DAFTAR PUSTAKA

[1] Z. Xu, "Analysis and experimental study of zero-shot capabilities in vision-language models: an in-depth exploration of contrastive and masked methods," *Theoretical and Natural Science*, vol. 74, pp. 104–111, 2024.

[2] C. Li, H. Liu, L. H. Li, P. Zhang, J. Aneja, J. Yang, P. Jin, Y. J. Lee, H. Hu, Z. Liu, and J. Gao, "Elevater: a benchmark and toolkit for evaluating language-augmented visual models," 2022.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of ICML*, 2021, cLIP.

[4] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon *et al.*, "Openclip," GitHub repository, 2021. [Online]. Available: https://github.com/mlfoundations/open_clip

[5] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *NeurIPS Datasets and Benchmarks*, 2022.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, D. Rolland *et al.*, "Segment anything," in *ICCV*, 2023.

[7] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, "CALIP: Zero-shot enhancement of CLIP with parameter-free attention," *Proc. Conf. AAAI Artif. Intell.*, vol. 37, no. 1, pp. 746–754, Jun. 2023.

[8] S. Liu, W. Pu, W. Xu, Z. Huang, Q. Li, H. Wang, C. Lin, and C. Shen, "A comprehensive survey of multimodal large language models: concept, application and safety," 2024.

[9] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality helps unimodality: cross-modal few-shot learning with multimodal models," 2023.

[10] X. Zhou, H. Yu, S. Yang, J. Huo, and P. Tian, "Learning from orthogonal space with multimodal large models for generalized few-shot segmentation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2025.

[11] S. Yin and L. Jiang, "Distilling knowledge from multiple foundation models for zero-shot image classification," *PLoS One*, vol. 19, no. 9, p. e0310730, Sep. 2024.

[12] H. Sun, Z. Zhen, Y. Liu, X. Zhang, X. Han, and P. Zhang, "Embedded zero-shot image classification based on bidirectional feature mapping," *Appl. Sci. (Basel)*, vol. 14, no. 12, p. 5230, Jun. 2024.

[13] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2501.02189

[14] D. Jiang, Y. Liu, S. Liu, J. Zhao, H. Zhang, Z. Gao, X. Zhang, J. Li, and H. Xiong, "From clip to dino: Visual encoders shout in multi-modal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2310.08825

[15] T. Kerssies, D. de Geus, and G. Dubbelman, "How to benchmark vision foundation models for semantic segmentation?" 2024. [Online]. Available: https://arxiv.org/abs/2404.12172

[16] Z. Mai, A. Chowdhury, Z. Wang, S. Jeon, L. Wang, J. Hou, and W.-L. Chao, "Ava-bench: Atomic visual ability benchmark for vision foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2506.09082

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, p. 2337–2348, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1007/s11263-022-01653-1

[18] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. de Charette, "Clip's visual embedding projector is a few-shot cornucopia," 2025. [Online]. Available: https://arxiv.org/abs/2410.05270

[19] M. J. Mirza, L. Karlinsky, W. Lin, S. Doveh, J. Micorek, M. Kozinski, H. Kuehne, and H. Possegger, "Meta-prompting for automating zero-shot visual recognition with llms," 2024, mPVR.

[20] A. Paderno, A. Rau, N. Bedi, P. Bossi, G. Mercante, C. Piazza, and F. C. Holsinger, "Computer vision foundation models in endoscopy: Proof of concept in oropharyngeal cancer," *The Laryngoscope*, 2024.

[21] L. Yuan, D. Chen, Y. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, "Florence: A new foundation model for computer vision," 2021.

[22] Z. Li, S. Wu, Y. Zhang, and W. Xu, "Anomaly detection model based on few-shot learning and memory modules," *Journal of Electronic Imaging*, 2022.

[23] Z. Zhang, Q. Wu, S. Ding, X. Wang, and J. Ye, "Echo-vision-fm: A pre-training and fine-tuning framework for echocardiogram video vision foundation model," 2024.

[24] M. A. Chia, F. Antaki, Y. Zhou, A. Turner, A. Lee, and P. A. Keane, "Foundation models in ophthalmology," *British Journal of Ophthalmology*, 2024.

[25] S. Yan, Q. Zeng, Y. Qi, L. Lu, W. Dong, and L. Yu, "Research on zero-shot learning based on generative model," 2024.

[26] G. Ramesh, "A review on nlp zero-shot and few-shot learning: methods and applications," *Discov Appl Sci*, vol. 7, 2025.

[27] M. Dabbah and R. El-Yaniv, "Using fictitious class representations to boost discriminative zero-shot learners," 2021.

[28] B. Baiju, P. Suresh, G. Subathra, P. Keerthika, K. Sadasivuni, and K. Logeswaran, "Unlocking the future of healthcare," pp. 258–280, 2024.

[29] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu, "Zero-shot out-of-distribution detection based on the pre-trained model clip," *Proceedings of the Aaai Conference on Artificial Intelligence*, vol. 36, pp. 6568–6576, 2022.

[30] S. Yan, L. Hong, H. Xu, J. Han, T. Tuytelaars, Z. Li, and X. He, "Generative negative text replay for continual vision-language pretraining," 2022.

[31] H. Sun, Z. Zhen, Y. Liu, X. Zhang, X. Han, and P. Zhang, "Embedded zero-shot image classification based on bidirectional feature mapping," *Applied Sciences*, vol. 14, p. 5230, 2024.

[32] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, and L. Beyer, "Lit: zero-shot transfer with locked-image text tuning," 2021.

[33] J. Li, S. Savarese, and S. Hoi, "Masked unsupervised self-training for zero-shot image classification," 2022.

[34] H. Steck, C. Ekanadham, and N. Kallus, "Is cosine-similarity of embeddings really about similarity?" in *Companion Proceedings of the ACM Web Conference 2024*, ser. WWW '24. ACM, May 2024, p. 887–890. [Online]. Available: http://dx.doi.org/10.1145/3589335.3651526

[35] V. Gabeff, M. Rußwurm, D. Tuia, and A. Mathis, "Wildclip: scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models," 2023.

[36] M. Hall, L. Gustafson, A. Adcock, I. Misra, and C. Ross, "Vision-language models performing zero-shot tasks exhibit gender-based disparities," 2023.

[37] M. Rahhal, Y. Bazi, H. ElGibreen, and M. Zuair, "Vision-language models for zero-shot classification of remote sensing images," *Applied Sciences*, vol. 13, p. 12462, 2023.

[38] L. Yuan, D. Chen, Y. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang, "Florence: a new foundation model for computer vision," 2021.

[39] P. Purwono, A. Ma'arif, W. Rahmaniar, H. Fathurrahman, A. Frisky, and Q. Haq, "Understanding of convolutional neural network (cnn): a review," *International Journal of Robotics and Control Systems*, vol. 2, pp. 739–748, 2023.

[40] R. Mo, "A survey of image classification algorithms based on convolution neural network," *Highlights in Science, Engineering and Technology*, vol. 15, pp. 191–198, 2022.

[41] X. Chen, "The study for convolutional neural network and corresponding applications," *Theoretical and Natural Science*, vol. 5, pp. 182–187, 2023.

[42] A. Alsajri and A. V. Hacimahmud, "Review of deep learning: convolutional neural network algorithm," *Babylonian Journal of Machine Learning*, vol. 2023, pp. 19–25, 2023.

[43] S. R. Dewi, F. Ramadhani, and S. Djasmayena, "Klasifikasi jenis jerawat berdasarkan gambar menggunakan algoritma cnn (convolutional neural network)," *Hello World Jurnal Ilmu Komputer*, vol. 3, pp. 68–73, 2024.

[44] S. Alamgunawan and Y. Kristian, "Klasifikasi tekstur serat kayu pada citra mikroskopik veneer memanfaatkan deep convolutional neural network," *Journal of Intelligent System and Computation*, vol. 2, pp. 06–11, 2021.

[45] T. Bariyah, M. A. Rasyidi, and N. Ngatini, "Convolutional neural network untuk metode klasifikasi multi-label pada motif batik," *Techno.Com*, vol. 20, pp. 155–165, 2021.

[46] D. L. Tyas, F. R. Rumambi, A. Patanduk, and R. C. J. Mailangkay, "Klasifikasi jenis tumor otak melalui citra mri dengan menggunakan convolutional neural network," *Informatik : Jurnal Ilmu Komputer*, vol. 21, pp. 26–34, 2025.

[47] C. M. Bachri and W. Gunawan, "Deteksi email spam menggunakan algoritma convolutional neural network (cnn)," *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, vol. 10, p. 88, 2024.

[48] M. F. Fauzi, "Klasifikasi image tinggi tanaman jagung dengan menggunakan algoritma convolution neural network (cnn)"klasifikasi image tinggi tanaman jagung dengan menggunakan algoritma convolution neural network (cnn)," *Jurnal Informatika Dan Teknik Elektro Terapan*, vol. 12, 2024.

[49] H. P. A. Tjahyaningtijas, M. A. Nashrullah, P. Puspitaningayu, L. Rakhmawati, Y. Yamasari, and J. R. Paragas, "Biomedical image classification using vision transformer," *E3S Web of Conferences*, vol. 640, p. 02021, 2025.

[50] X. Dai, Z. Li, L. Li, S. Xue, X. Huang, and X. Yang, "Hypertransxnet: learning both global and local dynamics with a dual dynamic token mixer for hyperspectral image classification," *Remote Sensing*, vol. 17, p. 2361, 2025.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[52] ——, "Identity mappings in deep residual networks," in *ECCV*, 2016.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[55] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," 2022. [Online]. Available: https://arxiv.org/abs/2201.03545

[56] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019. [Online]. Available: https://openai.com/blog/language-unsupervised

[57] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, N. Shinn, J. Schulman, D. Amodei *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[58] R. Bommasani, D. A. Hudson, S. Adolphs, and et al., "Opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[59] F. Petroni, T. Rocktäschel, P. Lewis, S. Riedel, A. Miller, and L. Zettlemoyer, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[61] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, D. Li, P. Dollár, and C. Feichtenhofer, "Perception encoder: The best visual embeddings are not at the output of the network," 2024, work from Meta FAIR.

[62] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," 2023.

[63] F. Chollet, *Deep Learning with Python*. Manning Publications, 2017.

[64] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[65] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.

[66] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[67] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019. [Online]. Available: https://arxiv.org/abs/1903.12261

[68] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning — the good, the bad and the ugly," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[69] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[71] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3498–3505.

[72] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.