REFERENCES

[1] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2011–2018.

[2] V. M. Araújo, A. S. Britto Jr, L. S. Oliveira, and A. L. Koerich, "Two-view fine-grained classification of plant species," *Neurocomputing*, vol. 467, pp. 427–441, 2022.

[3] Z. Fang, K. Kuang, Y. Lin, F. Wu, and Y.-F. Yao, "Concept-based explanation for fine-grained images and its application in infectious keratitis classification," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 700–708.

[4] W. Park and J. Ryu, "Fine-grained self-supervised learning with jigsaw puzzles for medical image classification," *Computers in Biology and Medicine*, vol. 174, p. 108460, 2024.

[5] Q. Cao, N. Du, L. Yu, M. Zuo, J. Lin, N. Liu, E. Zhong, Z. Liu, Q. Chen, Y. Shen *et al.*, "Practical fine-grained learning based anomaly classification for ecg image," *Artificial Intelligence in Medicine*, vol. 119, p. 102130, 2021.

[6] I. Baz, E. Yoruk, and M. Cetin, "Context-aware hybrid classification system for fine-grained retail product recognition," in *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2016, pp. 1–5.

[7] B. Santra, A. K. Shaw, and D. P. Mukherjee, "Part-based annotation-free fine-grained classification of images of retail products," *Pattern Recognition*, vol. 121, p. 108257, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image

is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[11] Y. Hu, X. Jin, Y. Zhang, H. Hong, J. Zhang, Y. He, and H. Xue, "Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4239–4248.

[12] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "Transfg: A transformer architecture for fine-grained recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 852–860.

[13] Y. Zhang, J. Cao, L. Zhang, X. Liu, Z. Wang, F. Ling, and W. Chen, "A free lunch from vit: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3234–3238.

[14] Q. Xu, J. Wang, B. Jiang, and B. Luo, "Fine-grained visual classification via internal ensemble learning transformer," *IEEE Transactions on Multimedia*, 2023.

[15] S. Ye, S. Yu, Y. Wang, and X. You, "R2-trans: Fine-grained visual categorization with redundancy reduction," *Image and Vision Computing*, vol. 143, p. 104923, 2024.

[16] J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid, "Convolutional kernel networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/81ca0262c82e712e50c580c032d99b60-Paper.pdf

[17] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8927–8948, 2021.

[18] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[19] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[20] F. Petersen, H. Kuehne, C. Borgelt, and O. Deussen, "Differentiable top-k classification learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 656–17 668.

[21] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[22] 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*, 2016, pp. 770–778.

[23] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5157–5166.

[24] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-cnn: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 579–590, 2019.

[25] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1025–1034.

[26] J. Wang, X. Yu, and Y. Gao, "Feature fusion vision transformer for fine-grained visual categorization," *British Machine Vision Conference*, 2021.

[27] H. Sun, X. He, and Y. Peng, "Sim-trans: Structure information modeling transformer for fine-grained visual categorization," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5853–5861.

[28] X. Hu, S. Zhu, and T. Peng, "Hierarchical attention vision transformer for fine-grained visual classification," *Journal of Visual Communication and Image Representation*, vol. 91, p. 103755, 2023.

[29] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[30] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 130–13 137.