Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification

(Supplementary Materials)

Fusheng Hao[1,2] Fengxiang He[3] Liu Liu[4] Fuxiang Wu[1,2] Dacheng Tao[4] Jun Cheng[1,2*]

[1] Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China

[2] The Chinese University of Hong Kong, Hong Kong, China

[3] AIAI, School of Informatics, University of Edinburgh, United Kingdom

[4] School of Computer Science, Faculty of Engineering, The University of Sydney, Australia

Additional details or results are collected in this appendix.

A. Datasets

Experiments are conducted on four popular benchmark datasets: miniImageNet [20, 15], tieredImageNet [16], CIFAR-FS [1], and FC-100 [13].

miniImageNet. This dataset is very popular in few-shot image classification. It is a subset of ImageNet [17] and contains 100 classes with 600 images per class. We use the split provided in [15] to divide the 100 classes into disjoint 64 training, 16 validation, and 20 test classes.

tieredImageNet. This dataset is also a subset of ImageNet [17] and the hierarchical structure of ImageNet is used when it is created. It includes 608 classes from 34 super-classes, with a total of 779,165 images. The 34 super-classes are divided into disjoint 20 training, 6 validation and 8 test super-classes to achieve better separation. Correspondingly, the 608 classes are divided into disjoint 351 training, 97 validation, and 160 test classes.

CIFAR-FS. This dataset is created based on CIFAR100 [8]. It includes 100 classes with 600 images per class. The 100 classes are divided into disjoint 64 training, 16 validation, and 20 test classes.

FC-100. This dataset is also created based on CIFAR100 [8] and includes 100 classes with 600 images per class. It uses a split strategy similar to tieredImageNet to increase the difficulty of the resulting few-shot image classification tasks. Correspondingly, the 100 classes are divided into disjoint 60 training, 20 validation, and 20 test classes.

B. Implementation details

Our training procedure consists of two stages: pretraining and finetuning. It is to be noted that only the training set of the corresponding dataset is used for pretraining and finetuning.

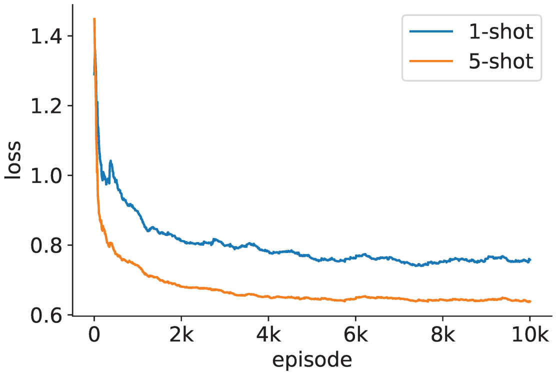
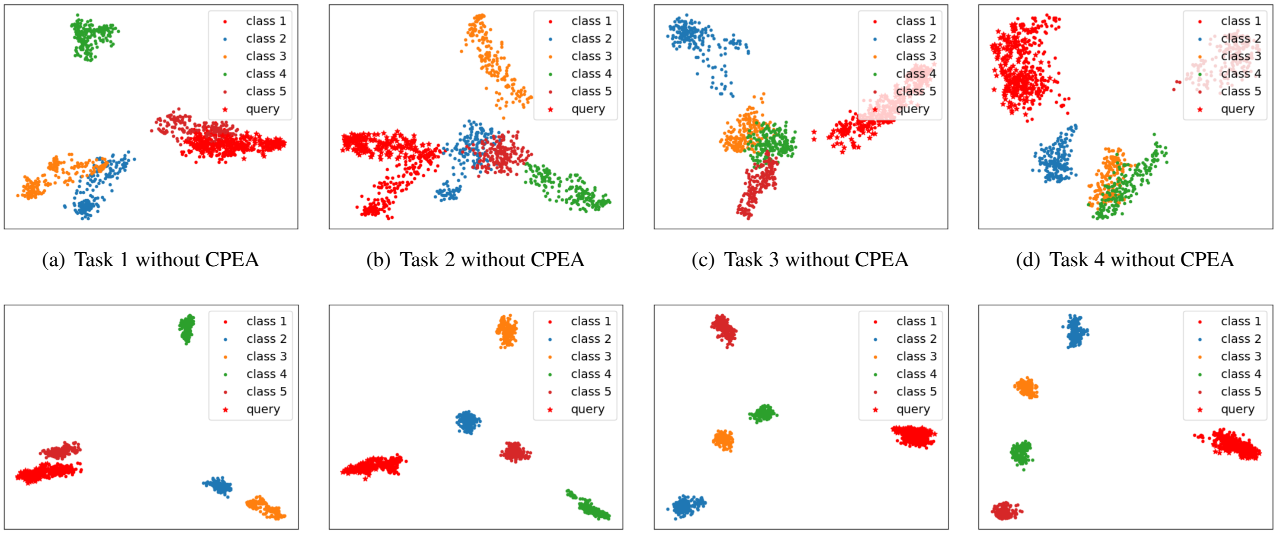


Figure A1. Illustration of finetuning losses of our CPEA. Experiments are conducted on miniImageNet.

Pretraining. We pretrain the backbone, *i.e.*, ViTS/16, by using the strategy proposed in [25, 6] and mostly sticking to the hyperparameter settings reported. Specifically, two global crops with a crop scale of (0.4, 1.0) and 10 local crops with a crop scale of (0.05, 0.4) are used. The output dimension of the project head is set to be 8,192-d. The prediction ratios and variances of random Masked Image Modelling is set to be (0, 0.3) and (0, 0.2), respectively. The resolution of input images is set to be 224×224. Four A100 40G GPUs are used to pretrain the ViT-S/16 and the total number of training epochs is set to be 1,600. To match our computing resources, the batch size is set to be 512. The optimizer used is AdamW [9] and the linearly ramped-up learning rate of $5\mathrm{e}{-4} \times \mathrm{batchsize}/256$ is carried out in the first 10 epochs.

Finetuning. After pretraining, the project head with an output dimension of 8,192-d is removed and a new project head with an output dimension of 384-d is added to the ViTS/16. The whole pipeline is finetuned by minimizing Eq. (6) using the episodes randomly sampled from the train-

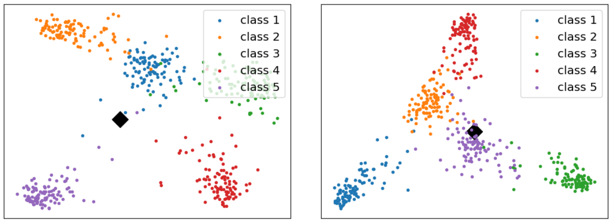
[*] Corresponding author (email: jun.cheng@siat.ac.cn).



(a) Task 1 without CPEA (b) Task 2 without CPEA (c) Task 3 without CPEA (d) Task 4 without CPEA

(e) Task 1 with CPEA (f) Task 2 with CPEA (g) Task 3 with CPEA (h) Task 4 with CPEA

Figure A2. Patch embedding visualization of four more randomly sampled 5-way 1-shot classification tasks with one query image per class. (a), (b), (c), and (d) show the visualization results without CPEA. (e), (f), (g), and (h) show the corresponding visualization results with CPEA. CPEA concentrates patch embeddings by class, thus making them class-relevant. Experiments are conducted on miniImageNet.



(a) (b)

Figure A3. Class-aware embedding visualization. Two more different sampling results are given in (a) and (b), respectively, with 100 class-aware embeddings per class. The class-agnostic embedding is denoted by the black "diamond". After interacting with images from different classes, the output states of class-agnostic embedding are class-aware. Experiments are conducted on miniImageNet.

ing classes. The optimizer used is Adam [7]. The global initial learning rate is set to be 0.001, which is halved every 500 episodes, and the learning rate of the ViT-S/16 is always kept to be one percent of the global learning rate. The weight decay is set to be 0.001 and the total number of episodes is set to be 10,000.

C. Ablation study

Convergence property. Figure A1 shows the finetuning losses of our CPEA. It can be observed that our CPEA converges after 8k episodes in both the 1-shot and 5-shot settings. These observations indicate that our CPEA could converge in a proper time.

Patch embedding visualization. More patch embedding visualization results are shown in Figure A2. It can be observed that with CPEA, the patch embeddings are clustered by class. This means that the patch embeddings are made class-relevant by CPEA. It is to be noted that PCA is used for visualization.

Method	Backbone	#Params	5-shot
ProtoNet [19]	ResNet-12	≈ 12.4M	79.46±0.48
FEAT [22]	ResNet-12	≈ 12.4M	82.05±0.14
DeepEMD [23]	ResNet-12	≈ 12.4M	82.41±0.56
COSOC [10]	ResNet-12	≈ 12.4M	85.16±0.42
DeepBDC [21]	ResNet-12	≈ 12.4M	84.46±0.28
LEO [18]	WRN-28-10	≈ 36.5M	77.59±0.12
CC+rot [5]	WRN-28-10	≈ 36.5M	79.87±0.33
FEAT [22]	WRN-28-10	≈ 36.5M	81.11±0.14
PSST [3]	WRN-28-10	≈ 36.5M	80.64±0.32
MetaQDA [24]	WRN-28-10	≈ 36.5M	84.28±0.69
OM [14]	WRN-28-10	≈ 36.5M	85.29±0.41
FewTURE [6]	ViT-T/16	≈ 5.0M	81.10±0.61
FewTURE [6]	ViT-S/16	≈ 22.0M	84.51±0.53
CPEA (ours)	ViT-T/16	≈ 5.0M	84.62±0.39
CPEA (ours)	ViT-S/16	≈ 22.0M	**87.06±0.38**

Table A1. Impact of the model size on the few-shot classification performance. Experiments are conducted on miniImageNet.

Class-aware embedding visualization. More class-aware embedding visualization results of images from different classes are shown in Figure A3. After interacting with images from different classes, the output states of class-agnostic embedding are class-aware. It is to be noted that PCA is used for visualization.

Model size. Since the few-shot training sets are com-

Model	Image resolution	5-shot
DeepEMD [23]	84×84 → 224×224	82.41 → 78.12
BML [26]	84×84 → 224×224	83.59 → 81.57
IE [12]	84×84 → 224×224	84.35 → 79.12
CPEA (ours)	224×224	87.06

Table A2. Impact of the image resolution on the few-shot classification performance. Experiments are conducted on miniImageNet.

parably small (*e.g.* 38.4K images in miniImageNet [20] vs. 1.28M images in ImageNet [17]), it has been shown that adopting bigger networks does not help improve performance [2, 11]. In other words, increasing the number of parameters on its own does not lead to better few-shot classification performance. Table A1 shows the impact of the model size on few-shot classification performance. It can be observed that 1) With less than one seventh of the number of parameters of WRN-28-10, our CPEA achieves comparable performance. 2) With roughly the same number of parameters of WRN-28-10, our CPEA outperforms the counterparts by a large margin. 3) Our CPEA benefits from the increased model size.

Image resolution. We follow the common practice of ViTs [4] to take images with a resolution of 224×224 as input, which is higher than traditional CNN-based few-shot image classification methods [23, 26, 12] (*e.g.*, 84×84). It is to be noted that increasing the image resolution on its own does not lead to better few-shot classification performance. Table A2 shows the impact of the image resolution on the few-shot classification performance. It can be observed that higher resolution always leads to performance degradation.

References

[1] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[3] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. Pareto self-supervised training for few-shot learning. In *CVPR*, 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[5] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *ICCV*, 2019.

[6] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. In *NeurIPS*, 2022.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[8] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[10] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background for few-shot learning. In *NeurIPS*, 2021.

[11] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020.

[12] Rizve Mamshad Nayeem, Khan Salman, Khan Fahad Shahbaz, and Shah Mubarak. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *CVPR*, 2021.

[13] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

[14] Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *ICCV*, 2021.

[15] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[16] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv:1803.00676*, 2018.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[18] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2018.

[19] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[20] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

[21] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *CVPR*, 2022.

[22] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, 2020.

[23] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *CVPR*, 2020.

[24] Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. Shallow bayesian meta learning for real-world few-shot recognition. In *ICCV*, 2021.

[25] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022.

[26] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *ICCV*, 2021.