gouk, and Timothy Hospedales. "Why Do Self-Supervised Models Transfer? Investigating the Impact of Invariance on Downstream Tasks". In: *BMVC*. 2021.

[145] Aleksandr Ermolov et al. "Whitening and Coloring Batch Transform for GANs". In: *CVPR* (2021).

[146] SM Eslami et al. "Neural Scene Representation and Rendering". In: *Science* 360.6394 (2018), pp. 1204–1210.

[147] Stefano Esposito et al. "KiloNeuS: A Versatile Neural Implicit Surface Representation for Real-Time Rendering". In: *arXiv preprint arXiv:2211.14183*. 2022.

[148] Patrick Esser, Robin Rombach, and Björn Ommer. *Taming Transformers for High-Resolution Image Synthesis*. 2021. arXiv: 2012.09841 [cs.CV]. URL: https://arxiv.org/abs/2012.09841.

[149] Patrick Esser et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". In: *arXiv preprint arXiv:2403.03206* (2024). URL: https://arxiv.org/abs/2403.03206.

[150] Mark Everingham et al. "The PASCAL visual object classes (VOC) challenge". In: *International Journal of Computer Vision* 88.2 (2010), pp. 303–338. DOI: 10.1007/s11263-009-0275-4.

[151] Haoqi Fan et al. "Multiscale Vision Transformers". In: *ICCV*. 2021.

[152] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. "A Point Set Generation Network for 3D Object Reconstruction from a Single Image". In: *CVPR*. 2017. arXiv: 1612.00603 [cs.CV].

[153] Christoph Feichtenhofer. "X3D: Expanding Architectures for Efficient Video Recognition". In: *CVPR*. 2020.

[154] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. "What Have We Learned from Deep Representations for Action Recognition?" In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7844–7853. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Feichtenhofer_What_Have_We_CVPR_2018_paper.html.

[155] Christoph Feichtenhofer et al. “A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning”. In: *CVPR*. 2021.

[156] Christoph Feichtenhofer et al. “Deep Insights into Convolutional Networks for Video Recognition”. In: *International Journal of Computer Vision (IJCV)* 128.2 (2019), pp. 420–439. DOI: 10.1007/s11263-019-01247-2. URL: https://link.springer.com/article/10.1007/s11263-019-01247-2.

[157] Christoph Feichtenhofer et al. “Masked Autoencoders as Spatiotemporal Learners”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[158] Christoph Feichtenhofer et al. “SlowFast Networks for Video Recognition”. In: *ICCV*. 2019.

[159] Martin A Fischler and Robert A Elschlager. “The representation and matching of pictorial structures”. In: *IEEE Transactions on Computers* C-22.1 (1973), pp. 67–92. DOI: 10.1109/T-C.1973.223602.

[160] Sara Fridovich-Keil et al. *Plenoxels: Radiance Fields without Neural Networks*. 2021. arXiv: 2112.05131 [cs.CV]. URL: https://arxiv.org/abs/2112.05131.

[161] Tsu-Jui Fu, Linjie Li, et al. “VIOLET: End-to-End Video-Language Transformers with Masked Visual-token Modeling”. In: *arXiv:2111.12681*. 2021.

[162] Yifan Fu et al. “COLMAP-Free 3D Gaussian Splatting”. In: *CVPR*. 2024.

[163] Kunihiko Fukushima. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. In: *Biological Cybernetics* 36.4 (1980), pp. 193–202.

[164] Samir Yitzhak Gadre et al. *DataComp: In search of the next generation of multimodal datasets*. 2023. arXiv: 2304.14108 [cs.CV]. URL: https://arxiv.org/abs/2304.14108.

[165] Chen Gao et al. “NeRF-Editing: Geometry Editing of Neural Radiance Fields”. In: *arXiv preprint arXiv:2205.01627* (2022).

[166] Jiyang Gao et al. “TALL: Temporal Activity Localization via Language Query”. In: *ICCV*. 2017.

[167] Peng Gao et al. *LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model*. 2023. arXiv: 2304.15010 [cs.CV]. URL: https://arxiv.org/abs/2304.15010.

[168] Itai Gat et al. *Discrete Flow Matching*. 2024. arXiv: 2407.15595 [cs.LG]. URL: https://arxiv.org/abs/2407.15595.

[169] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “Image Style Transfer Using Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2414–2423.

[170] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. *Texture Synthesis Using Convolutional Neural Networks*. 2015. arXiv: 1505.07376 [cs.CV]. URL: https://arxiv.org/abs/1505.07376.

[171] Golnaz Ghiasi et al. *Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation*. 2021. arXiv: 2012.07177 [cs.CV]. URL: https://arxiv.org/abs/2012.07177.

[172] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised Representation Learning by Predicting Image Rotations". In: *International Conference on Learning Representations (ICLR)*. 2018.

[173] Rohit Girdhar et al. "ImageBind: One Embedding Space to Bind Them All". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[174] Rohit Girdhar et al. "OmniMAE: Single Model Masked Pretraining on Images and Videos". In: *arXiv preprint arXiv:2206.08356* (2023).

[175] Ross Girshick. *Fast R-CNN*. 2015. arXiv: 1504.08083 [cs.CV]. URL: https://arxiv.org/abs/1504.08083.

[176] Ross Girshick et al. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[177] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. *Mesh R-CNN*. 2020. arXiv: 1906.02739 [cs.CV]. URL: https://arxiv.org/abs/1906.02739.

[178] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Vol. 9. Proceedings of Machine Learning Research. PMLR, 2010, pp. 249–256.

[179] Gabriel Goh et al. "Multimodal neurons in artificial neural networks". In: *Distill* 6.3 (2021), e30.

[180] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[181] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *ICLR* (2015). arXiv: 1412.6572 [stat.ML].

[182] Jianping Gou et al. "Knowledge Distillation: A Survey". In: *CoRR* abs/2006.05525 (2020). arXiv: 2006.05525. URL: https://arxiv.org/abs/2006.05525.

[183] Benjamin Graham. *Fractional Max-Pooling*. 2015. arXiv: 1412.6071 [cs.CV]. URL: https://arxiv.org/abs/1412.6071.

[184] Benjamin Graham and Laurens van der Maaten. *Submanifold Sparse Convolutional Networks*. 2017. arXiv: 1706.01307 [cs.NE]. URL: https://arxiv.org/abs/1706.01307.

[185] Will Grathwohl et al. "FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models". In: *International Conference on Learning Representations (ICLR)*. 2019. URL: https://openreview.net/forum?id=rJxFNnCqFQ.

[186] Kristen Grauman et al. "Ego4D: Around the World in 3,000 Hours of Egocentric Video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 18995–19012. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

[187] Karol Gregor et al. *DRAW: A Recurrent Neural Network For Image Generation*. 2015. arXiv: 1502.04623 [cs.CV].

[188] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[189] Amos Gropp et al. *Implicit Geometric Regularization for Learning Shapes*. 2020. arXiv: 2002.10099 [cs.LG]. URL: https://arxiv.org/abs/2002.10099.

[190] Albert Gu et al. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces”. In: *arXiv preprint* arXiv:2312.00752 (2023). URL: https://arxiv.org/abs/2312.00752.

[191] Chunhui Gu et al. “AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6047–6056. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Gu_AVA_A_Video_CVPR_2018_paper.html.

[192] Jiatao Gu et al. “Non-autoregressive neural machine translation”. In: *International Conference on Learning Representations (ICLR)*. 2018.

[193] Xiuye Gu et al. *Open-vocabulary Object Detection via Vision and Language Knowledge Distillation*. 2022. arXiv: 2104.13921 [cs.CV]. URL: https://arxiv.org/abs/2104.13921.

[194] Ishaan Gulrajani et al. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG]. URL: https://arxiv.org/abs/1704.00028.

[195] Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 1321–1330.

[196] Agrim Gupta, Piotr Dollar, and Ross Girshick. “LVIS: A Dataset for Large Vocabulary Instance Segmentation”. In: *CVPR* (2019).

[197] Agrim Gupta, Piotr Dollar, Ross Girshick, et al. *ODinW: Evaluating and Harnessing Object Detection in the Wild*. 2022. arXiv: 2206.14268 [cs.CV]. URL: https://arxiv.org/abs/2206.14268.

[198] Agrim Gupta et al. “Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 2255–2264.

[199] Raia Hadsell, Sumit Chopra, and Yann LeCun. “Dimensionality Reduction by Learning an Invariant Mapping”. In: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. IEEE Computer Society, 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[200] Kai Han et al. *GhostNet: More Features from Cheap Operations*. 2020. arXiv: 1911.11907 [cs.CV]. URL: https://arxiv.org/abs/1911.11907.

[201] Tengda Han, Weidi Xie, and Andrew Zisserman. “Memory-augmented Dense Predictive Coding for Video Representation Learning”. In: *ECCV*. 2020.

[202] Ayaan Haque et al. *Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions*. 2023. arXiv: 2303.12789 [cs.CV]. URL: https://arxiv.org/abs/2303.12789.

[203] Chris Harris and Mike Stephens. “A combined corner and edge detector”. In: *Proceedings of the Alvey Vision Conference*. Citeseer. 1988, pp. 147–151.

[204] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. *LoRA+: Efficient Low Rank Adaptation of Large Models*. 2024. arXiv: 2402.12354 [cs.LG]. URL: https://arxiv.org/abs/2402.12354.

[205] Kaiming He, Ross Girshick, and Piotr Dollár. *Rethinking ImageNet Pre-training*. 2018. arXiv: 1811.08883 [cs.CV]. URL: https://arxiv.org/abs/1811.08883.

[206] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: (2016), pp. 770–778.

[207] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[208] Kaiming He et al. "Identity mappings in deep residual networks". In: *arXiv preprint arXiv:1603.05027* (2016).

[209] Kaiming He et al. "Mask R-CNN". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 42.2 (2020). Originally published in 2017 in ICCV, pp. 386–397. DOI: 10.1109/TPAMI.2018.2844175.

[210] Kaiming He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16000–16009.

[211] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[212] Tong He et al. *Bag of Tricks for Image Classification with Convolutional Neural Networks*. 2018. arXiv: 1812.01187 [cs.CV]. URL: https://arxiv.org/abs/1812.01187.

[213] Peter Hedman et al. "Deep Blending for Free-Viewpoint Image-Based Rendering". In: *SIGGRAPH Asia*. 2018.

[214] Olivier J. Hénaff et al. *Data-Efficient Image Recognition with Contrastive Predictive Coding*. 2020. arXiv: 1905.09272 [cs.CV]. URL: https://arxiv.org/abs/1905.09272.

[215] Dan Hendrycks and Kevin Gimpel. "Gaussian error linear units (gelus)". In: *arXiv preprint arXiv:1606.08415* (2016).

[216] Byeongho Heo et al. *Rotary Position Embedding for Vision Transformer*. 2024. arXiv: 2403.13298 [cs.CV]. URL: https://arxiv.org/abs/2403.13298.

[217] Amir Hertz et al. *Prompt-to-Prompt Image Editing with Cross Attention Control*. 2022. arXiv: 2208.01626 [cs.CV]. URL: https://arxiv.org/abs/2208.01626.

[218] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems* 30 (2017). URL: https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[219] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML]. URL: https://arxiv.org/abs/1503.02531.

[220] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507. DOI: 10.1126/science.1127647.

[221] Ester Hlav. “Kaiming He Initialization in Neural Networks — Math Proof”. In: *Medium* (2023). https://medium.com/towards-data-science/kaiming-he-initialization-in-neural-networks-math-proof-73b9a0d845c4.

[222] Ester Hlav. “Xavier Glorot Initialization in Neural Networks: Math Proof”. In: *Medium* (2023). https://medium.com/towards-data-science/xavier-glorot-initialization-in-neural-networks-math-proof-4682bf5c6ec3.

[223] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. URL: https://arxiv.org/abs/2006.11239.

[224] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: 2207.12598 [cs.LG]. URL: https://arxiv.org/abs/2207.12598.

[225] Jonathan Ho et al. *Cascaded Diffusion Models for High Fidelity Image Generation*. 2021. arXiv: 2106.15282 [cs.CV]. URL: https://arxiv.org/abs/2106.15282.

[226] Quan Hoang et al. “MGAN: Training generative adversarial nets with multiple generators”. In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://openreview.net/forum?id=SJ8wKh0c-.

[227] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[228] Peter Holderrieth et al. “Generator Matching: Generative Modeling with Arbitrary Markov Processes”. In: *arXiv preprint arXiv:2410.20587* (2024). URL: https://arxiv.org/abs/2410.20587.

[229] Andrew Howard et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 484–492. DOI: 10.48550/arXiv.1704.04861.

[230] Andrew Howard et al. “Searching for MobileNetV3”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019, pp. 1314–1324. DOI: 10.1109/ICCV.2019.00140.

[231] Jeremy Howard and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification”. In: *ACL* (2018).

[232] Ting-I Hsieh et al. *One-Shot Object Detection with Co-Attention and Co-Excitation*. 2019. arXiv: 1911.12529 [cs.CV]. URL: https://arxiv.org/abs/1911.12529.

[233] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[234] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-Excitation Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745. URL: https://arxiv.org/abs/1709.01507.

[235] Qianjiang Hu et al. “Adversarial Contrast for Efficient Learning of Unsupervised Representations from Self-Trained Negative Adversaries”. In: *CVPR*. 2021. arXiv: 2011.08435 [cs.LG].

[236] Weiming Hu et al. *SiamMask: A Framework for Fast Online Object Tracking and Segmentation*. 2022. arXiv: 2207.02088 [cs.CV]. URL: https://arxiv.org/abs/2207.02088.

[237] Wenbo Hu et al. *Tri-MipRF: Tri-Mip Representation for Efficient Anti-Aliasing Neural Radiance Fields*. 2023. arXiv: 2307.11335 [cs.CV]. URL: https://arxiv.org/abs/2307.11335.

[238] Xuefeng Hu et al. *ReCLIP: Refine Contrastive Language Image Pre-Training with Source Free Domain Adaptation*. 2023. arXiv: 2308.03793 [cs.CV]. URL: https://arxiv.org/abs/2308.03793.

[239] Ying Hu, Chenyi Zhuang, and Pan Gao. "DiffuseST: Unleashing the Capability of the Diffusion Model for Style Transfer". In: *Proceedings of the 6th ACM International Conference on Multimedia in Asia*. 2024, pp. 1–1.

[240] Yushi Hu et al. "PromptCap: Prompt-guided Task-aware Image Captioning". In: *CVPR*. 2023.

[241] Zhengdong Hu et al. "DAC-DETR: Divide the Attention Layers and Conquer". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023. URL: https://openreview.net/forum?id=8JMexYVcXB.

[242] Gao Huang et al. *Deep Networks with Stochastic Depth*. 2016. arXiv: 1603.09382 [cs.LG]. URL: https://arxiv.org/abs/1603.09382.

[243] Gao Huang et al. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV]. URL: https://arxiv.org/abs/1608.06993.

[244] Hai Huang and Randall Balestriero. "ALLoRA: Adaptive Learning Rate Mitigates LoRA Fatal Flaws". In: *arXiv preprint arXiv:2410.09692* (2024).

[245] Rongjie Huang et al. *AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head*. 2023. arXiv: 2304.12995 [cs.CL]. URL: https://arxiv.org/abs/2304.12995.

[246] Shihua Huang et al. *Real-Time Object Detection Meets DINOv3*. 2025. arXiv: 2509.20787 [cs.CV]. URL: https://arxiv.org/abs/2509.20787.

[247] Xiao Shi Huang et al. "Improving Transformer Optimization Through Better Initialization". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4475–4483. URL: https://proceedings.mlr.press/v119/huang20f.html.

[248] Xun Huang and Serge Belongie. "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 1501–1510. URL: https://arxiv.org/abs/1703.06868.

[249] David H. Hubel and Torsten N. Wiesel. "Receptive fields of single neurones in the cat's striate cortex". In: *The Journal of Physiology* 148.3 (1959), pp. 574–591. DOI: 10.1113/jphysiol.1959.sp006308.

[250] Jonathan Hui. *StyleGAN and StyleGAN2: Learn to generate and control images*. https://jonathan-hui.medium.com/gan-stylegan-stylegan2-479bdf256299. Accessed: 2025-04-12. 2020.

[251] Becoming Human. "All About Normalization". In: *Becoming Human* (2018). URL: https://becominghuman.ai/all-about-normalization-6ea79e70894b.

[252] IDEA-Research. *Grounded SAM 2: Ground and Track Anything in Videos*. https://github.com/IDEA-Research/Grounded-SAM-2. Code repository; supports text-grounded detection, segmentation, and tracking with SAM 2. 2024.

[253] Intellindust AI Lab. *DEIMv2: Real-Time Object Detection Meets DINOv3 (Code)*. 2025. URL: https://github.com/Intellindust-AI-Lab/DEIMv2.

[254] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 2015, pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html.

[255] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5967–5976.

[256] Jerome Jabri, Jean-Baptiste Alayrac, and Andrew Zisserman. "STC: Self-Supervised Tracking for Correspondence". In: *European Conference on Computer Vision (ECCV)*. 2020.

[257] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018. URL: https://arxiv.org/abs/1806.07572.

[258] Ajay Jain et al. "Zero-Shot Text-Guided Object Generation with Dream Fields". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 867–876. DOI: 10.1109/CVPR52688.2022.00095. URL: https://arxiv.org/abs/2112.01455.

[259] Jitesh Jain et al. "OneFormer: One Transformer To Rule Universal Image Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 2989–2998. arXiv: 2211.06220 [cs.CV]. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Jain_OneFormer_One_Transformer_To_Rule_Universal_Image_Segmentation_CVPR_2023_paper.html.

[260] Clement Jambon et al. *NeRFshop: Interactive Editing of Neural Radiance Fields*. 2023. arXiv: 2303.11537 [cs.GR]. URL: https://arxiv.org/abs/2303.11537.

[261] Eric Jang, Shixiang Gu, and Ben Poole. *Categorical Reparameterization with Gumbel-Softmax*. 2017. arXiv: 1611.01144 [stat.ML]. URL: https://arxiv.org/abs/1611.01144.

[262] Rasmus Jensen et al. "Large scale multi-view stereopsis evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 406–413.

[263] Shuiwang Ji et al. "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.1 (2010), pp. 220–233. DOI: 10.1109/TPAMI.2009.23.

[264] Chao Jia et al. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. 2021. arXiv: 2102.05918 [cs.CV]. URL: https://arxiv.org/abs/2102.05918.

[265] Chao Jia et al. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. 2021. arXiv: 2102.05918 [cs.CV]. URL: https://arxiv.org/abs/2102.05918.

[266] Yu-Gang Jiang et al. “THUMOS Challenge: Action Recognition with a Large Number of Classes”. In: *ECCV Workshop on Large-Scale Video Classification*. 2014.

[267] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. “GeoNeRF: Generalizing NeRF with Geometry Priors”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. arXiv: 2111.13539. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Johari_GeoNeRF_Generalizing_NeRF_With_Geometry_Priors_CVPR_2022_paper.pdf.

[268] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual Losses for Real-Time Style Transfer and Super-Resolution”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 694–711.

[269] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. *DenseCap: Fully Convolutional Localization Networks for Dense Captioning*. 2015. arXiv: 1511.07571 [cs.CV]. URL: https://arxiv.org/abs/1511.07571.

[270] Justin Johnson et al. *Inferring and Executing Programs for Visual Reasoning*. 2017. arXiv: 1705.03633 [cs.CV]. URL: https://arxiv.org/abs/1705.03633.

[271] Damjan Kalajdzievski. “A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA”. In: *arXiv preprint arXiv:2312.03732* (2023).

[272] Aishwarya Kamath et al. “MDETR: Modulated Detection for End-to-End Multi-Modal Understanding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1780–1790.

[273] Minguk Kang et al. *Scaling up GANs for Text-to-Image Synthesis*. 2023. arXiv: 2303.05511 [cs.CV]. URL: https://arxiv.org/abs/2303.05511.

[274] Andrej Karpathy and Li Fei-Fei. “Deep Visual-Semantic Alignments for Generating Image Descriptions”. In: *Computer Vision and Pattern Recognition (CVPR)* (2015).

[275] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. *Visualizing and Understanding Recurrent Networks*. 2015. arXiv: 1506.02078 [cs.LG]. URL: https://arxiv.org/abs/1506.02078.

[276] Andrej Karpathy et al. “Large-scale Video Classification with Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1725–1732. DOI: 10.1109/CVPR.2014.223.

[277] Andrej Karpathy et al. “Large-scale video classification with convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2014, pp. 1725–1732.

[278] Tero Karras, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4401–4410. DOI: 10.1109/CVPR.2019.00453. URL: https://arxiv.org/abs/1812.04948.

[279] Tero Karras et al. “Alias-Free Generative Adversarial Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021. URL: https://arxiv.org/abs/2106.12423.

[280] Tero Karras et al. “Analyzing and Improving the Image Quality of StyleGAN”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8110–8119. URL: https://arxiv.org/abs/1912.04958.

[281] Tero Karras et al. “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://arxiv.org/abs/1710.10196.

[282] Will Kay et al. “The Kinetics Human Action Video Dataset”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017. arXiv: 1705.06950. URL: https://arxiv.org/abs/1705.06950.

[283] Amirhossein Kazemnejad. “Transformer Architecture: The Positional Encoding”. In: *kazemnejad.com* (2019). URL: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/.

[284] Guolin Ke, Di He, et al. *Rethinking Positional Encoding in Language Pre-training*. 2021. arXiv: 2006.15595 [cs.CL].

[285] Lei Ke et al. “Segment Anything in High Quality”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023. arXiv: 2306.01567 [cs.CV]. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/5f828e38160f31935cfe9f67503ad17c-Paper-Conference.pdf.

[286] Bernhard Kerbl, Georgios Kopanas, and George Drettakis. “Gaussian Surfels for Real-Time Rendering of Point Clouds”. In: *arXiv:2404.19702* (2024).

[287] Bernhard Kerbl et al. *3D Gaussian Splatting for Real-Time Radiance Field Rendering*. 2023. arXiv: 2308.04079 [cs.GR]. URL: https://arxiv.org/abs/2308.04079.

[288] Justin Kerr et al. “LERF: Language Embedded Radiance Fields”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. arXiv: 2303.09553 [cs.CV]. URL: https://arxiv.org/abs/2303.09553.

[289] Nitish Shirish Keskar et al. *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*. 2017. arXiv: 1609.04836 [cs.LG]. URL: https://arxiv.org/abs/1609.04836.

[290] Valentin Khrulkov and Ivan Oseledets. “Geometry Score: A Method for Comparing Generative Adversarial Networks”. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018, pp. 2621–2629. URL: http://proceedings.mlr.press/v80/khrulkov18a/khrulkov18a.pdf.

[291] Yannic Kilcher. *Flow Matching: A Unified Framework for Generative Models*. YouTube video. 2022. URL: https://www.youtube.com/watch?v=7NNxK3CqaDk.

[292] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. 2014. URL: https://arxiv.org/abs/1312.6114.

[293] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[294] Diederik P. Kingma and Prafulla Dhariwal. *Glow: Generative Flow with Invertible 1x1 Convolutions*. 2018. arXiv: 1807.03039 [stat.ML]. URL: https://arxiv.org/abs/1807.03039.

[295] Alexander Kirillov et al. “PointRend: Image Segmentation as Rendering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9799–9808. DOI: 10.1109/CVPR42600.2020.00982. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Kirillov_PointRend_Image_Segmentation_as_Rendering_CVPR_2020_paper.html.

[296] Alexander Kirillov et al. “Segment Anything”. In: (2023). arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643.

[297] Alexander Kirillov et al. “Segment Anything”. In: *arXiv preprint arXiv:2304.02643* (2023).

[298] Günter Klambauer et al. “Self-normalizing neural networks”. In: *Advances in neural information processing systems* 30 (2017).

[299] Arno Knapitsch et al. “Tanks and temples: benchmarking large-scale scene reconstruction”. In: *ACM Trans. Graph.* 36.4 (July 2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073599. URL: https://doi.org/10.1145/3072959.3073599.

[300] Sebastian Koch et al. “ABC: A Big CAD Model Dataset For Geometric Deep Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[301] Jonas Kohler et al. “Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 806–815.

[302] Nikita Kornilov et al. *Optimal Flow Matching: Learning Straight Trajectories in Just One Step*. 2024. arXiv: 2403.13117 [cs.LG]. URL: https://arxiv.org/abs/2403.13117.

[303] Jonathan Krause et al. “3D Object Representations for Fine-Grained Categorization”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2013, pp. 554–561. DOI: 10.1109/ICCVW.2013.77. URL: https://ieeexplore.ieee.org/document/6755945.

[304] Ranjay Krishna et al. “Dense-Captioning Events in Videos”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017.

[305] Raghuraman Krishnamoorthi. *Quantizing deep convolutional networks for efficient inference: A whitepaper*. 2018. arXiv: 1806.08342 [cs.LG]. URL: https://arxiv.org/abs/1806.08342.

[306] Alex Krizhevsky and Geoffrey Hinton. “Learning Multiple Layers of Features from Tiny Images”. In: *Technical Report, University of Toronto*. 2009.

[307] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 25. 2012, pp. 1097–1105.

[308] Tejas D. Kulkarni et al. “Deep Convolutional Inverse Graphics Network”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2015, pp. 2539–2547. URL: https://papers.nips.cc/paper_files/paper/2015/hash/09d3b6c4c2c8f1d159d905f9b3d6fe4a-Abstract.html.

[309] Ananya Kumar et al. *Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution*. 2022. arXiv: 2202.10054 [cs.LG]. URL: https://arxiv.org/abs/2202.10054.

[310] Rohit Kundu. *The Beginner's Guide to Contrastive Learning*. V7 Labs Blog. Accessed: 2025-06-15. May 2022.

[311] Weicheng Kuo et al. *F-VLM: Open-Vocabulary Object Detection upon Frozen Vision and Language Models*. 2023. arXiv: 2209.15639 [cs.CV]. URL: https://arxiv.org/abs/2209.15639.

[312] Weicheng Kuo et al. *FindIt: Generalized Localization with Natural Language Queries*. 2022. arXiv: 2203.17273 [cs.CV]. URL: https://arxiv.org/abs/2203.17273.

[313] Hyunsu Kwon et al. "Diffusion-GAN: Training GANs with Diffusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[314] Zihang Lai, Erika Lu, and Weidi Xie. "MAST: A Memory-Augmented Self-Supervised Tracker". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[315] Magdalena Lazova et al. *Control-NeRF: Editable Feature Volumes for Scene Rendering and Manipulation*. 2022. arXiv: 2204.10850 [cs.CV]. URL: https://arxiv.org/abs/2204.10850.

[316] Yann LeCun et al. "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[317] Christian Ledig et al. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. 2017. arXiv: 1609.04802 [cs.CV].

[318] Doyup Lee et al. "Draft-and-Revise: Effective Image Generation with Contextual RQ-Transformer". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 30127–30138. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c276c3303c0723c83a43b95a44a1fcbf-Paper-Conference.pdf.

[319] Kuang-Huei Lee et al. *Compressive Visual Representations*. 2021. arXiv: 2109.12909 [cs.LG]. URL: https://arxiv.org/abs/2109.12909.

[320] Pil Sun Lee et al. "From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[321] Sangwoo Lee et al. "StyleGAN-T: Unlocking the Power of GANs with Transformer Backbones". In: *CVPR* (2023). arXiv: 2206.00260 [cs.CV].

[322] Jie Lei et al. "Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling". In: *CVPR*. 2021.

[323] Jie Lei, Tamara L. Berg, and Mohit Bansal. "Q&A Highlights: Detecting Moments and Highlights from Text Queries". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[324] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2020. arXiv: 1910.13461 [cs.CL].

[325] Bo Li et al. *LLaVA-OneVision: Easy Visual Task Transfer*. 2024. arXiv: 2408.03326 [cs.CV]. URL: https://arxiv.org/abs/2408.03326.

[326] Boyi Li et al. “Language-Driven Semantic Segmentation”. In: *International Conference on Learning Representations (ICLR)*. 2022. URL: https://arxiv.org/abs/2201.03546.

[327] Feng Li et al. *DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection*. 2022. arXiv: 2203.03605 [cs.CV]. URL: https://arxiv.org/abs/2203.03605.

[328] Feng Li et al. “DN-DETR: Accelerate DETR Training with Decoupled Denoising Anchor Boxes”. In: *European Conference on Computer Vision (ECCV)*. 2022.

[329] Feng Li et al. *LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models*. 2024. arXiv: 2407.07895 [cs.CV]. URL: https://arxiv.org/abs/2407.07895.

[330] Feng Li et al. “Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation”. In: (2022). arXiv: 2206.02777 [cs.CV]. URL: https://arxiv.org/abs/2206.02777.

[331] Hao Li et al. *Visualizing the Loss Landscape of Neural Nets*. 2018. arXiv: 1712.09913 [cs.LG]. URL: https://arxiv.org/abs/1712.09913.

[332] Junnan Li et al. “Align Before Fuse: Vision and Language Representation Learning With Momentum Distillation”. In: *NeurIPS*. 2021. URL: https://arxiv.org/abs/2107.07651.

[333] Junnan Li et al. “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”. In: *arXiv preprint arXiv:2301.12597* (2023).

[334] Junnan Li et al. “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation”. In: *International Conference on Machine Learning (ICML)*. 2022. URL: https://arxiv.org/abs/2201.12086.

[335] KunChang Li et al. *VideoChat: Chat-Centric Video Understanding*. 2024. arXiv: 2305.06355 [cs.CV]. URL: https://arxiv.org/abs/2305.06355.

[336] Kunchang Li et al. “MVBench: A Comprehensive Multi-modal Video Understanding Benchmark”. In: *CVPR*. 2024.

[337] Kunchang Li et al. *Unmasked Teacher: Towards Training-Efficient Video Foundation Models*. 2024. arXiv: 2303.16058 [cs.CV]. URL: https://arxiv.org/abs/2303.16058.

[338] Liunian Harold Li et al. *Grounded Language-Image Pre-Training*. 2022. arXiv: 2112.03857 [cs.CV]. URL: https://arxiv.org/abs/2112.03857.

[339] Xian Li et al. “Temporal Aggregation Network for Video Recognition”. In: *CVPR*. 2021.

[340] Xiang Li et al. “Understanding and Improving Layer Normalization”. In: *ICLR* (2022).

[341] Xiujun Li, Xi Yin, and et al. “OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks”. In: *ECCV*. 2020. URL: https://arxiv.org/abs/2004.06165.

[342] Yali Li et al. “TEA: Temporal Excitation and Aggregation for Action Recognition”. In: *CVPR*. 2020.

[343] Yang Li et al. “Learnable Fourier Features for Multi-dimensional Spatial Positional Encoding”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021. URL: https://openreview.net/forum?id=ROh3NUMao_U.

[344] Yanghao Li et al. “Improved Multiscale Vision Transformers for Classification and Detection”. In: *arXiv preprint arXiv:2111.11429* (2021).

[345] Yuhui Li et al. “UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer”. In: *arXiv preprint arXiv:2211.09552* (2022). arXiv: 2211.09552 [cs.CV]. URL: https://arxiv.org/abs/2211.09552.

[346] Yuxi Li et al. “Detailed 2D-3D Joint Representation for Human Actions”. In: *ECCV*. 2020.

[347] Yuzhe Li et al. “TEINet: Towards an Efficient Architecture for Video Recognition”. In: *AAAI*. 2021.

[348] Yuzhuo Li et al. “UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning”. In: *ICLR*. 2022.

[349] Zhengqi Li et al. “Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 6498–6508. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Li_Neural_Scene_Flow_Fields_for_Space-Time_View_Synthesis_of_Dynamic_CVPR_2021_paper.html.

[350] Zhizhong Li and Derek Hoiem. “Learning without Forgetting”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 614–629. DOI: 10.1007/978-3-319-46493-0_37.

[351] Vladislav Lialin et al. *Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning*. 2024. arXiv: 2303.15647 [cs.CL]. URL: https://arxiv.org/abs/2303.15647.

[352] Chen-Hsuan Lin et al. *BARF: Bundle-Adjusting Neural Radiance Fields*. 2021. arXiv: 2104.06405 [cs.CV]. URL: https://arxiv.org/abs/2104.06405.

[353] Chen-Hsuan Lin et al. “Magic3D: High-Resolution Text-to-3D Content Creation”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023.

[354] Ji Lin, Chuang Gan, and Song Han. “TSM: Temporal Shift Module for Efficient Video Understanding”. In: *ICCV*. 2019.

[355] Jing Lin et al. *One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer*. 2023. arXiv: 2303.16160 [cs.CV]. URL: https://arxiv.org/abs/2303.16160.

[356] Tianwei Lin et al. “BMN: Boundary-Matching Network for Temporal Action Proposal Generation”. In: *ICCV*. 2019.

[357] Tianwei Lin et al. “Learning Salient Boundary Feature for Anchor-free Temporal Action Localization”. In: *CVPR*. 2021.

[358] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. “Microsoft COCO: Common objects in context”. In: *European Conference on Computer Vision (ECCV)* (2014).

[359] Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 936–944. DOI: 10.1109/CVPR.2017.106.

[360] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV]. URL: https://arxiv.org/abs/1708.02002.

[361] Yutong Lin et al. *DETR Doesn't Need Multi-Scale or Locality Design*. 2023. arXiv: 2308.01904 [cs.CV]. URL: https://arxiv.org/abs/2308.01904.

[362] Zinan Lin et al. "PacGAN: The power of two samples in generative adversarial networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 31. 2018. URL: https://papers.nips.cc/paper_files/paper/2018/hash/8313e4bde5847d539d2700627c655a3b-Abstract.html.

[363] Yaron Lipman et al. *Flow Matching Guide and Code*. 2024. arXiv: 2412.06264 [cs.LG]. URL: https://arxiv.org/abs/2412.06264.

[364] Yotam Gideon Lipman et al. "Flow Matching for Generative Modeling". In: *International Conference on Machine Learning (ICML)*. 2022. URL: https://arxiv.org/abs/2206.08791.

[365] Geert Litjens et al. "A Survey on Deep Learning in Medical Image Analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

[366] Hao Liu et al. *World Model on Million-Length Video And Language With Blockwise RingAttention*. 2025. arXiv: 2402.08268 [cs.LG]. URL: https://arxiv.org/abs/2402.08268.

[367] Haotian Liu and et al. *LLaVA-NeXT-Video: A Comprehensive Video Understanding Extension of LLaVA-NeXT*. https://github.com/LLaVA-VL/LLaVA-NeXT. Use the project URL or replace with the official arXiv entry if you track it. 2024.

[368] Haotian Liu et al. "Visual Instruction Tuning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023. URL: https://arxiv.org/abs/2304.08485.

[369] Lingjie Liu et al. "A Survey on Neural Radiance Fields". In: *Computer Graphics Forum* (2023).

[370] Lingjie Liu et al. "Neural Sparse Voxel Fields". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 15651–15663. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/b4b758962f17808746e9bb832a6fa4b8-Paper.pdf.

[371] Meng Liu et al. "Temporal Adaptive Module for Video Recognition". In: *ICCV*. 2021.

[372] Ruyang Liu et al. "One for all: Video conversation is feasible without video instruction tuning". In: *arXiv preprint arXiv:2309.15785* (2023).

[373] Shih-Yang Liu et al. *DoRA: Weight-Decomposed Low-Rank Adaptation*. 2024. arXiv: 2402.09353 [cs.CL]. URL: https://arxiv.org/abs/2402.09353.

[374] Shilong Liu et al. "DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[375] Shilong Liu et al. *DQ-DETR: Dual Query Detection Transformer for Phrase Extraction and Grounding*. 2022. arXiv: 2211.15516 [cs.CV]. URL: https://arxiv.org/abs/2211.15516.

[376] Shilong Liu et al. "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection". In: (2023). arXiv: 2303.05499 [cs.CV]. URL: https://arxiv.org/abs/2303.05499.

[377] Yu-Tao Liu et al. “NeUDF: Learning Neural Unsigned Distance Fields with Volume Rendering”. In: *arXiv preprint arXiv:2302.04416* (2023).

[378] Weiyang Liu et al. “Large-Margin Softmax Loss for Convolutional Neural Networks”. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. 2016.

[379] Weiyang Liu et al. “SphereFace: Deep Hypersphere Embedding for Face Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2017, pp. 6738–6746. DOI: 10.1109/CVPR.2017.713. URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Liu_SphereFace_Deep_Hypersphere_CVPR_2017_paper.pdf.

[380] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

[381] Yuan Liu et al. “TS2-Net: Token Shift and Selection Transformer for Video-Language Pretraining”. In: *CVPR*. 2022.

[382] Yuan Liu et al. “Neural Rays for Occlusion-Aware Image-Based Rendering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. arXiv: 2107.13421. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Liu_Neural_Rays_for_Occlusion-Aware_Image-Based_Rendering_CVPR_2022_paper.pdf.

[383] Ze Liu et al. “End-to-End Temporal Action Detection with Transformer”. In: *ECCV*. 2022.

[384] Ze Liu, Jia Ning, and et al. “Video Swin Transformer”. In: *CVPR*. 2022.

[385] Ze Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution”. In: *CVPR*. 2022.

[386] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV]. URL: https://arxiv.org/abs/2103.14030.

[387] Zhaoyang Liu et al. “TAM: Temporal Adaptive Module for Video Recognition”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 13708–13718. URL: https://arxiv.org/abs/2005.06803.

[388] Zhuang Liu et al. *A ConvNet for the 2020s*. 2022. arXiv: 2201.03545 [cs.CV]. URL: https://arxiv.org/abs/2201.03545.

[389] Stephen Lombardi et al. “Neural Volumes: Learning Dynamic Renderable Volumes from Images”. In: *ACM Transactions on Graphics* 38.4 (2019), 65:1–65:14. DOI: 10.1145/3306346.3323020. URL: https://dl.acm.org/doi/10.1145/3306346.3323020.

[390] Jonathan Long, Evan Shelhamer, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. 2015. arXiv: 1411.4038 [cs.CV].

[391] William E. Lorensen and Harvey E. Cline. “Marching cubes: A high resolution 3D surface construction algorithm”. In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’87. New York, NY, USA: Association for Computing Machinery, 1987, pp. 163–169. ISBN: 0897912276. DOI: 10.1145/37401.37422. URL: https://doi.org/10.1145/37401.37422.

[392] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations (ICLR)* (2019). URL: https://arxiv.org/abs/1711.05101.

[393] David G Lowe. “Three-dimensional object recognition from single two-dimensional images”. In: *Proceedings of the IEEE International Joint Conference on Artificial Intelligence (IJCAI)*. 1987, pp. 535–540.

[394] David G. Lowe. “Object Recognition from Local Scale-Invariant Features”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1999, pp. 1150–1157. DOI: 10.1109/ICCV.1999.790410.

[395] Cheng Lu et al. “DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 5775–5787. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/260a14acce2a89dad36adc8eefe7c59e-Paper-Conference.pdf.

[396] Erika Lu et al. “Omnimatte: Associating Objects and Their Effects in Video”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 4507–4516. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Lu_Omnimatte_Associating_Objects_and_Their_Effects_in_Video_CVPR_2021_paper.html.

[397] Erika Lu et al. “OmnimatteRF: Dampened Global Transport for Layered Neural Rendering”. In: *arXiv preprint arXiv:2306.00989* (2023). arXiv: 2306.00989 [cs.CV]. URL: https://arxiv.org/abs/2306.00989.

[398] Pan Lu et al. *Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models*. 2023. arXiv: 2304.09842 [cs.CL]. URL: https://arxiv.org/abs/2304.09842.

[399] Pan Lu et al. *Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering*. 2022. arXiv: 2209.09513 [cs.CL]. URL: https://arxiv.org/abs/2209.09513.

[400] Mario Lucic et al. “Are GANs Created Equal? A Large-Scale Study”. In: *Advances in Neural Information Processing Systems* 31 (2018). URL: https://proceedings.neurips.cc/paper/2018/file/7a982f24ad653db0e0e92fd2bba507a3-Paper.pdf.

[401] Mario Lucic et al. “High-fidelity image generation with fewer labels”. In: *International Conference on Machine Learning (ICML)*. 2019, pp. 4180–4189. URL: http://proceedings.mlr.press/v97/lucic19a.html.

[402] Calvin Luo. *Diffusion Models from Scratch*. https://calvinluo.com/2022/08/26/diffusion-tutorial.html. Tutorial blog post. 2022. URL: https://calvinluo.com/2022/08/26/diffusion-tutorial.html.

[403] Honglu Luo et al. “CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval”. In: *NeurIPS Datasets and Benchmarks*. 2022.

[404] Ruipu Luo et al. *Valley: Video Assistant with Large Language model Enhanced abilitY*. 2025. arXiv: 2306.07207 [cs.CV]. URL: https://arxiv.org/abs/2306.07207.

[405] Wenjie Luo et al. *Understanding the Effective Receptive Field in Deep Convolutional Neural Networks*. 2017. arXiv: 1701.04128 [cs.CV]. URL: https://arxiv.org/abs/1701.04128.

[406] Wenyu Lv et al. *RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer*. 2024. arXiv: 2407.17140 [cs.CV]. URL: https://arxiv.org/abs/2407.17140.

[407] Bin Ma et al. "X-CLIP: End-to-End Multi-Granular Contrastive Learning for Video-Text Retrieval". In: *ACM MM*. 2022.

[408] Ningning Ma et al. "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 122–138. DOI: 10.1007/978-3-030-01264-9_8.

[409] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[410] Muhammad Maaz et al. *Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models*. 2024. arXiv: 2306.05424 [cs.CV]. URL: https://arxiv.org/abs/2306.05424.

[411] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *ICLR*. 2018.

[412] Mahin Khan Mahadi et al. "Gated recurrent unit (GRU)-based deep learning method for spectrum estimation and inverse modeling in plasmonic devices". In: *Applied Physics A* 130.784 (2024). DOI: 10.1007/s00339-024-07956-z. URL: https://link.springer.com/article/10.1007/s00339-024-07956-z.

[413] Turki Al Malki et al. *Mega-NeRF: Scalable Construction of Large-Scale Neural Radiance Fields*. 2022. arXiv: 2112.10752 [cs.CV]. URL: https://arxiv.org/abs/2112.10752.

[414] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. *EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding*. 2023. arXiv: 2308.09126 [cs.CV]. URL: https://arxiv.org/abs/2308.09126.

[415] Kevis-Kokitsi Maninis et al. *TIPS: Text-Image Pretraining with Spatial awareness*. 2025. arXiv: 2410.16512 [cs.CV]. URL: https://arxiv.org/abs/2410.16512.

[416] Xudong Mao et al. "Least Squares Generative Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[417] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, 1982.

[418] Ricardo Martin-Brualla et al. *NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections*. 2021. arXiv: 2008.02268 [cs.CV]. URL: https://arxiv.org/abs/2008.02268.

[419] Robert J. McCann. "A Convexity Principle for Interacting Gases". In: *Advances in Mathematics* 128.1 (1997), pp. 153–179. DOI: 10.1006/aima.1997.1634.

[420] Medium Contributor. *NT–Xent Loss: Normalized Temperature-Scaled Cross-Entropy Loss*. https://medium.com/self-supervised-learning/nt-xent-loss-normalized-temperature-scaled-cross-entropy-loss-ea5a1ede7c40. Accessed: 2025-06-15. 2021.

[421] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. "Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences". In: *AAAI Conference on Artificial Intelligence*. 2016.

[422] Chenlin Meng et al. *SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations*. 2022. arXiv: 2108.01073 [cs.CV]. URL: https://arxiv.org/abs/2108.01073.

[423] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. *PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models*. 2025. arXiv: 2404.02948 [cs.LG]. URL: https://arxiv.org/abs/2404.02948.

[424] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which training methods for GANs do actually converge?" In: *International conference on machine learning*. PMLR. 2018, pp. 3481–3490.

[425] Lars Mescheder et al. "Occupancy networks: Learning 3d reconstruction in function space". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4460–4470.

[426] Moustafa Meshry et al. "Neural Rerendering in the Wild". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019, pp. 6878–6887. DOI: 10.1109/CVPR.2019.00704.

[427] Gal Metzer et al. *Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures*. 2022. arXiv: 2211.07600 [cs.CV]. URL: https://arxiv.org/abs/2211.07600.

[428] Ben Mildenhall et al. *Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines*. 2019. arXiv: 1905.00889 [cs.CV]. URL: https://arxiv.org/abs/1905.00889.

[429] Ben Mildenhall et al. *NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*. 2020. arXiv: 2003.08934 [cs.CV]. URL: https://arxiv.org/abs/2003.08934.

[430] Ben Mildenhall et al. *RawNeRF: Neural Radiance Fields from Noisy Raw Images*. 2022. arXiv: 2111.13679 [cs.CV]. URL: https://arxiv.org/abs/2111.13679.

[431] Matthias Minderer, Alexey Dosovitskiy, Xiaohua Zhai, et al. "Scaling Open-Vocabulary Object Detection". In: *arXiv preprint arXiv:2402.11682* (2024). URL: https://arxiv.org/abs/2402.11682.

[432] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. *Scaling Open-Vocabulary Object Detection*. 2024. arXiv: 2306.09683 [cs.CV]. URL: https://arxiv.org/abs/2306.09683.

[433] Matthias Minderer et al. "Simple Open-Vocabulary Object Detection with Vision Transformers". In: *European Conference on Computer Vision (ECCV)*. 2022. URL: https://arxiv.org/abs/2205.06230.

[434] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[435] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. arXiv: 1411.1784 [cs.LG]. URL: https://arxiv.org/abs/1411.1784.

[436] Ishan Misra and Laurens van der Maaten. *Self-Supervised Learning of Pretext-Invariant Representations*. 2019. arXiv: 1912.01991 [cs.CV]. URL: https://arxiv.org/abs/1912.01991.

[437] Jovana Mitrovic et al. *Representation Learning via Invariant Causal Mechanisms*. 2020. arXiv: 2010.07922 [cs.LG]. URL: https://arxiv.org/abs/2010.07922.

[438] Takeru Miyato et al. "Spectral Normalization for Generative Adversarial Networks". In: *International Conference on Learning Representations (ICLR)*. 2018.

[439] Kaichun Mo et al. "PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding". In: *CVPR*. 2019.

[440] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "Universal Adversarial Perturbations". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1765–1773. DOI: 10.1109/CVPR.2017.190.

[441] Alexander Mordvintsev, Chris Olah, and Mike Tyka. *Inceptionism: Going Deeper into Neural Networks*. Google Research Blog. 2015. URL: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

[442] Tianxiang Mou et al. *T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models*. 2023. arXiv: 2302.08453 [cs.CV]. URL: https://arxiv.org/abs/2302.08453.

[443] Thomas Müller et al. *Instant Neural Graphics Primitives with a Multiresolution Hash Encoding*. 2022. arXiv: 2201.05989 [cs.GR]. URL: https://arxiv.org/abs/2201.05989.

[444] Muhammad Ferjad Naeem et al. *SILC: Improving Vision Language Pretraining with Self-Distillation*. 2023. arXiv: 2310.13355 [cs.CV]. URL: https://arxiv.org/abs/2310.13355.

[445] Preetum Nakkiran et al. "Deep Double Descent: Where Bigger Models and More Data Hurt". In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12 (2021), p. 124003. DOI: 10.1088/1742-5468/ac2db2. URL: https://arxiv.org/abs/1912.02292.

[446] Daniel Neimark et al. "Video Transformer Network". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021. arXiv: 2102.00719. URL: https://openaccess.thecvf.com/content/ICCV2021W/CVEU/papers/Neimark_Video_Transformer_Network_ICCVW_2021_paper.pdf.

[447] Anh Nguyen, Jason Yosinski, and Jeff Clune. *Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks*. 2016. arXiv: 1602.03616 [cs.NE]. URL: https://arxiv.org/abs/1602.03616.

[448] Bao Nguyen, Binh Nguyen, and Viet Anh Nguyen. *Bellman Optimal Stepsize Straightening of Flow-Matching Models*. 2023. arXiv: 2312.16414 [cs.LG]. URL: https://arxiv.org/abs/2312.16414.

[449] Alex Nichol and Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2102.09672 [cs.LG]. URL: https://arxiv.org/abs/2102.09672.

[450] Alex Nichol et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. 2022. arXiv: 2112.10741 [cs.CV]. URL: https://arxiv.org/abs/2112.10741.

[451] Michael Niemeyer et al. “Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[452] Michael Niemeyer et al. “RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[453] Jim Nilsson and Tomas Akenine-Möller. “Understanding SSIM”. In: (2020). arXiv: 2006.13846 [cs.CV]. URL: https://arxiv.org/abs/2006.13846.

[454] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. *Learning Deconvolution Network for Semantic Segmentation*. 2015. arXiv: 1505.04366 [cs.CV].

[455] Michael Oechsle, Songyou Peng, and Andreas Geiger. *UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction*. 2021. arXiv: 2104.10078 [cs.CV]. URL: https://arxiv.org/abs/2104.10078.

[456] Seoung Wug Oh et al. “Video Object Segmentation using Space–Time Memory Networks”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. arXiv: 1904.00607 [cs.CV].

[457] Ozan Oktay et al. *Attention U-Net: Learning Where to Look for the Pancreas*. 2018. arXiv: 1804.03999 [cs.CV]. URL: https://arxiv.org/abs/1804.03999.

[458] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. *Pixel Recurrent Neural Networks*. 2016. arXiv: 1601.06759 [cs.CV]. URL: https://arxiv.org/abs/1601.06759.

[459] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. 2019. arXiv: 1807.03748 [cs.LG]. URL: https://arxiv.org/abs/1807.03748.

[460] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. *Neural Discrete Representation Learning*. 2018. arXiv: 1711.00937 [cs.LG]. URL: https://arxiv.org/abs/1711.00937.

[461] Aaron van den Oord et al. *Conditional Image Generation with PixelCNN Decoders*. 2016. arXiv: 1606.05328 [cs.CV]. URL: https://arxiv.org/abs/1606.05328.

[462] OpenAI. *GPT-4V(ision) System Card*. Tech. rep. OpenAI, 2023. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[463] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2023. arXiv: 2304.07193 [cs.CV]. URL: https://arxiv.org/abs/2304.07193.

[464] et al. Oriol Vinyals. “Show and Tell: A Neural Image Caption Generator”. In: *Computer Vision and Pattern Recognition (CVPR)* (2015).

[465] Junting Pan et al. “Actor-Context-Actor Relation Network for Spatio-Temporal Action Detection”. In: *CVPR*. 2021.

[466] Tianyu Pan et al. “VideoMoCo: Contrastive Video Representation Learning with Temporally Adversarial Examples”. In: *CVPR*. 2021.

[467] Jeong Joon Park et al. *DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation*. 2019. arXiv: 1901.05103 [cs.CV]. URL: https://arxiv.org/abs/1901.05103.

[468] Keunhong Park et al. *HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields*. 2021. arXiv: 2106.13228 [cs.CV]. URL: https://arxiv.org/abs/2106.13228.

[469] Keunhong Park et al. *Nerfies: Deformable Neural Radiance Fields*. 2021. arXiv: 2108.01454 [cs.CV]. URL: https://arxiv.org/abs/2108.01454.

[470] Taesung Park et al. "Semantic Image Synthesis with Spatially-Adaptive Normalization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2337–2346.

[471] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the Difficulty of Training Recurrent Neural Networks". In: *International Conference on Machine Learning (ICML)*. 2013.

[472] Deepak Pathak et al. "Context Encoders: Feature Learning by Inpainting". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2536–2544. DOI: 10.1109/CVPR.2016.278.

[473] Kaushik Patnaik. *ROI Pool and Align: PyTorch Implementation*. 2020. URL: https://kaushikpatnaik.github.io/annotated/papers/2020/07/04/ROI-Pool-and-Align-Pytorch-Implementation.html.

[474] Viorica Pătrăucean et al. "Perception Test: A Diagnostic Benchmark for Multimodal Video Models". In: *arXiv:2305.13786* (2023).

[475] Mandela Patrick et al. "Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers". In: *NeurIPS*. 2021.

[476] Mandela Patrick et al. "MFormer: Masked Transformer for Video Action Recognition". In: *NeurIPS*. 2021.

[477] Judea Pearl. *Causality*. Cambridge university press, 2009.

[478] William Peebles and Saining Xie. *Scalable Diffusion Models with Transformers*. 2023. arXiv: 2212.09748 [cs.CV]. URL: https://arxiv.org/abs/2212.09748.

[479] Ethan Perez et al. *FiLM: Visual Reasoning with a General Conditioning Layer*. 2017. arXiv: 1709.07871 [cs.CV]. URL: https://arxiv.org/abs/1709.07871.

[480] Lawrence Perko. *Differential Equations and Dynamical Systems*. 3rd. Vol. 7. Texts in Applied Mathematics. Springer, 2013. DOI: 10.1007/978-1-4419-7254-8.

[481] Ed Pizzi et al. *A Self-Supervised Descriptor for Image Copy Detection*. 2022. arXiv: 2202.10261 [cs.CV]. URL: https://arxiv.org/abs/2202.10261.

[482] Dustin Podell et al. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. URL: https://arxiv.org/abs/2307.01952.

[483] Tomaso Poggio et al. "Why and when can deep—but not shallow—networks avoid the curse of dimensionality: A review". In: *International Journal of Automation and Computing* 14.5 (2017), pp. 503–519.

[484] B. T. Polyak and A. B. Juditsky. “Acceleration of Stochastic Approximation by Averaging”. In: *SIAM Journal on Control and Optimization* 30.4 (1992), pp. 838–855. DOI: 10.1137/0330046. eprint: https://doi.org/10.1137/0330046. URL: https://doi.org/10.1137/0330046.

[485] Jordi Pont-Tuset et al. “The 2017 DAVIS Challenge on Video Object Segmentation”. In: *arXiv preprint arXiv:1704.00675* (2017).

[486] Aram-Alexandre Pooladian et al. “Multisample Flow Matching: Straightening Flows with Minibatch Couplings”. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. 2023, pp. 28100–28127. URL: https://proceedings.mlr.press/v202/pooladian23a.html.

[487] Ben Poole et al. *DreamFusion: Text-to-3D using 2D Diffusion*. 2022. arXiv: 2209.14988 [cs.CV]. URL: https://arxiv.org/abs/2209.14988.

[488] Albert Pumarola et al. “D-NeRF: Neural Radiance Fields for Dynamic Scenes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Preprint at arXiv:2011.13961. IEEE, 2021, pp. 10313–10322. DOI: 10.1109/CVPR46437.2021.01018. URL: https://ieeexplore.ieee.org/document/9578753.

[489] Charles R. Qi et al. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[490] Charles R. Qi et al. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[491] Guocheng Qian et al. “PointNeXt: Revisiting PointNet++ with Improved Training and Scaling Strategies”. In: *arXiv preprint* (2022). arXiv: 2206.04670 [cs.CV].

[492] Rui Qian, Joseph Tighe, and et al. “Spatiotemporal Contrastive Video Representation Learning”. In: *CVPR*. 2021.

[493] Danfeng Qin et al. *MobileNetV4 – Universal Models for the Mobile Ecosystem*. 2024. arXiv: 2404.10518 [cs.CV]. URL: https://arxiv.org/abs/2404.10518.

[494] Qwen et al. *Qwen2.5 Technical Report*. 2025. arXiv: 2412.15115 [cs.CL]. URL: https://arxiv.org/abs/2412.15115.

[495] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. In: *International Conference on Learning Representations (ICLR)*. 2016. URL: https://arxiv.org/abs/1511.06434.

[496] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[497] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2021).

[498] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[499] Ilija Radosavovic et al. *Designing Network Design Spaces*. 2020. arXiv: 2003.13678 [cs.CV]. URL: https://arxiv.org/abs/2003.13678.

[500] Ilija Radosavovic et al. *Designing Network Design Spaces*. 2020. arXiv: 2003.13678 [cs.CV]. URL: https://arxiv.org/abs/2003.13678.

[501] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2020. arXiv: 1910.10683 [cs.LG].

[502] Nasim Rahaman et al. "On the Spectral Bias of Neural Networks". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2019.

[503] Prajit Ramachandran, Barret Zoph, and Quoc V Le. "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941* (2017).

[504] Prajit Ramachandran, Barret Zoph, and Quoc V Le. "Swish: a self-gated activation function". In: *arXiv preprint arXiv:1710.05941* 7.1 (2017), p. 5.

[505] Prajit Ramachandran et al. *Stand-Alone Self-Attention in Vision Models*. 2019. arXiv: 1906.05909 [cs.CV]. URL: https://arxiv.org/abs/1906.05909.

[506] Sameera Ramasinghe, Lachlan Macdonald, and Simon Lucey. "On the Frequency Bias of Coordinate-MLPs". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[507] Aditya Ramesh et al. "DALL-E: Creating Images from Text Descriptions". In: *OpenAI Blog* 1.1 (2021). URL: https://openai.com/dall-e.

[508] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV]. URL: https://arxiv.org/abs/2204.06125.

[509] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV]. URL: https://arxiv.org/abs/2102.12092.

[510] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision Transformers for Dense Prediction". In: *IEEE International Conference on Computer Vision (ICCV)*. 2021.

[511] René Ranftl et al. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.3 (2022).

[512] Yongming Rao et al. *Dynamic Spatial Sparsification for Efficient Vision Transformers and Convolutional Neural Networks*. 2022. arXiv: 2207.01580 [cs.CV]. URL: https://arxiv.org/abs/2207.01580.

[513] Nikhila Ravi et al. "SAM 2: Segment Anything in Images and Videos". In: (2024). arXiv: 2408.00714 [cs.CV]. URL: https://arxiv.org/abs/2408.00714.

[514] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. *Generating Diverse High-Fidelity Images with VQ-VAE-2*. 2019. arXiv: 1906.00446 [cs.LG]. URL: https://arxiv.org/abs/1906.00446.

[515] Albert Recasens et al. "Broaden Your Views for Self-Supervised Video Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Recasens_Broaden_Your_Views_for_Self-Supervised_Video_Learning_ICCV_2021_paper.pdf.

[516] Joseph Redmon and Ali Farhadi. "YOLO9000: better, faster, stronger". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.

[517] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).

[518] Joseph Redmon et al. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: 1506.02640 [cs.CV]. URL: https://arxiv.org/abs/1506.02640.

[519] Scott E. Reed et al. “Generative Adversarial Text to Image Synthesis”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2016, pp. 4985–4993.

[520] Scott E. Reed et al. “Learning What and Where to Draw: Generative Adversarial What-Where Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, pp. 217–225.

[521] Jeremy Reizenstein et al. “CO3D: Common Objects in 3D for Few-Shot View Synthesis”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[522] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Neural Information Processing Systems (NeurIPS)* (2015).

[523] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV]. URL: https://arxiv.org/abs/1506.01497.

[524] Tianhe Ren et al. “Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks”. In: *arXiv preprint arXiv:2401.14159* (2024). DOI: 10.48550/arXiv.2401.14159. URL: https://arxiv.org/abs/2401.14159.

[525] Tianhe Ren et al. *Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection*. 2024. arXiv: 2405.10300 [cs.CV]. URL: https://arxiv.org/abs/2405.10300.

[526] Jerome Revaud et al. “Learning with Average Precision: Training Image Retrieval with a Listwise Loss”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019. arXiv: 1906.07589 [cs.CV].

[527] Hamid Rezatofighi et al. *Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression*. 2019. arXiv: 1902.09630 [cs.CV]. URL: https://arxiv.org/abs/1902.09630.

[528] Pierre H. Richemond et al. “BYOL works even without batch statistics”. In: *arXiv preprint* arXiv:2010.10241 (2020). Preprint.

[529] Lawrence G. Roberts. “Machine Perception of Three-Dimensional Solids”. PhD Thesis. Massachusetts Institute of Technology, 1963.

[530] Renan A. Rojas-Gomez et al. *SASSL: Enhancing Self-Supervised Learning via Neural Style Transfer*. 2024. arXiv: 2312.01187 [cs.CV]. URL: https://arxiv.org/abs/2312.01187.

[531] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV]. URL: https://arxiv.org/abs/2112.10752.

[532] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[533] Frank Rosenblatt. *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*. Vol. 65. 1958, pp. 386–408.

[534] Antoni Rosinol et al. “SplaTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM”. In: *CVPR*. 2024.

[535] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “"GrabCut": interactive foreground extraction using iterated graph cuts”. In: *ACM SIGGRAPH 2004 Papers*. SIGGRAPH ’04. Los Angeles, California: Association for Computing Machinery, 2004, pp. 309–314. ISBN: 9781450378239. DOI: 10.1145/1186562.1015720. URL: https://doi.org/10.1145/1186562.1015720.

[536] W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. ISBN: 9780070856134. URL: https://books.google.co.il/books?id=kwqzPAAACAAJ.

[537] Nataniel Ruiz et al. “DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation”. In: *arXiv preprint arXiv:2208.12242* (2022).

[538] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning Representations by Back-Propagating Errors”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Ed. by David E. Rumelhart, James L. McClelland, and the PDP Research Group. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362. ISBN: 0-262-18120-7.

[539] David Sage et al. “Logo-GAN-AE: Large-scale conditional generation from small data”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Workshop paper. 2018.

[540] Chitwan Saharia et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: 2205.11487 [cs.CV]. URL: https://arxiv.org/abs/2205.11487.

[541] Mehdi S. M. Sajjadi et al. “Assessing Generative Models via Precision and Recall”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018. URL: https://proceedings.neurips.cc/paper/2018/file/e97bcd67bba095d137723f6c7f3e2430-Paper.pdf.

[542] Tim Salimans and Jonathan Ho. *Progressive Distillation for Fast Sampling of Diffusion Models*. 2022. arXiv: 2202.00512 [cs.LG]. URL: https://arxiv.org/abs/2202.00512.

[543] Tim Salimans et al. “Improved Techniques for Training GANs”. In: *Advances in Neural Information Processing Systems* 29 (2016). URL: https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761aa6e4f85fa48-Paper.pdf.

[544] Tim Salimans et al. “PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications”. In: *ICLR*. 2017.

[545] *SAM 2: Segment Anything in Images and Videos (official repository)*. https://github.com/facebookresearch/sam2. Accessed 2025-10-08.

[546] et al. Sander Dieleman. “Rotation-Invariant Convolutional Neural Networks for Galaxy Morphology Prediction”. In: *Monthly Notices of the Royal Astronomical Society (MNRAS)* (2014).

[547] Mark Sandler et al. “MobileNetV2: Inverted Residuals and Linear Bottlenecks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[548] Aditya Sanghi et al. “CLIP-Mesh: Generating Text-to-Mesh with Text-Supervised Training”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022. URL: https://arxiv.org/abs/2203.13333.

[549] Shibani Santurkar et al. “How Does Batch Normalization Help Optimization?” In: *Advances in Neural Information Processing Systems (NeurIPS)* (2018). URL: https://arxiv.org/abs/1805.11604.

[550] Shibani Santurkar et al. *How Does Batch Normalization Help Optimization?* 2019. arXiv: 1805.11604 [stat.ML]. URL: https://arxiv.org/abs/1805.11604.

[551] Axel Sauer, Katja Schwarz, and Andreas Geiger. *StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets*. 2022. arXiv: 2202.00273 [cs.LG]. URL: https://arxiv.org/abs/2202.00273.

[552] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. “Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2014.

[553] Johannes L. Schonberger and Jan-Michael Frahm. “Structure-From-Motion Revisited”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[554] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015, pp. 815–823. DOI: 10.1109/cvpr.2015.7298682. URL: http://dx.doi.org/10.1109/CVPR.2015.7298682.

[555] Christoph Schuhmann et al. *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*. 2021. arXiv: 2111.02114 [cs.CV]. URL: https://arxiv.org/abs/2111.02114.

[556] Christoph Schuhmann et al. “LAION-5B: An open large-scale dataset for training next generation image-text models”. In: *arXiv preprint arXiv:2210.08402* (2022).

[557] Towards Data Science. *A Basic Introduction to Separable Convolutions*. Accessed: 2025-02-09. 2023. URL: https://medium.com/towards-data-science/a-basic-introduction-to-separable-convolutions-b99ec3102728.

[558] Steven M. Seitz et al. “A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2006.

[559] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[560] Zhenwei Shao et al. “Prompting Large Language Models with Answer Heuristics for Knowledge-based VQA”. In: *CVPR*. 2023.

[561] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. *Self-attention with relative position representations*. NAACL-HLT. 2018.

[562] Noam Shazeer. *Fast Transformer Decoding: One Write-Head is All You Need*. 2019. arXiv: 1911.02150 [cs.NE]. URL: https://arxiv.org/abs/1911.02150.

[563] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905. DOI: 10.1109/34.868688.

[564] Wenzhe Shi et al. *Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network*. 2016. arXiv: 1609.05158 [cs.CV]. URL: https://arxiv.org/abs/1609.05158.

[565] Daeyun Shin, Charless C. Fowlkes, and Derek Hoiem. *Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction*. 2018. arXiv: 1804.06032 [cs.CV]. URL: https://arxiv.org/abs/1804.06032.

[566] Yao Shu et al. *Unifying and Boosting Gradient-Based Training-Free Neural Architecture Search*. 2022. arXiv: 2201.09785 [cs.LG]. URL: https://arxiv.org/abs/2201.09785.

[567] Tim Shuttleworth et al. "An Illusion of Equivalence: Revisiting Low-Rank Adaptation and Full Fine-Tuning". In: *arXiv preprint arXiv:2410.21228* (2024).

[568] K. Simek. *Understanding Camera Calibration and Intrinsics*. Accessed July 2025. 2013. URL: https://ksimek.github.io/2013/08/13/intrinsic/.

[569] Oriane Simeoni et al. *DINOv3*. 2025. arXiv: 2508.10104 [cs.CV]. URL: https://arxiv.org/abs/2508.10104.

[570] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. 2014. arXiv: 1312.6034 [cs.CV]. URL: https://arxiv.org/abs/1312.6034.

[571] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos". In: *Neural Information Processing Systems (NeurIPS)*. 2014.

[572] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICLR)*. 2015. URL: https://arxiv.org/abs/1409.1556.

[573] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. *Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations*. 2019. arXiv: 1906.01618 [cs.CV]. URL: https://arxiv.org/abs/1906.01618.

[574] Edward J. Smith et al. *GEOMetrics: Exploiting Geometric Structure for Graph-Encoded Objects*. 2019. arXiv: 1901.11461 [cs.CV]. URL: https://arxiv.org/abs/1901.11461.

[575] Leslie N. Smith. "Cyclical Learning Rates for Training Neural Networks". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 464–472. DOI: 10.1109/WACV.2017.58. URL: https://arxiv.org/abs/1506.01186.

[576] Leslie N. Smith and Nicholay Topin. "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates". In: *arXiv preprint arXiv:1708.07120* (2018). URL: https://arxiv.org/abs/1708.07120.

[577] Noah Snavely, Steven M. Seitz, and Richard Szeliski. "Modeling the World from Internet Photo Collections". In: *International Journal of Computer Vision* 80.2 (Nov. 2008), pp. 189–210. DOI: 10.1007/s11263-007-0107-3.

[578] Jascha Sohl-Dickstein et al. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)* (2015), pp. 2256–2265. URL: https://arxiv.org/abs/1503.03585.

[579] Pavithra Solai. *Convolutions and Backpropagation*. 2023. URL: https://pavisj.medium.com/convolutions-and-backpropagations-46026a8f5d2c.

[580] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2010.02502.

[581] Yang Song et al. "Consistency Models". In: *NeurIPS*. 2023. arXiv: 2303.01469 [cs.LG].

[582] Yang Song et al. *From Signal to Noise: Understanding Diffusion Models*. https://cvpr2022-tutorial-diffusion-models.github.io/. CVPR 2022 Tutorial. 2022.

[583] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG]. URL: https://arxiv.org/abs/2011.13456.

[584] Jost Tobias Springenberg et al. *Striving for Simplicity: The All Convolutional Net*. 2015. arXiv: 1412.6806 [cs.LG]. URL: https://arxiv.org/abs/1412.6806.

[585] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[586] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Training Very Deep Networks". In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015, pp. 2377–2385.

[587] Stability AI. *Stable Diffusion Image Variations (official release)*. https://github.com/Stability-AI/stable-diffusion-image-variation. 2022.

[588] Stability AI. *Stable Diffusion unCLIP*. https://github.com/Stability-AI/stable-diffusion-unclip. 2022.

[589] Andreas Steiner et al. *How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers*. 2022. arXiv: 2106.10270 [cs.CV]. URL: https://arxiv.org/abs/2106.10270.

[590] Jane Street. "L2 Regularization and Batch Normalization: How They Interact". In: *Jane Street Blog* (2020). URL: https://blog.janestreet.com/l2-regularization-and-batch-norm/.

[591] Cheng Sun, Min Sun, and Hwann-Tzong Chen. *Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction*. 2022. arXiv: 2111.11215 [cs.CV]. URL: https://arxiv.org/abs/2111.11215.

[592] Cheng Sun, Min Sun, and Hwann-Tzong Chen. *Improved Direct Voxel Grid Optimization for Radiance Fields Reconstruction*. 2022. arXiv: 2206.05085 [cs.GR]. URL: https://arxiv.org/abs/2206.05085.

[593] Xingyuan Sun et al. *Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling*. 2018. arXiv: 1804.04610 [cs.CV]. URL: https://arxiv.org/abs/1804.04610.

[594] Zhongyi Sun et al. “NMS Strikes Back: Suppressing Overconfident Incorrect Queries in DETR”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023.

[595] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems (NeurIPS)*. 2014, pp. 3104–3112.

[596] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[597] Yaniv Taigman et al. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220.

[598] Hao Tan et al. “VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning”. In: *NeurIPS*. 2021.

[599] Mingxing Tan and Quoc Le. “EfficientNetV2: Smaller Models and Faster Training”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 10096–10106. URL: https://proceedings.mlr.press/v139/tan21a.html.

[600] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2019, pp. 6105–6114. DOI: 10.48550/arXiv.1905.11946.

[601] Mingxing Tan et al. *MnasNet: Platform-Aware Neural Architecture Search for Mobile*. 2019. arXiv: 1807.11626 [cs.CV]. URL: https://arxiv.org/abs/1807.11626.

[602] Matthew Tancik et al. *Block-NeRF: Scalable Large Scene Neural View Synthesis*. 2022. arXiv: 2202.05263 [cs.CV]. URL: https://arxiv.org/abs/2202.05263.

[603] Matthew Tancik et al. *Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains*. 2020. arXiv: 2006.10739 [cs.CV]. URL: https://arxiv.org/abs/2006.10739.

[604] Matthew Tancik et al. “Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[605] Matthew Tancik et al. *Nerfacto: A Fast Hash-Grid NeRF Baseline*. 2023. arXiv: 2302.01964 [cs.CV]. URL: https://arxiv.org/abs/2302.01964.

[606] Jiaxiang Tang et al. “DreamFields: Physically Plausible Text-to-3D Generation with Radiance Fields”. In: *ACM SIGGRAPH Asia Conference Papers*. 2023.

[607] Jiaxiang Tang et al. “DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation”. In: *arXiv:2309.16651* (2023).

[608] Ming Tao et al. *DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis*. 2022. arXiv: 2008.05865 [cs.CV]. URL: https://arxiv.org/abs/2008.05865.

[609] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. “Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2088–2096.

[610] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. "Single-View to Multi-View: Reconstructing Unseen Views with a Convolutional Network". In: *arXiv preprint arXiv:1511.06702* (2015).

[611] Maxim Tatarchenko et al. *What Do Single-view 3D Reconstruction Networks Learn?* 2019. arXiv: 1905.03678 [cs.CV]. URL: https://arxiv.org/abs/1905.03678.

[612] DeepFloyd Team. *DeepFloyd IF: A Cascaded Diffusion Model for Text-to-Image Synthesis*. 2023. URL: https://github.com/deep-floyd/IF.

[613] TensorFlow Team. *Higher Accuracy on Vision Models with EfficientNet-Lite*. 2020. URL: https://blog.tensorflow.org/2020/03/higher-accuracy-on-vision-models-with-efficientnet-lite.html.

[614] Mikhail Telgarsky. "Benefits of depth in neural networks". In: *Proceedings of the 29th Annual Conference on Learning Theory (COLT)* (2016), pp. 1517–1539.

[615] Keyu Tian et al. *Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction*. 2024. arXiv: 2404.02905 [cs.CV]. URL: https://arxiv.org/abs/2404.02905.

[616] Yuandong Tian, Xinlei Chen, and Surya Ganguli. "Understanding self-supervised learning dynamics without contrastive pairs". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10268–10278.

[617] Zhi Tian et al. *FCOS: Fully Convolutional One-Stage Object Detection*. 2019. arXiv: 1904.01355 [cs.CV]. URL: https://arxiv.org/abs/1904.01355.

[618] Pavel Tokmakov, Jie Li, and Adrien Gaidon. "Breaking the "Object" in Video Object Segmentation". In: *arXiv preprint arXiv:2212.06200* (2022). VOST dataset.

[619] Ilya Tolstikhin et al. *MLP-Mixer: An all-MLP Architecture for Vision*. 2021. arXiv: 2105.01601 [cs.CV]. URL: https://arxiv.org/abs/2105.01601.

[620] Nenad Tomasev et al. *Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet?* 2022. arXiv: 2201.05119 [cs.CV]. URL: https://arxiv.org/abs/2201.05119.

[621] Alexander Tong et al. "TrajectoryNet: Learning Continuous Dynamics for Optimal Transport". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 16298–16309. URL: https://proceedings.neurips.cc/paper/2020/hash/8f8383b0f3c4e3f235df4ef4d9e8f9Abstract.html.

[622] Zhan Tong et al. "VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *European Conference on Computer Vision (ECCV)*. 2022. URL: https://arxiv.org/abs/2203.12602.

[623] Alexander Toshev and Christian Szegedy. "DeepPose: Human Pose Estimation via Deep Neural Networks". In: *Computer Vision and Pattern Recognition (CVPR)* (2014).

[624] H. Touvron et al. "Training Data-efficient Image Transformers & Distillation through Attention". In: *ICML*. 2021.

[625] Hugo Touvron, Matthieu Cord, and Hervé Jégou. *DeiT III: Revenge of the ViT*. 2022. arXiv: 2204.07118 [cs.CV]. URL: https://arxiv.org/abs/2204.07118.

[626] Hugo Touvron et al. *Fixing the train-test resolution discrepancy*. 2022. arXiv: 1906.06423 [cs.CV]. URL: https://arxiv.org/abs/1906.06423.

[627] Hugo Touvron et al. *Going deeper with Image Transformers*. 2021. arXiv: 2103.17239 [cs.CV]. URL: https://arxiv.org/abs/2103.17239.

[628] Florian Tramèr et al. "On Adaptive Attacks to Adversarial Example Defenses". In: *arXiv preprint arXiv:2002.08307* (2020). URL: https://arxiv.org/abs/2002.08307.

[629] Du Tran et al. "Video Classification with Channel-Separated Convolutional Networks". In: *ICCV*. 2019.

[630] Du Tran et al. *A Closer Look at Spatiotemporal Convolutions for Action Recognition*. 2018. arXiv: 1711.11248 [cs.CV]. URL: https://arxiv.org/abs/1711.11248.

[631] Du Tran et al. "Learning Spatiotemporal Features with 3D Convolutional Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 4489–4497. DOI: 10.1109/ICCV.2015.510.

[632] Alex Trevithick and Bo Yang. "GRF: Learning a General Radiance Field for 3D Representation and Rendering". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. arXiv: 2010.04595. URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Trevithick_GRF_Learning_a_General_Radiance_Field_for_3D_Representation_and_ICCV_2021_paper.pdf.

[633] Prune Truong et al. "SPARF: Neural Radiance Fields from Sparse and Noisy Poses". In: *CVPR*. 2023. URL: https://arxiv.org/abs/2211.11738.

[634] Sh-Tsang. "Review: Group Normalization (GN) for Image Classification". In: *Medium* (2018). URL: https://sh-tsang.medium.com/review-group-norm-gn-group-normalization-image-classification-5f7fe0f58eb6.

[635] Sik-Ho Tsang. *Review — BYOL: Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning*. Online; accessed 2025-06-23.

[636] Michael Tschannen et al. *SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features*. 2025. arXiv: 2502.14786 [cs.CV]. URL: https://arxiv.org/abs/2502.14786.

[637] Vishaal Udandarao et al. *Active Data Curation Effectively Distills Large-Scale Multimodal Models*. 2025. arXiv: 2411.18674 [cs.CV]. URL: https://arxiv.org/abs/2411.18674.

[638] J. R. R. Uijlings et al. "Selective Search for Object Recognition". In: *International Journal of Computer Vision* 104.2 (2013), pp. 154–171. URL: https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013.

[639] Ultralytics. *Data Augmentation using Ultralytics YOLO*. 2025. URL: https://docs.ultralytics.com/guides/yolo-data-augmentation/.

[640] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. *Instance Normalization: The Missing Ingredient for Fast Stylization*. 2017. arXiv: 1607.08022 [cs.CV]. URL: https://arxiv.org/abs/1607.08022.

[641] Mojtaba Valipour et al. "DyLoRA: Parameter-Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation". In: *NeurIPS*. 2022.

[642] VantAI. *Flow Matching and Generative Modeling: Deep Dive and Code Walkthrough*. YouTube video. 2022. URL: https://www.youtube.com/watch?v=5ZSwYogAxYg.

[643] VantAI. *Training Flows with the Continuity Equation*. https://www.youtube.com/watch?v=5ZSwYogAxYg. 2023.

[644] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 5998–6008.

[645] Andreas Veit, Michael Wilber, and Serge Belongie. *Residual Networks Behave Like Ensembles of Relatively Shallow Networks*. 2016. arXiv: 1605.06431 [cs.CV]. URL: https://arxiv.org/abs/1605.06431.

[646] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN: 9783540710493. URL: https://books.google.co.il/books?id=NZXiNAEACAAJ.

[647] Pascal Vincent. "A connection between score matching and denoising autoencoders". In: *Neural Computation* 23.7 (2011), pp. 1661–1674. DOI: 10.1162/NECO_a_00142.

[648] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2015, pp. 3156–3164.

[649] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. 2001, pp. I–511. DOI: 10.1109/CVPR.2001.990517.

[650] Elena Voita et al. *Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned*. ACL. 2019.

[651] Catherine Wah et al. *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. Pasadena, CA, USA: California Institute of Technology, 2011.

[652] Bo Wan et al. *LocCa: Visual Pretraining with Location-aware Captioners*. 2024. arXiv: 2403.19596 [cs.CV]. URL: https://arxiv.org/abs/2403.19596.

[653] Li Wan et al. "Regularization of Neural Networks using DropConnect". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 1058–1066. URL: https://proceedings.mlr.press/v28/wan13.html.

[654] Boyang Wang et al. *IBRNet: Learning Multi-View Image-Based Rendering*. 2021. arXiv: 2102.13090 [cs.CV]. URL: https://arxiv.org/abs/2102.13090.

[655] Feng Wang et al. "Additive Margin Softmax for Face Verification". In: *arXiv preprint arXiv:1801.05599* (2018). arXiv: 1801.05599 [cs.CV]. URL: https://arxiv.org/abs/1801.05599.

[656] Hao Wang et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[657] Heng Wang et al. "Video Modeling with Correlation Networks". In: *arXiv preprint arXiv:1906.03349* (2019). URL: https://arxiv.org/abs/1906.03349.

[658] Jianfeng Wang et al. “All-in-One: Exploring Unified Video-Language Pre-Training”. In: *CVPR*. 2022.

[659] Jianfeng Wang et al. “GIT: A Generative Image-to-Text Transformer for Vision and Language”. In: *arXiv preprint arXiv:2205.14100* (2022).

[660] Jiaqi Wang et al. “V3Det: Vast Vocabulary Visual Detection Dataset”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 19787–19797. DOI: 10.1109/ICCV51070.2023.01817.

[661] Jinghao Wang et al. “BEVT: BERT Pretraining of Video Transformers”. In: *CVPR*. 2022.

[662] Junke Wang et al. *OmniVL:One Foundation Model for Image-Language and Video-Language Tasks*. 2022. arXiv: 2209.07526 [cs.CV]. URL: https://arxiv.org/abs/2209.07526.

[663] Limin Wang et al. “TDN: Temporal Difference Networks for Efficient Video Understanding”. In: *CVPR*. 2021.

[664] Limin Wang et al. “VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking”. In: *arXiv preprint arXiv:2303.16727* (2023). URL: https://arxiv.org/abs/2303.16727.

[665] Nanyang Wang et al. *Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images*. 2018. arXiv: 1804.01654 [cs.CV]. URL: https://arxiv.org/abs/1804.01654.

[666] Peng Wang et al. “BAD-NeRF: Bundle Adjusted Deblur Neural Radiance Fields”. In: *CVPR*. 2023. URL: https://arxiv.org/abs/2211.12853.

[667] Peng Wang et al. “Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction”. In: *arXiv preprint arXiv:2106.10689* (2021).

[668] Peng Wang et al. *Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution*. 2024. arXiv: 2409.12191 [cs.CV]. URL: https://arxiv.org/abs/2409.12191.

[669] Rui Wang et al. “Masked Video Distillation: Rethinking Masked Feature Modeling for Self-Supervised Video Representation Learning”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. URL: https://arxiv.org/abs/2212.04500.

[670] Weiyao Wang et al. “Unidentified Video Objects: A Benchmark for Dense, Open-World Segmentation”. In: *ICCV*. 2021.

[671] Wenhao Wang et al. “InternVideo: General Video Foundation Models via Data, Architecture and Pretraining”. In: *arXiv preprint arXiv:2212.03191* (2022). URL: https://arxiv.org/abs/2212.03191.

[672] Wenhui Wang et al. “Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks”. In: *CVPR*. 2023. URL: https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_Image_as_a_Foreign_Language_BEiT_Pretraining_for_Vision_and_CVPR_2023_paper.pdf.

[673] Xiao Wang and Guo-Jun Qi. “Contrastive Learning with Stronger Augmentations”. In: *ICLR*. 2021. arXiv: 2104.07713 [cs.CV].

[674] Xiaolong Wang and Abhinav Gupta. "Videos as Space-Time Region Graphs". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 399–417. URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Xiaolong_Wang_Videos_as_Space-Time_ECCV_2018_paper.html.

[675] Xiaolong Wang et al. "Non-local Neural Networks". In: *CVPR*. 2018.

[676] Xiaolong Wang et al. *Non-local Neural Networks*. 2018. arXiv: 1711.07971 [cs.CV]. URL: https://arxiv.org/abs/1711.07971.

[677] Yi Wang et al. *InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation*. 2024. arXiv: 2307.06942 [cs.CV]. URL: https://arxiv.org/abs/2307.06942.

[678] Yi Wang et al. "InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding". In: *arXiv preprint arXiv:2403.15377* (2024). URL: https://arxiv.org/abs/2403.15377.

[679] Yi Wang et al. "InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling". In: *arXiv preprint arXiv:2501.12386* (2025). v3, Jul 13, 2025. DOI: 10.48550/arXiv.2501.12386. URL: https://arxiv.org/abs/2501.12386.

[680] Yifan Wang et al. "GS-IR: 3D Gaussian Splatting for Inverse Rendering". In: *CVPR*. 2024.

[681] Yiming Wang et al. "NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction". In: *ICCV*. 2023.

[682] Yiqun Wang et al. "HF-NeuS: Improved Surface Reconstruction Using High-Frequency Details Robust to Noise". In: *NeurIPS*. 2022.

[683] Yiqun Wang, Ivan Skorokhodov, and Peter Wonka. "PET-NeuS: Positional Encoding Tri-Planes for Neural Surfaces". In: *arXiv preprint arXiv:2302.01871*. 2023.

[684] Yue Wang et al. "Dynamic Graph CNN for Learning on Point Clouds". In: *ACM Transactions on Graphics (TOG)*. 2019.

[685] Zhizhong Wang, Lei Zhao, and Wei Xing. *StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models*. 2023. arXiv: 2308.07863 [cs.CV]. URL: https://arxiv.org/abs/2308.07863.

[686] Zirui Wang et al. "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision". In: *arXiv preprint arXiv:2108.10904* (2021). URL: https://arxiv.org/abs/2108.10904.

[687] Chen Wei et al. "Masked Feature Prediction for Self-Supervised Visual Pre-Training". In: *CVPR*. 2022.

[688] Li-Yi Wei and Marc Levoy. "Fast Texture Synthesis using Tree-structured Vector Quantization". In: *Proceedings of SIGGRAPH*. 2000, pp. 479–488.

[689] Chao Wen et al. *Pixel2Mesh++: Multi-View 3D Mesh Generation via Deformation*. 2019. arXiv: 1908.01491 [cs.CV]. URL: https://arxiv.org/abs/1908.01491.

[690] Yuetian Weng et al. "LongVLM: Efficient Long Video Understanding via Large Language Models". In: *European Conference on Computer Vision (ECCV)*. 2024.

[691] Wikipedia contributors. *Aliasing — Wikipedia, The Free Encyclopedia*. [Online; accessed 27-February-2025]. 2004. URL: https://www.wikiwand.com/en/articles/Aliasing.

[692] Wikipedia contributors. *Sine and cosine — Wikipedia, The Free Encyclopedia*. Accessed: 2025-03-14. 2024. URL: https://www.wikiwand.com/en/articles/Sine_and_cosine.

[693] Ronald J Williams. "Simple statistical gradient-following algorithms for connectionist reinforcement learning". In: *Machine learning* 8 (1992), pp. 229–256.

[694] Ronald J. Williams and Jing Peng. "An Efficient Gradient-Based Algorithm for On-Line Training of Recurrent Network Trajectories". In: *Neural Computation*. Vol. 2. 4. 1990, pp. 490–501.

[695] Samuel Williams, Andrew Waterman, and David Patterson. "Roofline: an insightful visual performance model for multicore architectures". In: *Commun. ACM* 52.4 (Apr. 2009), pp. 65–76. ISSN: 0001-0782. DOI: 10.1145/1498765.1498785. URL: https://doi.org/10.1145/1498765.1498785.

[696] Sarah Wolf. *ProGAN – How NVIDIA Generated Images of Unprecedented Quality*. Medium. https://medium.com/data-science/progan-how-nvidia-generated-images-of-unprecedented-quality-51c98ec2cbd2. 2019.

[697] Sanghyun Woo et al. "CBAM: Convolutional Block Attention Module". In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.

[698] Sanghyun Woo et al. *ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*. 2023. arXiv: 2301.00808 [cs.CV]. URL: https://arxiv.org/abs/2301.00808.

[699] Mitchell Wortsman et al. *Robust fine-tuning of zero-shot models*. 2022. arXiv: 2109.01903 [cs.CV]. URL: https://arxiv.org/abs/2109.01903.

[700] Chao-Yi Wu et al. "MeMViT: Memory-Augmented Multiscale Vision Transformers for Efficient Long-Term Video Recognition". In: *arXiv preprint arXiv:2201.08383* (2022). URL: https://arxiv.org/abs/2201.08383.

[701] Chao-Yuan Wu et al. "Multi-Scale Feature Aggregation for Spatio-Temporal Action Detection". In: *ECCV*. 2020.

[702] Xiaoyang Wu et al. "Point Transformer V2: Grouped Vector Attention and Partition-based Pooling". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[703] Zhen Wu et al. "4D Gaussian Splatting for Real-Time Dynamic Scene Rendering". In: *arXiv:2310.08528* (2023).

[704] Zheng Wu et al. "AIA++: Advanced Interaction Aggregation for Spatio-Temporal Action Detection". In: *CVPR Workshops*. 2022.

[705] Zhirong Wu et al. *3D ShapeNets: A Deep Representation for Volumetric Shapes*. 2015. arXiv: 1406.5670 [cs.CV]. URL: https://arxiv.org/abs/1406.5670.

[706] et al. Xiaoxiao Guo. "Deep Reinforcement Learning for Playing Atari Games". In: *Neural Information Processing Systems (NeurIPS)* (2014).

[707] Junyuan Xie, Ross Girshick, and Ali Farhadi. "Unsupervised Deep Embedding for Clustering Analysis". In: *arXiv* abs/1511.06335 (2019). Preprint, archived on arXiv. URL: https://arxiv.org/abs/1511.06335.

[708] Saining Xie et al. *Aggregated Residual Transformations for Deep Neural Networks*. 2017. arXiv: 1611.05431 [cs.CV]. URL: https://arxiv.org/abs/1611.05431.

[709] Huijuan Xu and Kate Saenko. "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering". In: *European Conference on Computer Vision (ECCV)*. 2016.

[710] Kelvin Xu et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *International Conference on Machine Learning (ICML)*. 2015, pp. 2048–2057.

[711] Mengmeng Xu et al. "Boundary-Sensitive Pre-Training for Temporal Localization in Videos". In: *ICCV*. 2021.

[712] Mingze Xu et al. "G-TAD: Sub-Graph Localization for Temporal Action Detection". In: *CVPR*. (RTD-Net results often reported alongside). 2020.

[713] Ning Xu et al. "YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark". In: *ECCV*. 2018, pp. 603–619. DOI: 10.1007/978-3-030-01228-1_36.

[714] Qiangeng Xu et al. *Point-NeRF: Point-based Neural Radiance Fields*. 2022. arXiv: 2201.08845 [cs.CV]. URL: https://arxiv.org/abs/2201.08845.

[715] Tao Xu et al. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 1316–1324.

[716] Xingqian Xu et al. *Prompt-Free Diffusion: Taking "Text" out of Text-to-Image Diffusion Models*. 2023. arXiv: 2305.16223 [cs.CV]. URL: https://arxiv.org/abs/2305.16223.

[717] Xingqian Xu et al. *Versatile Diffusion: Text, Images and Variations All in One Diffusion Model*. 2024. arXiv: 2211.08332 [cs.CV]. URL: https://arxiv.org/abs/2211.08332.

[718] Youssef Xu, Alain Durmus, and Youssef Mroueh. "Sliced Wasserstein Generative Models". In: *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 5453–5462. URL: https://proceedings.mlr.press/v80/xu18e.html.

[719] Zhiliang Xu et al. "LanguageBind: Extending Video-Language Pretraining to Multiple Modalities". In: *arXiv preprint arXiv:2307.12367* (2023). arXiv: 2307.12367 [cs.CV]. URL: https://arxiv.org/abs/2307.12367.

[720] Seung-Hwan Yan, Saining Xie, Kaiming He, et al. "Multiview Transformers for Video Recognition". In: *CVPR*. 2022.

[721] Shen Yan et al. *VideoCoCa: Video-Text Modeling with Zero-Shot Transfer from Contrastive Captioners*. 2023. arXiv: 2212.04979 [cs.CV]. URL: https://arxiv.org/abs/2212.04979.

[722] Antoine Yang et al. *Zero-Shot Video Question Answering via Frozen Bidirectional Language Models*. 2022. arXiv: 2206.08155 [cs.CV]. URL: https://arxiv.org/abs/2206.08155.

[723] Cheng-Yen Yang et al. "SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory". In: *arXiv preprint arXiv:2411.11922* (2024). URL: https://arxiv.org/abs/2411.11922.

[724] Ling Yang et al. *Consistency Flow Matching: Defining Straight Flows with Velocity Consistency*. 2024. arXiv: 2407.02398 [cs.LG]. URL: https://arxiv.org/abs/2407.02398.

[725] Linjie Yang, Yuchen Fan, Ning Xu, et al. "Video Instance Segmentation". In: *IEEE International Conference on Computer Vision (ICCV)*. 2019.

[726] Lewei Yao et al. *DetCLIP: Dictionary-Enriched Visual-Concept Paralleled Pre-training for Open-world Detection*. 2022. arXiv: 2209.09407 [cs.CV]. URL: https://arxiv.org/abs/2209.09407.

[727] Lewei Yao et al. *DetCLIPv2: Scalable Open-Vocabulary Object Detection Pre-Training via Word-Region Alignment*. 2023. arXiv: 2304.04514 [cs.CV]. URL: https://arxiv.org/abs/2304.04514.

[728] Lewei Yao et al. "FILIP: Fine-grained Interactive Language-Image Pre-Training". In: *ICLR*. 2022. URL: https://openreview.net/forum?id=cpDhcsEDC2.

[729] Zhiyuan Yao, Dengxin Zhao, and Chang Xu. "Improving parallel decoding for text generation with speculation". In: *arXiv preprint* (2022). eprint: 2211.04206.

[730] Zhuyu Yao et al. *Efficient DETR: Improving End-to-End Object Detector with Dense Prior*. 2021. arXiv: 2104.01318 [cs.CV]. URL: https://arxiv.org/abs/2104.01318.

[731] Lior Yariv et al. *Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance*. 2020. arXiv: 2003.09852 [cs.CV]. URL: https://arxiv.org/abs/2003.09852.

[732] Lior Yariv et al. *Volume Rendering of Neural Implicit Surfaces*. 2021. arXiv: 2106.12052 [cs.CV]. URL: https://arxiv.org/abs/2106.12052.

[733] Hu Ye et al. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. 2023. arXiv: 2308.06721 [cs.CV]. URL: https://arxiv.org/abs/2308.06721.

[734] Qinghao Ye et al. *mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality*. 2024. arXiv: 2304.14178 [cs.CL]. URL: https://arxiv.org/abs/2304.14178.

[735] Okan Yenigün. *Overfitting vs. Underfitting*. AI Mind. URL: https://miro.medium.com/v2/resize:fit:1400/1*T5FDrB9yAWNkOOODpuOEBQ.png.

[736] Xin Yi, Ekta Walia, and Paul Babyn. "Generative Adversarial Network in Medical Imaging: A Review". In: *Medical Image Analysis* 58 (2019), p. 101552.

[737] Jason Yosinski et al. *Understanding Neural Networks Through Deep Visualization*. 2015. arXiv: 1506.06579 [cs.CV]. URL: https://arxiv.org/abs/1506.06579.

[738] Yang You, Igor Gitman, and Boris Ginsburg. "Large batch training of convolutional networks". In: *arXiv preprint arXiv:1708.03888* (2017). URL: https://arxiv.org/abs/1708.03888.

[739] Alex Yu et al. *pixelNeRF: Neural Radiance Fields from One or Few Images*. 2020. arXiv: 2012.02190 [cs.CV]. URL: https://arxiv.org/abs/2012.02190.

[740] Alex Yu et al. “PlenOctrees for Real-Time Rendering of Neural Radiance Fields”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. arXiv: 2103.14024. URL: https://openaccess.thecvf.com/content/ICCV2021/papers/Yu_PlenOctrees_for_Real-Time_Rendering_of_Neural_Radiance_Fields_ICCV_2021_paper.pdf.

[741] Jiahui Yu et al. “CoCa: Contrastive Captioners are Image-Text Foundation Models”. In: *CVPR*. 2022.

[742] Jiahui Yu et al. *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation*. 2022. arXiv: 2206.10789 [cs.CV]. URL: https://arxiv.org/abs/2206.10789.

[743] Jiahui Yu et al. *Vector-quantized Image Modeling with Improved VQGAN*. 2022. arXiv: 2110.04627 [cs.CV]. URL: https://arxiv.org/abs/2110.04627.

[744] Meng Yu et al. “GSDF: 3DGS Meets SDF for Improved Neural Rendering”. In: *NeurIPS*. 2024.

[745] Lu Yuan and et al. “Florence: A New Foundation Model for Computer Vision”. In: *arXiv:2111.11432* (2021). URL: https://arxiv.org/abs/2111.11432.

[746] Zijian Yue et al. “V4D: 4D Convolutional Neural Networks for Video”. In: *ICLR*. 2020.

[747] Sangdoo Yun et al. “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features”. In: *ICCV*. 2019.

[748] Sergey Zagoruyko and Nikos Komodakis. “DiracNets: Training Very Deep Neural Networks Without Skip-Connections”. In: *International Conference on Learning Representations*. 2017.

[749] Kevin Zakka. *Understanding Batch Normalization Backpropagation*. 2016. URL: https://kevinzakka.github.io/2016/09/14/batch_normalization/.

[750] Yuchen Zang et al. “Open-Vocabulary DETR with Conditional Matching”. In: *European Conference on Computer Vision*. Springer, 2022, pp. 106–122.

[751] Jure Zbontar, Li Jing, Ishan Misra, et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. In: *International Conference on Machine Learning (ICML)* (2021).

[752] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53. arXiv: 1311.2901 [cs.CV].

[753] Rowan Zellers et al. “MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound”. In: *CVPR*. 2022.

[754] Rowan Zellers et al. “MERLOT: Multimodal Neural Script Knowledge Models”. In: *NeurIPS*. 2021.

[755] Runhao Zeng et al. “DCAN: Dual Context Aggregation Network for Temporal Action Detection”. In: *CVPR*. 2021.

[756] Xiang Zhai, Aravind Srinivas, Stella X. Yu, et al. *Large-Scale Evaluation of Self-Supervised Learning for Visual Tasks*. Tech. rep. UC Berkeley, 2019.

[757] Xiaohua Zhai et al. “LiT: Zero-Shot Transfer with Locked-Image Text Tuning”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[758] Xiaohua Zhai et al. *Sigmoid Loss for Language Image Pre-Training*. 2023. arXiv: 2303.15343 [cs.CV]. URL: https://arxiv.org/abs/2303.15343.

[759] Boqiang Zhang et al. *VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding*. 2025. arXiv: 2501.13106 [cs.CV]. URL: https://arxiv.org/abs/2501.13106.

[760] Chen-Lin Zhang et al. "ActionFormer: Localizing Moments of Actions with Transformers". In: *ECCV*. 2022.

[761] Chen-Lin Zhang et al. "TALLFormer: Temporal Action Localization with Long-range Transformer". In: *ECCV*. 2022.

[762] et al. Zhang. "AdaLoRA: Adaptive Low-Rank Adaptation via Gradient-Based Rank Selection". In: *arXiv preprint arXiv:2210.07558* (2022).

[763] Han Zhang et al. "Self-Attention Generative Adversarial Networks". In: *International Conference on Machine Learning (ICML)*. 2019.

[764] Han Zhang et al. "Self-Attention Generative Adversarial Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7354–7363. URL: https://proceedings.mlr.press/v97/zhang19d.html.

[765] Han Zhang et al. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 5907–5915.

[766] Han Zhang et al. "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Vol. 41. 8. 2018, pp. 1947–1962.

[767] Hang Zhang, Xin Li, and Lidong Bing. *Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding*. 2023. arXiv: 2306.02858 [cs.CL]. URL: https://arxiv.org/abs/2306.02858.

[768] Hang Zhang et al. *ResNeSt: Split-Attention Networks*. 2020. arXiv: 2004.08955 [cs.CV]. URL: https://arxiv.org/abs/2004.08955.

[769] Haotian Zhang et al. *GLIPv2: Unifying Localization and Vision-Language Understanding*. 2022. arXiv: 2206.05836 [cs.CV]. URL: https://arxiv.org/abs/2206.05836.

[770] Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. *Fixup Initialization: Residual Learning Without Normalization*. 2019. arXiv: 1901.09321 [cs.LG]. URL: https://arxiv.org/abs/1901.09321.

[771] Hongyi Zhang et al. *mixup: Beyond Empirical Risk Minimization*. 2018. arXiv: 1710.09412 [cs.LG]. URL: https://arxiv.org/abs/1710.09412.

[772] Kai Zhang et al. *NeRF++: Analyzing and Improving Neural Radiance Fields*. 2020. arXiv: 2010.07492 [cs.CV]. URL: https://arxiv.org/abs/2010.07492.

[773] Lvmin Zhang and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". In: *arXiv preprint arXiv:2302.05543* (2023).

[774] Pengchuan Zhang, Xiujun Li, and et al. "VinVL: Making Visual Representations Matter in Vision-Language Models". In: *CVPR*. 2021. URL: https://arxiv.org/abs/2101.00529.

[775] Renrui Zhang et al. “LLaMA-Adapter: Efficient Fine-Tuning of Language Models with Zero-Initiated Attention”. In: *arXiv preprint arXiv:2303.16199* (2023).

[776] Richard Zhang. *Making Convolutional Networks Shift-Invariant Again*. 2019. arXiv: 1904.11486 [cs.CV]. URL: https://arxiv.org/abs/1904.11486.

[777] Richard Zhang, Phillip Isola, and Alexei A. Efros. “Colorful Image Colorization”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 649–666. DOI: 10.1007/978-3-319-46487-9_40.

[778] Richard Zhang et al. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. 2018. arXiv: 1801.03924 [cs.CV]. URL: https://arxiv.org/abs/1801.03924.

[779] Shuhan Zhang et al. “Revisiting Photometric Consistency in Multi-View Stereo”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[780] Xiangyu Zhang et al. “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6848–6856. DOI: 10.1109/CVPR.2018.00716.

[781] Yiming Zhang et al. “SuGaR: Surface-Aligned Gaussian Reconstruction for High-Fidelity Meshes”. In: *CVPR*. 2024.

[782] Youcai Zhang et al. *Recognize Anything: A Strong Image Tagging Model*. 2023. arXiv: 2306.03514 [cs.CV]. URL: https://arxiv.org/abs/2306.03514.

[783] Yuxin Zhang et al. *Inversion-Based Style Transfer with Diffusion Models*. 2023. arXiv: 2211.13203 [cs.CV]. URL: https://arxiv.org/abs/2211.13203.

[784] Zhengxu Zhang, Qingjie Liu, and Yunhong Wang. *Road Extraction by Deep Residual U-Net*. 2018. arXiv: 1711.10684 [cs.CV].

[785] Zhengyou Zhang and Olivier Faugeras. “3-D Shape and Motion Recovery Under Varying Illumination”. In: *International Journal of Computer Vision (IJCV)* (2001).

[786] Zhuosheng Zhang et al. *Multimodal Chain-of-Thought Reasoning in Language Models*. 2024. arXiv: 2302.00923 [cs.CL]. URL: https://arxiv.org/abs/2302.00923.

[787] Chen Zhao et al. “TubeR: Tubelet Transformer for Action Detection”. In: *CVPR*. 2022.

[788] Hang Zhao, Zhicheng Yan, Lorenzo Torresani, et al. “HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization”. In: *ICCV*. 2019.

[789] Hengshuang Zhao et al. “Point Transformer”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 16259–16268.

[790] Hengshuang Zhao et al. *Pyramid Scene Parsing Network*. 2017. arXiv: 1612.01105 [cs.CV]. URL: https://arxiv.org/abs/1612.01105.

[791] Long Zhao et al. *VideoPrism: A Foundational Visual Encoder for Video Understanding*. 2025. arXiv: 2402.13217 [cs.CV]. URL: https://arxiv.org/abs/2402.13217.

[792] Shihao Zhao et al. *Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models*. 2023. arXiv: 2305.16322 [cs.CV]. URL: https://arxiv.org/abs/2305.16322.

[793] Tiancheng Zhao et al. *Real-time Transformer-based Open-Vocabulary Detection with Efficient Fusion Head*. 2024. arXiv: 2403.06892 [cs.CV]. URL: https://arxiv.org/abs/2403.06892.

[794] Yue Zhao et al. *Learning Video Representations from Large Language Models*. 2022. arXiv: 2212.04501 [cs.CV]. URL: https://arxiv.org/abs/2212.04501.

[795] Yiwu Zhong et al. *RegionCLIP: Region-based Language-Image Pretraining*. 2021. arXiv: 2112.09106 [cs.CV]. URL: https://arxiv.org/abs/2112.09106.

[796] Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929.

[797] Bolei Zhou et al. "Places: A 10 Million Image Database for Scene Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 40.6 (2017), pp. 1452–1464. DOI: 10.1109/TPAMI.2017.2723009.

[798] Chunting Zhou et al. *Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model*. 2024. arXiv: 2408.11039 [cs.AI]. URL: https://arxiv.org/abs/2408.11039.

[799] Jinghao Zhou et al. "iBOT: Image BERT Pre-Training with Online Tokenizer". In: *Proc. ICLR*. 2022. URL: https://arxiv.org/abs/2111.07832.

[800] Kaiyang Zhou et al. "Conditional prompt learning for vision-language models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16816–16825.

[801] Luowei Zhou et al. *Unified Vision-Language Pre-Training for Image Captioning and VQA*. 2019. arXiv: 1909.11059 [cs.CV]. URL: https://arxiv.org/abs/1909.11059.

[802] Zongwei Zhou et al. *Unet++: A nested u-net architecture for medical image segmentation*. 2018. arXiv: 1807.10165 [cs.CV].

[803] Deyao Zhu et al. *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models*. 2023. arXiv: 2304.10592 [cs.CV]. URL: https://arxiv.org/abs/2304.10592.

[804] Feng Zhu et al. "RE-DETR: Accelerating Deformable DETR with Reference Points". In: *International Conference on Learning Representations (ICLR)*. 2023.

[805] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2223–2232.

[806] Minfeng Zhu et al. "DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5802–5810.

[807] Wanhua Zhu et al. "Enriching Local and Global Contexts for Temporal Action Detection". In: *ICCV*. 2021.

[808] Xizhou Zhu et al. *Deformable DETR: Deformable Transformers for End-to-End Object Detection*. 2021. arXiv: 2010.04159 [cs.CV]. URL: https://arxiv.org/abs/2010.04159.

[809] Yihang Zhu et al. *Transfusion: Cross-modal Diffusion Models for Reference-based Image Editing and Generation*. 2023. arXiv: 2312.02445 [cs.CV]. URL: https://arxiv.org/abs/2312.02445.

[810] Yuchen Zhu et al. “Chameleon: Mixed-Modal Early-Fusion Foundation Models”. In: *arXiv preprint arXiv:2405.09818* (2024). URL: https://arxiv.org/abs/2405.09818.

[811] C Lawrence Zitnick and Piotr Dollár. “Edge boxes: Locating object proposals from edges”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 391–405.

[812] Barret Zoph and Quoc V. Le. “Neural Architecture Search with Reinforcement Learning”. In: *ICLR*. 2017.

[813] Barret Zoph et al. “Learning Transferable Architectures for Scalable Image Recognition”. In: *CVPR*. 2018.

[814] Xueyan Zou et al. “Generalized Decoding for Pixel, Image, and Language”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023. arXiv: 2212.11270 [cs.CV]. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Zou_Generalized_Decoding_for_Pixel_Image_and_Language_CVPR_2023_paper.html.