

REFERENCES

- [1] J. Yu, M. Tan, H. Zhang, Y. Rui, and D. Tao, “Hierarchical deep click feature prediction for fine-grained image recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 563–578, 2022.
- [2] S. Huang, Z. Xu, D. Tao, and Y. Zhang, “Part-stacked cnn for fine-grained visual categorization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [3] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, “Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization,” *Pattern Recognition*, vol. 76, pp. 704–714, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320317303990>
- [4] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, “Learning to navigate for fine-grained classification,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [5] Y. Rao, G. Chen, J. Lu, and J. Zhou, “Counterfactual attention learning for fine-grained visual categorization and re-identification,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1005–1014.
- [6] M. Liu, C. Zhang, H. Bai, R. Zhang, and Y. Zhao, “Cross-part learning for fine-grained image classification,” *IEEE Transactions on Image Processing*, vol. 31, pp. 748–758, 2022.
- [7] K. Xu, R. Lai, L. Gu, and Y. Li, “Multiresolution discriminative mixup network for fine-grained visual categorization,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 7, pp. 3488–3500, 2023.
- [8] 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,” *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>

- [9] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, “Deepvit: Towards deeper vision transformer,” *arXiv preprint arXiv:2103.11886*, 2021.
- [10] K. Kim, B. Wu, X. Dai, P. Zhang, Z. Yan, P. Vajda, and S. Kim, “Rethinking the self-attention in vision transformers,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 3065–3069.
- [11] W. Luo, X. Yang, X. Mo, Y. Lu, L. Davis, J. Li, J. Yang, and S.-N. Lim, “Cross-x learning for fine-grained visual categorization,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8241–8250.
- [12] J. Wang, X. Yu, and Y. Gao, “Feature fusion vision transformer for fine-grained visual categorization,” in *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*. BMVA Press, 2021, p. 170. [Online]. Available: <https://www.bmvc2021-virtualconference.com/assets/papers/0685.pdf>
- [13] M. Tan, F. Yuan, J. Yu, G. Wang, and X. Gu, “Fine-grained image classification via multi-scale selective hierarchical biquadratic pooling,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 1s, Jan. 2022. [Online]. Available: <https://doi.org/10.1145/3492221>
- [14] Q. Xu, J. Wang, B. Jiang, and B. Luo, “Fine-grained visual classification via internal ensemble learning transformer,” *IEEE Transactions on Multimedia*, vol. 25, pp. 9015–9028, 2023.
- [15] X. Liu, L. Wang, and X. Han, “Transformer with peak suppression and knowledge guidance for fine-grained image recognition,” *Neurocomputing*, vol. 492, pp. 137–149, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231222004052>
- [16] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, “Cats and dogs,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [17] J. Song and R. Yang, “Feature boosting, suppression, and diversification for fine-grained visual classification,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.
- [18] G. Sun, H. Cholakkal, S. Khan, F. Khan, and L. Shao, “Fine-grained recognition: Accounting for subtle differences between similar

- classes,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 047–12 054, Apr. 2020. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/6882>
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: <https://doi.org/10.18653/v1/n19-1423>
- [21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [22] J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467–4480, 2020.
- [23] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, “Transfg: A transformer architecture for fine-grained recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 852–860, Jun. 2022. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/19967>
- [24] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *Proceedings of the Thirteenth International Conference on International*

Conference on Machine Learning, ser. ICML'96. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996, p. 148–156.

- [25] M. Hossin and S. M.N, “A review on evaluation metrics for data classification evaluations,” *International Journal of Data Mining Knowledge Management Process*, vol. 5, pp. 01–11, 03 2015.
- [26] F. Petersen, H. Kuehne, C. Borgelt, and O. Deussen, “Differentiable top-k classification learning,” in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 17 656–17 668. [Online]. Available: <https://proceedings.mlr.press/v162/petersen22a.html>
- [27] Y. Peng, X. He, and J. Zhao, “Object-part attention model for fine-grained image classification,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018.
- [28] M. Simon and E. Rodner, “Neural activation constellations: Unsupervised part model discovery with convolutional networks,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1143–1151.
- [29] J. Feng, B. Price, S. Cohen, and S.-F. Chang, “Interactive segmentation on rgb-d images via cue selection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.