# Supplemental Material

Haoquan Li*[1], Laoming Zhang*[2], Daoan Zhang[1], Lang Fu[2], Peng Yang[2,3], and Jianguo Zhang†[1,4]

[1] Research Institute of Trustworthy Autonomous Systems, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
[2] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
[3] Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China
[4] Peng Cheng Lab, Shenzhen, China
{12032492, 12032505, 12032503, 12032485}@mail.sustech.edu.cn, {yangp, zhangjg}@sustech.edu.cn

## 1 Implementation Details

### 1.1 Pre-training

The default setting is in Table 1. Most of the configurations are consistent with MAE [9]. For example, a standard ViT [6] model is used as an encoder. Due to the small dataset and GPU limitation, we adjust batch size to 128. We use xavier_uniform [7] to initialize all transformer blocks and use the linear lr scaling rule [8]: $lr = base\_lr \times batch\_size/256$.

### 1.2 Fine-tuning

Our fine-tuning strategy follows the common practice of supervised ViT training. The default setting is in Table 4. We adjust the learning rate to 7e-4 and use layer-wise lr decay [4] following BEiT [1].

## 2 Image Reconstruction

The target of MAE pre-training is to reconstruct masked patches, and the image quality of the reconstruction can reflect the degree of pre-training learning. To explore how well the MAE pre-trained model performs on its own and other datasets, we visualize them in Fig. 1.

From the figure, MAE pre-trained models produce distinct reconstructions of images from the same dataset (pre-trained dataset) and other datasets. Meanwhile, the model pre-trained on *mini*ImageNet performs better than on CIFAR-FS. It is consistent with our previous cross-domain results. This is probably

* H. Li and L. Zhang made equal contributions to this work.
† J. Zhang is the corresponding author.

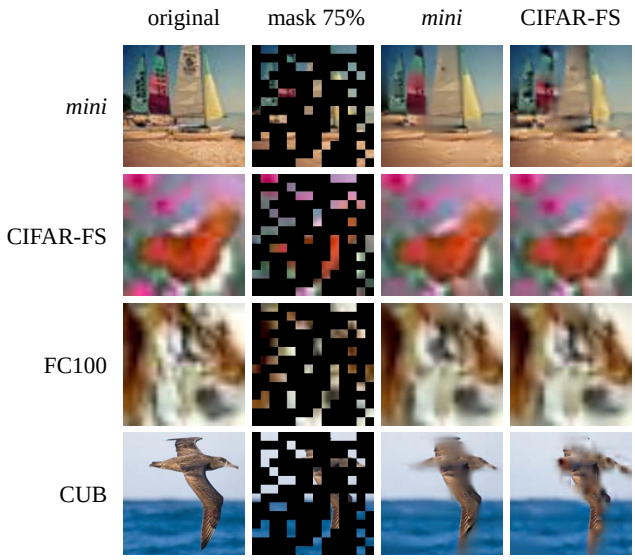


**Fig. 1.** The visualizations of image reconstructions by MAE pre-trained models. Each image is first resized and cropped into the size of 224x224 before masking, and then reconstructed by two models pre-trained on the training set of *mini*ImageNet and CIFAR-FS respectively. Each row represents the data source of the image (from validation set) to be reconstructed. Columns from left to right represent original images, masked images, reconstructions of model pre-trained on *mini*ImageNet, and reconstructions of model pre-trained on CIFAR-FS

**Table 1.** Pre-training setting

| config | value |
|---|---|
| image resize | $224 \times 224$ |
| patch size | 16 |
| optimizer | AdamW [14] |
| base learning rate | 1.5e-4 |
| weight decay | 0.05 |
| optimizer momentum | $\beta_1, \beta_2 = 0.9, 0.95$ [3] |
| batch size | 128 |
| training epochs | 1600 |
| learning rate schedule | cosine decay [13] |
| warmup epochs [8] | 40 |
| augmentation | RandomResizedCrop |
| masking ratio | 0.75 |
| encoder layers | 12 |
| encoder dim | 768 |
| decoder layers | 4 |
| decoder dim | 384 |

because the high-resolution images contains more information which is helpful for MAE pre-training.

## 3 Comparison of ViT and ResNet for Few-Shot Learning

We compare ViT with ResNet (an architecture based on CNN) on a few-shot dataset in Table 2. From the table, ResNet is less prone to overfitting and performs better with the same number of parameters, while ViT overfit easily. This is due to the local prior knowledge in CNN. Therefore, to avoid overfitting, ViT often need a large amount of target data or to be pre-trained in advance.

**Table 2.** Comparisons of ResNet [10] and ViT [6] for few-shot learning. For a fair comparison, we make the numbers of parameters for ResNet and ViT close. Then we test their accuracies for base classes and novel classes (1-shot, 5-shot). Base classes are those classes for training

| model | params | base classes | 1-shot | 5-shot |
|---|---|---|---|---|
| ResNet12 | 8 M | 82.34 | 58.89 | 78.21 |
| ResNet101 | 86 M | 76.61 | 55.98 | 73.47 |
| ViT-Tiny | 8 M | 58.52 | 51.46 | 68.08 |
| ViT-Base | 86 M | 26.22 | 34.55 | 45.73 |

## 4 The Effect of Grouping on NeXtVLAD

**Table 3.** The influence of grouping number on the number of NeXtVLAD parameters and *mini*ImageNet 5-way performance

| Grouping | Params | 1-shot | 5-shot |
|---|---|---|---|
| 1 | 153.8M | 68.44±0.60 | 83.36±0.11 |
| 4 | 39.0M | 69.18±0.81 | 84.40±0.09 |
| 8 | 22.9M | 68.24±0.60 | 84.42±0.24 |
| 12 | 17.4M | 68.35±0.58 | 84.82±0.07 |

Grouping was proposed by NeXtVLAD[12] to reduce the number of parameters without decreasing performance. Without the grouping method, the last projection layer of NeXtVLAD will contain a huge amount of parameters. We test the effect of grouping on performance in Table 3. It could be seen that NeXtVLAD without grouping (grouping number 1) still works well (although with a huge number of parameters); and more importantly, when the grouping

number increases, the performance remains very reliable, however, the number of parameters is significantly reduced.

## 5 Future Work

Although our method has improved a lot compared to the baseline (MAE pre-training and fine-tuning directly), there is still a long way to go for the combination of few-shot learning and transformer-based models. First, the potential of transformers for cross-domain few-shot learning has not been fully released. More in-depth researches should be carried out, e.g., under the Meta-Dataset scenario [16]. Second, random masking can eliminate supervision bias. It is possibly able to be combined with selected masking [2] method which focuses more on the key patches useful for representing images. Finally, NeXtVLAD is a module proposed for frame-level representations aggregation. There should be a more appropriate aggregation method for patch-level features. In addition, this paper studied in the inductive few-shot setting, where no extra unlabeled data is considered. Another popular setting, the transductive few-shot setting [18], where test data can be taken into account during training by seeing them as unlabeled data, can be further studied.

**Table 4.** Fine-tuning setting

| config | value |
|---|---|
| image resize | $224 \times 224$ |
| patch size | 16 |
| optimizer | AdamW |
| base learning rate | 7e-4 |
| weight decay | 0.05 |
| optimizer momentum | $\beta_1, \beta_2 = 0.9, 0.999$ |
| layer-wise lr decay [4,1] | 0.75 |
| batch size | 128 |
| training epochs | 100 |
| learning rate schedule | cosine decay |
| label smoothing [15] | 0.1 |
| mixup [19] | 0.8 |
| cutmix [17] | 1 |
| augmentation | RandAug (9, 0.5) [5] |
| drop path [11] | 0.1 |
| masking ratio | 0.7 |
| soft focal loss $\gamma$ | 2 |
| NeXtVLAD $\lambda$ | 4 |
| NeXtVLAD $K$ | 64 |
| NeXtVLAD $G$ | 8 |

## References


1. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Chen, H., Li, H., Li, Y., Chen, C.: Shaping visual representations with attributes for few-shot learning. arXiv preprint arXiv:2112.06398 (2021)
3. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020)
4. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
8. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: European conference on computer vision. pp. 646–661. Springer (2016)
12. Lin, R., Xiao, J., Fan, J.: Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
13. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
15. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems **32** (2019)
16. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.A., et al.: Meta-dataset: A dataset of datasets for learning to learn from few examples. arXiv preprint arXiv:1903.03096 (2019)
17. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)

18. Zhang, B., Li, X., Ye, Y., Huang, Z., Zhang, L.: Prototype completion with primitive knowledge for few-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3754–3762 (2021)
19. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)