References

* The highlighted references were mentioned in the proposal *

[1] Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., ... & Ballas, N. (2022). Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*.

[2] Bachmann, R., Mizrahi, D., Atanov, A., & Zamir, A. (2022). MultiMAE: Multi-modal Multi-task Masked Autoencoders. *arXiv preprint arXiv:2204.01678*.

[3] Bao, H., Dong, L., & Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

[4] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, *33*, 9912-9924.

[5] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9650-9660).

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR.

[7] Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15750-15758).

[8] Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9640-9649).

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

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[11] Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., & Liu, Z. (2021). Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*.

[12] Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

[13] Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789-1819.

[14] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., ... & He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

[15] Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, *33*, 21271-21284.

[16] Hadsell, R., Chopra, S., & LeCun, Y. (2006, June). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 1735-1742). IEEE.

[17] Hao, Y., Dong, L., Wei, F., & Xu, K. (2021, May). Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 14, pp. 12963-12971).

[18] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000-16009).

[19] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729-9738).

[20] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... & Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8340-8349).

[21] Jing, L., Zhu, J., & LeCun, Y. (2022). Masked siamese convnets. *arXiv preprint arXiv:2206.07700*.

[22] Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., ... & Gao, J. (2021). Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*.

[23] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[24] Liu, X., Zhou, J., Kong, T., Lin, X., & Ji, R. (2022). Exploring Target Representations for Masked Autoencoders. *arXiv preprint arXiv:2209.03917*.

[25] Liu, Y., Zhang, W., & Wang, J. (2020). Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, *415*, 106-113.

[26] Misra, I., & Maaten, L. V. D. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6707-6717).

[27] Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3967-3976).

[28] Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, *30*.

[29] Tian, Y., Chen, X., & Ganguli, S. (2021, July). Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning* (pp. 10268-10278). PMLR.

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[31] Wei, C., Fan, H., Xie, S., Wu, C. Y., Yuille, A., & Feichtenhofer, C. (2022). Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14668-14678).

[32] Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., ... & Guo, B. (2022). Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation. *arXiv preprint arXiv:2205.14141*.

[33] Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, *34*, 30392-30400.

[34] Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10687-10698).

[35] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., ... & Hu, H. (2022). Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9653-9663).

[36] Yuan, L., Hou, Q., Jiang, Z., Feng, J., & Yan, S. (2022). Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[37] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.

[38] Zhou, P., Zhou, Y., Si, C., Yu, W., Ng, T. K., & Yan, S. (2022). Mugs: A Multi-Granular Self-Supervised Learning Framework. *arXiv preprint arXiv:2203.14415*.