Visual semantic alignment network based on pre-trained ViT for few-shot image classification

Jiaming Zhang* ,Jijie Wu* and Xiaoxu Li*
* University of Lanzhou University of Technology,Lanzhou,China
E-mail: lixiaoxu@lut.edu.cn

***Abstract*—Images often contain a wealth of information, but not all the information in a picture is related to image classification, and those modules that are not related to classification may be used as interference during classification, affecting the classification ability of the model. To solve this problem, we propose a class-aware feature alignment adaptive module (CAFASA). The key idea of CAFASA is first to combine patch embedding and class-aware embedding and add a feature alignment module on top of this foundation to help the model obtain better classification performance. This method is significantly superior to the existing state-of-the-art methods in CIFAR-FS, miniImageNet, tieredImageNet, and FC-100.**

I. Introduction

In most of the existing datasets, the representation of the image is single, and only part of the image information is included, for example, on the CIFRA[1] and miniImageNet[2] datasets, only part of the content of the image is described. Although annotations may be incomplete, this approach is acceptable in some cases, such as when there are enough tagged images per class, which vary significantly within the class to cover different situations within the class, so that even if the annotation focuses on only a portion of the image, the model can learn a large amount of data to identify and classify the pictures. However, in the case of small-shot classification, the model needs to identify new classes that are different from the training stage. Each class has only a minimal number of labeled images, which makes it difficult for the model to confirm which entities are the critical information to determine the image, resulting in a decrease in the generalization ability of the model.

A promising approach is VITs ability to capture complex relationships between different regions within an image through a self-attention mechanism, which is widely used in small-shot classification, and ViTs excels in tasks such as image detection and object classification[3]. ViTs typically require a large amount of data to train due to the lack of convolutional inductive bias found in CNNs. Some approaches use a single Transformer head in combination with a CNN[4], [5] to enhance the model's performance in few-shot learning, and recently feature[6] provides an effective solution that enables a fully ViT-based architecture to be successfully generalized to small-scale image datasets.

Another common approach is to align semantically relevant regions[6]–[8] . For example, SAML[9] uses an activation-based attention mechanism to highlight regions related to categories and suppress others, thereby enhancing the model's ability to generalize new categories. CNA[10], i.e., cross-attention network, the core idea of which is to enhance the model's ability to recognize category-related regions in the image through the attention mechanism. DeepEMD[7] uses differentiable EMD as a metric to calculate the structural distance between image regions to determine the correlation of images. CXT uses Transformer[8] to enhance further the expressive ability of features through the self-attention mechanism so that the model can capture the complex relationship between different regions in the imageFeature method dynamically selects the most informative region in the image. The weights of these regions in the feature representation are readjusted to optimize category recognition. While these techniques do a great job of reducing noise and avoiding over-reliance on training data, they also face challenges where the aligned regions may not be useful areas, and secondly, because the small number of samples in the new category means that the model may mistakenly focus on those regions that are not relevant to the target class, affecting the accuracy of the classification.

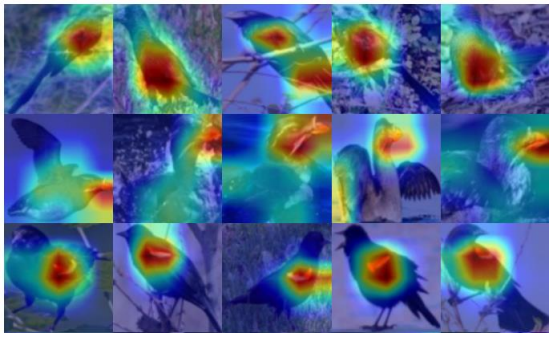

Fig. 1. The localized region visualization of birds in the miniImageNet dataset obtained using the method of this paper shows that the areas with higher energy in the images represent the more discriminative parts.

In this paper, we propose an innovative Category-Aware Feature Alignment Adaptive Module (CAFASA) that makes patch embedding class-related by integrating patch embedding with class-aware embedding[11], after which we introduce a feature alignment module, Reconstruct query set features with support sets and query sets to further process and refine features. As shown in Figure 1, based on the local area visualization returned by our model in miniImageNet, it can be observed that the feature alignment module enhances the model's ability to recognize category features. At the same time, to avoid the

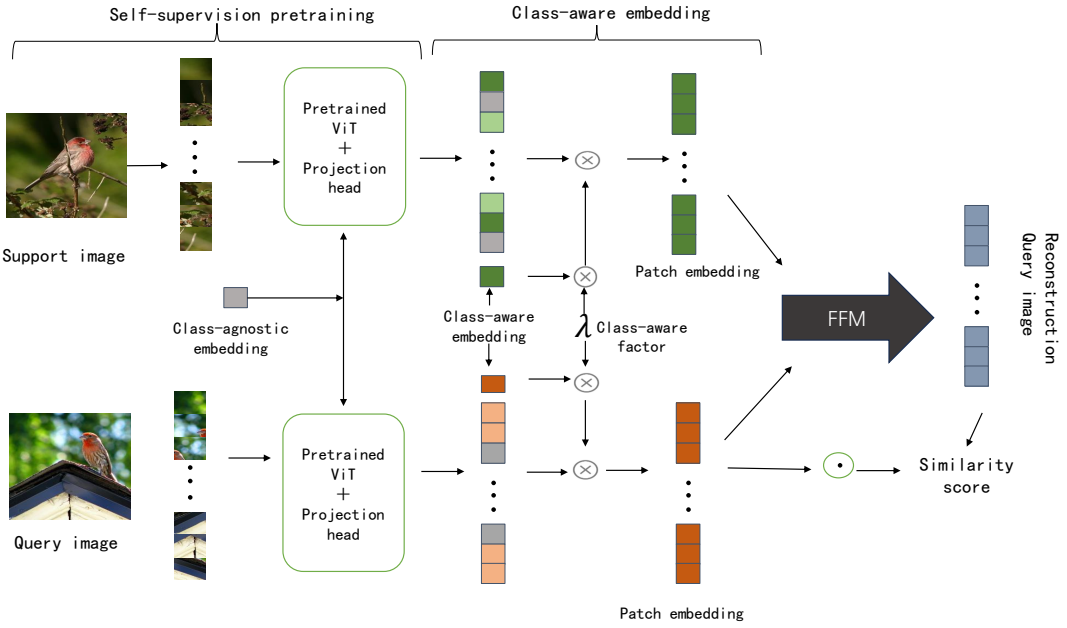

Fig. 2. illustrates the framework of the class-aware Feature fusion method. The Vision Transformer (ViT) is initially pre-trained using the Masked Image Modeling (MIM) approach. Subsequently, a class-aware embedding module is introduced, which interacts and learns continuously with the image patch embeddings within the ViT. At the model's output end, these patch embeddings are adjusted in conjunction with the class-aware embeddings to ensure they are associated with specific categories. Finally, we introduce a Feature fusion method that utilizes the reconstructed features for similarity measurement.

collapse of supervision, we use self-supervised pre-training instead of traditional pre-training. Specifically, this approach generates semantically meaningful patch embeddings through Masked Image Modeling (MIM)[12] as an excuse task. We introduce a category-independent embedding that is fed into the Transformer model along with the patch. This embedding retains the semantic information of the original feature and enhances its category relevance.

Our main contributions are summarized as follows:

1) We take an innovative approach to deal with the problems caused by single-label annotations in the few-shot learning environment.
2) By introducing category-aware embedding, CAFASA can generate feature representations that are relevant to categories. This enables the model to capture the differences between categories more accurately when performing classification, improving the accuracy and reliability of classification.
3) The feature information of the support set is used to enhance the query set, and the features of the query set are made more prosperous and more complete through feature alignment.

II. Method

In this section, we first define the small sample problem and then describe in detail the process of class-aware patch embedding adaptation and feature Feature fusion.

A. problem define

Few-shot image classification aims to apply the patterns learned from the limited training effectively set C_{train} to the test set C_{test}, $C_{train} \cap C_{test} = \emptyset$. In this process, we follow an N-way K-shot classification task. N represents the total number of categories, K represents the number of labeled images available in each category, and Q represents the number of images used for testing in each category. Precisely, it consists of two parts: the support set contains N categories, each with K images $X_s = (x_i, y_i)_{i=1}^{NK}$, which are used to train the model; the query set also contains N categories, but each category has only Q images, which are used to evaluate the generalization ability of the model $X_q = (x_i, y_i)_{i=1}^{NQ}$ so that a model that can learn from a small number of samples and generalize to new categories can be trained so that accurate predictions can still be made when facing new and unseen categories.

B. overview

As shown in Figure 2, We first divide the input image into multiple small patches and map these patches into a high-dimensional feature space through linear projection. To preserve the spatial structure information of the picture, we add position encoding to the feature representation of these patches. In addition, we introduce a category-independent embedding vector to enhance the model's understanding of the overall characteristics of the image. Based on this innovative process, we also introduce a Feature fusion mechanism. Feature fusion generates richer feature representations through the interaction and relearning of internal features.

C. Class-aware patch embedding adaptation

Self-supervised pre-training: In the few-shot image classification task, traditional supervised pre-training methods may cause overfitting[6], resulting in the inability of training data to generalize effectively to new categories. To address this problem, we adopt an innovative self-supervised pre-training method. Specifically, we divide the image into small patches, randomly mask some areas, and use a Vision Transformer (ViT) to encode these image patches to fuse the masked regions. This method, called Masked Image Modeling (MIM)[12], [13], not only generates semantically meaningful patch embedding[14]s but also forces the model to understand the overall structure and content of the image through the feature fusion task. Unlike self-supervised methods that rely on specific training courses, MIM aims to improve the model's deep understanding of images, thereby enhancing its generalization ability when facing new categories. The advantage of this method is that it reduces the dependence on limited labeled data while providing a more robust feature.

Class patch embedding: After the previous pre-training, a large number of patch embeddings can be obtained and processed in the model's feedforward process. However, these embeddings may not be related to any specific category at first. Still, by interacting with the class tokens in the model, they can be adjusted to reflect the characteristics of a particular category. That is, by interacting with the class token, the patch embedding can be adjusted to have class awareness and reflect the characteristics of a specific category. The pre-training process generates patch embeddings for a large number of input images, and the semantic content of these embeddings may not be directly associated with a specific category. To improve the category relevance of these embeddings, we propose a class-aware embedding adjustment method. First, the class tokens are category-agnostic before being input into ViT. Second, by interacting with the patch embeddings in ViT, the class tokens can gradually gain category awareness, and their final state can reflect the category information of the image. Although we can extract embeddings of many small blocks (i.e., patches) from an image, these embeddings themselves may not carry valuable information about the image category. Some methods directly use the obtained patches as images to alleviate the problem of the number of pictures, but the effect is not evident despite the improvement[15], [16]. We start from a new perspective and combine patch embeddings with class-aware embeddings:

$$\bar{z}_i = z_i + \lambda z_{class} \tag{1}$$

Where $\bar{z}_i$ represents the adaptive patch, z_i represents the i-th original patch extracted directly from the input data, and z_{class} represents the information related to the category. $\lambda > 0$ represents the class-aware factor of the correlation size, which adjusts the correlation between adaptive and class-aware embeddings. The larger the value of λ, the higher the correlation between the adapted patch embedding and the class of interest. This helps the model to utilize the information of known categories better when dealing with unknown categories.

D. FFM

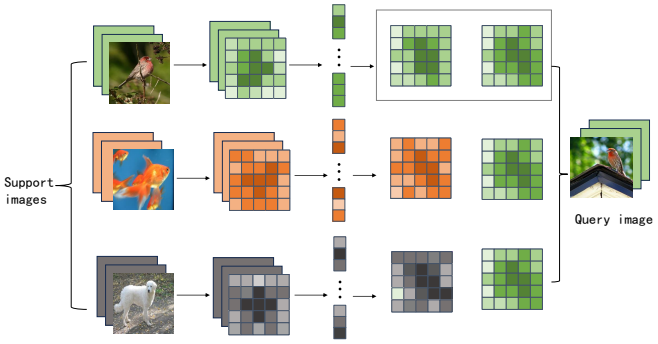

Fig. 3. Feature fusion module

For a C-way, K-shot classification task, after the initial patch input, we can obtain the support features of the Cth class, denoted as $F_s = [\bar{z}_i^s] \in \mathbb{R}^{r \times d}$, where d represents the number of channels, and r is the feature spatial resolution, which is the height of the image multiplied by its width. For each class c, we aggregate all features from the k support images into a single support image matrix, and the query features $F_q = [\bar{z}_j^q] \in \mathbb{R}^{kr \times d}$, Attention score S_{ij} is calculated as follows:

$$S_{ij} = \frac{(\bar{z}_j^q)^T \cdot \bar{z}_i^s}{\sqrt{D}} \cdot w \tag{2}$$

where w is the learnable weight parameter.

$$P_{ij} = \frac{\exp(S_{ij})}{\sum_{i=1}^{r} \exp(S_{ij})} \tag{3}$$

The attention probability P_{ij} is obtained by applying the softmax function to the attention scores S_{ij}, where the softmax function converts the raw scores into a probability distribution.

$$\hat{z}_j^q = \sum_{i=1}^{r} P_{ij} \bar{z}_i^s \tag{4}$$

where i indicates the ith support set feature, and j indicates the jth query set feature. The reconstructed query feature vector $\hat{Q}_i$ can be expressed as:

$$\hat{Q}_i = \begin{bmatrix} \hat{z}_1^q \\ \vdots \\ \hat{z}_r^q \end{bmatrix} \tag{5}$$

For the ith feature of the query set, the similarity matrix M between the reconstructed query set feature matrix $\hat{Q}_i$ and the original query set Q_i can be represented as:

$$M = \hat{Q}_i^T Q_i \tag{6}$$

We can define a loss function to measure the difference between $\hat{Q}_i$ and Q_i, such as the Mean Squared Error (MSE) loss:

TABLE I

TABLE 1: CLASSIFICATION ACCURACY OF SMALL SAMPLES SET BY THIS METHOD ON MINIIMAGENET, TIEREDIMAGENET, CIFAR-FS, FC100 FOR 5-WAY 1-SHOT AND 5-WAY 5-SHOT.

Model	**Backbone**	$\approx \#Params$	**miniImageNet**		**tieredImageNet**		**CIFAR-FS**		**FC100**	
			1-shot	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
SetFeat [17]	SetFeat-12	12.3 M	68.32±0.62	82.71±0.46	73.63±0.88	87.59±0.57	-	-	-	-
ProtoNet [18]	ResNet-12	12.4 M	62.29±0.33	79.46±0.48	68.25±0.23	84.01±0.56	-	-	41.54±0.76	57.08±0.76
FEAT [19]	ResNet-12	12.4 M	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16	-	-	-	-
DeepEMDT [7]	ResNet-12	12.4 M	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58	-	-	-	-
IEPT [20]	ResNet-12	12.4 M	67.05±0.44	82.90±0.30	72.24±0.50	86.73±0.34	-	-	-	-
MELR [21]	ResNet-12	12.4 M	67.40±0.43	83.40±0.28	72.14±0.51	87.01±0.35	-	-	-	-
FRN [22]	ResNet-12	12.4 M	66.45±0.19	82.83±0.13	72.06±0.22	86.89±0.14	-	-	-	-
CG [23]	ResNet-12	12.4 M	67.02±0.20	82.32±0.14	71.66±0.23	85.50±0.15	73.00±0.70	85.80±0.50	-	-
DMF [24]	ResNet-12	12.4 M	67.76±0.46	82.71±0.31	71.89±0.52	85.96±0.35	-	-	-	-
InfoPatch [25]	ResNet-12	12.4 M	67.67±0.45	82.44±0.31	-	-	-	-	-	-
BML [26]	ResNet-12	12.4 M	67.04±0.63	83.63±0.29	68.99±0.50	85.49±0.34	73.45±0.47	88.04±0.33	-	-
CNL [27]	ResNet-12	12.4 M	67.96±0.98	83.36±0.51	73.42±0.95	87.72±0.75	-	-	-	-
Meta-NVG [28]	ResNet-12	12.4 M	67.14±0.80	83.82±0.51	74.58±0.88	86.73±0.61	74.63±0.91	74.63±0.91	46.40±0.81	61.33±0.71
PAL [29]	ResNet-12	12.4 M	69.37±0.64	84.40±0.44	72.25±0.72	86.95±0.47	-	-	-	-
COSOC [30]	ResNet-12	12.4 M	69.28±0.49	85.16±0.42	73.57±0.43	87.57±0.10	-	-	-	-
Meta DeepBDC [23]	ResNet-12	12.4 M	67.34±0.43	84.46±0.28	72.34±0.49	87.31±0.32	-	-	-	-
LEO [31]	WRN-28-10	36.5 M	61.76±0.08	77.59±0.12	66.33±0.05	81.44±0.09	-	-	-	-
CC+rot [32]	WRN-28-10	36.5 M	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36	73.62±0.31	86.05±0.22	-	-
FEAT [19]	WRN-28-10	36.5 M	65.10±0.20	81.11±0.14	70.41±0.23	84.38±0.16	-	-	-	-
PSST [33]	WRN-28-10	36.5 M	64.16±0.44	80.64±0.32	-	-	77.02±0.38	88.45±0.35	-	-
MetaQDA [34]	WRN-28-10	36.5 M	67.83±0.64	84.28±0.69	74.33±0.65	89.56±0.79	75.83±0.88	88.79±0.75	-	-
OM [35]	WRN-28-10	36.5 M	66.78±0.30	85.29±0.41	71.54±0.29	87.79±0.46	-	-	-	-
FewTURE [6]	ViT-S/16	22 M	68.02±0.88	84.51±0.53	72.96±0.92	86.43±0.67	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73
CPEA [11]	ViT-S/16	22 M	71.97±0.65	87.06±0.38	76.93±0.70	90.12±0.45	77.82±0.66	88.98±0.45	**47.24±0.58**	65.02±0.60
Ours(CASAFA)	ViT-S/16	22 M	**73.19 + 0.63**	**88.23±0.36**	**77.11±0.72**	**90.39±0.41**	**78.46±0.68**	**89.95±0.44**	47.19±0.58	**66.36±0.57**

$$L_{\text{MSE}} = \frac{1}{Q}\sum_{j=1}^{r} \|\hat{z}_j^q - \bar{z}_j^q\|_2^2 \tag{7}$$

III. EXPERIMENTS

In this section, we first explain the purpose of the experiment and the results obtained, then conduct a comparative analysis with other researchers in the field, and finally perform an ablation experiment on the critical components of the experiment.

A. Experimental settings

Datasets. We tested our method on four widely recognized few-shot learning benchmark datasets: miniImageNet[36], tieredImageNet[31], CIFAR-FS[37], and FC100[38]. First, we split the datasets into training, validation, and test sets according to standard experimental practices. The class labels of these datasets are non-overlapping. That is, the categories observed in the training phase will not appear in the validation and test phases. This approach ensures that the knowledge learned by the model in the training phase can be effectively generalized to unseen categories.

Backbone. In this paper, we use the ViT-S/16 model as the core architecture of our research. This model was chosen based on its parameter count, which is comparable to other commonly used networks in small sample image classification. In this configuration, we use a multi-layer perception (MLP) as the projection head, which contains two hidden layers. In the first hidden layer, we applied the GELU activation function, and after the second hidden layer, we used LayerNorm for normalization. We then introduced an innovative feature fusion strategy, using the query vector and the critical vector to calculate the attention score. We normalized it through the softmax function and the scaling factor to obtain the attention probability. Through the calculated attention probability, we perform a weighted summation on the value vector to reconstruct the feature. Through the fusion mechanism, we have improved the model's generalization ability for samples.

Experimental details. We adopted the strategy proposed in the literature [12], strictly set parameters, pre-trained the ViT-S/16 model, and used 4 A100 40G GPUs to pre-train ViT-S/16. The total number of training times was 1600 and the batch size was 512. During the training process, we only used the trained part of the dataset to avoid overfitting the test data in the pre-training stage. After that, we trained an inverse ResNet as a decoder. The purpose of the decoder is to convert the high-dimensional feature values learned by the model back to the pixel space of the original image, with an output size of 3 × 84 × 84. At this stage, we used the Adam optimizer with an initial learning rate of 0.01. We set the batch size to 200 and the total number of training times to 1000.

Evaluation protocol. In both 5-way 1-shot and 5-way 5-shot image classification tasks, we evaluate the final performance of the model by randomly sampling 1000 samples from each test category and calculating their average accuracy, where there are 15 query images for each category.

B. Compariison result

Tables I show the performance comparison of different methods on four different benchmark datasets, including miniImageNet, tieredImageNet, CIFAR-FS, and FC100, using two settings: 5-way 1-shot and 5-way 5-shot. The experimental results show that our CAFASA method outperforms existing methods.

TABLE II
IMPACT OF PROJECTION HEAD ON SMALL SAMPLE CLASSIFICATION

Projection head			miniImageNet		tieredImageNet		CIFAR-FS		FC100	
a	b	c	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
		✓	72.88±0.62	87.91±0.35	76.95±0.69	90.17±0.39	77.69±0.65	89.89±0.42	46.65±0.55	65.77±0.53
✓		✓	73.06±0.63	87.99±0.37	77.02±0.71	**90.68±0.41**	77.43±0.64	89.95±0.42	46.77±0.58	65.83±0.56
	✓	✓	72.95±0.62	87.96±0.36	77.04±0.71	90.24±0.41	77.80±0.67	89.93±0.43	46.65±0.57	65.84±0.56
✓	✓	✓	**73.19±0.63**	**88.23±0.36**	**77.11±0.72**	90.39±0.41	**78.46±0.68**	**89.95±0.44**	**47.19±0.58**	**66.36±0.57**

TABLE III
DIFFERENT INPUTS

	miniImageNet		tieredImageNet		CIFAR-FS		FC100	
$\mathbf{x}$	73.01±0.63	87.99±0.37	77.03±0.72	90.11±0.38	78.33±0.67	89.86±0.42	46.98±0.56	66.21±0.57
$\lvert\mathbf{x}\rvert$	73.17±0.68	**88.33±0.36**	77.11±0.68	**90.45±0.42**	78.37±0.67	89.93±0.44	**47.39±0.61**	66.24±0.58
$\mathbf{x^2}$	**73.19±0.63**	88.23±0.36	**77.13±0.72**	90.39±0.41	**78.47±0.68**	**89.95±0.44**	47.13±0.58	**66.36±0.57**

C. Ablation study

Ablation experiments were performed on the critical components of the method proposed in this article, and experiments were conducted on CIFAR-FS, miniImageNet, tieredImageNet, and FC-100.

Projection head: The projection head increases the model's ability to express features by mapping the embedding vector to a higher-dimensional space, thereby improving classification accuracy in small-sample learning scenarios, as shown in TABLE II.

Different inputs: As shown in TABLE III, the results vary with other inputs to fc(2). Both squaring and taking the absolute value of the inputs can improve performance. Squaring increases intra-class similarity and decreases inter-class similarity, thereby achieving better performance.

D. Conclusion

We have introduced a new method called CAFASA, which enhances the model's efficiency in identifying critical features for image classification tasks through self-supervised learning and feature fusion. Experiments have shown that CAFASA performs exceptionally well across multiple datasets.

REFERENCES

[1] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," *Technical Report, University of Toronto*, 2009.

[2] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[3] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[4] J. Lai, S. Yang, W. Liu, *et al.*, "Tsf: Transformer-based semantic filter for few-shot learning," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–19.

[5] M. Ye, X. Lin, G. Burachas, A. Divakaran, and Y. Yao, "Hybrid consistency training with prototype adaptation for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2726–2735.

[6] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3582–3595, 2022.

[7] C. Zhang, G. Lin, and C. Shen, "Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213.

[8] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: Spatially-aware few-shot transfer," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 981–21 993, 2020.

[9] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2019, pp. 8460–8469.

[10] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in neural information processing systems*, vol. 32, 2019.

[11] F. Hao, F. He, L. Liu, F. Wu, D. Tao, and J. Cheng, "Class-aware patch embedding adaptation for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, IEEE/CVF, 2023, pp. 18 905–18 915.

[12] J. Zhou, C. Wei, H. Wang, *et al.*, "Ibot: Image bert pre-training with online tokenizer," *arXiv preprint arXiv:2111.07832*, 2021.

[13] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[14] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[15] Y. Lifchitz, Y. Avrithis, and S. Picard, "Local propagation for few-shot learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 10 457–10 464.

[16] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8642–8649.

[17] A. Afrasiyabi, H. Larochelle, J.-F. Lalonde, and C. Gagné, "Matching feature sets for few-shot image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9014–9024.

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[19] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8808–8817.

[20] M. Zhang, J. Zhang, Z. Lu, T. Xiang, M. Ding, and S. Huang, "Iept: Instance-level and episode-level pretext tasks for few-shot learning," in *International Conference on Learning Representations*, 2021.

[21] N. Fei, Z. Lu, T. Xiang, and S. Huang, "Melr: Meta-learning via modeling episode-level relationships for few-shot learning," in *International Conference on Learning Representations*, 2021.

[22] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8012–8021.

[23] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep brownian distance covariance for few-shot classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7972–7981.

[24] C. Xu, Y. Fu, C. Liu, *et al.*, "Learning dynamic alignment via meta-filter for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5182–5191.

[25] C. Liu, Y. Fu, C. Xu, *et al.*, "Learning a few-shot embedding model with contrastive learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 8635–8643.

[26] Z. Zhou, X. Qiu, J. Xie, J. Wu, and C. Zhang, "Binocular mutual learning for improving few-shot classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8402–8411.

[27] J. Zhao, Y. Yang, X. Lin, J. Yang, and L. He, "Looking wider for better adaptive representation in few-shot learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 10 981–10 989.

[28] C. Zhang, H. Ding, G. Lin, R. Li, C. Wang, and C. Shen, "Meta navigator: Search for a good adaptation policy for few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9435–9444.

[29] J. Ma, H. Xie, G. Han, S.-F. Chang, A. Galstyan, and W. Abd-Almageed, "Partner-assisted learning for few-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 573–10 582.

[30] X. Luo, L. Wei, L. Wen, *et al.*, "Rectifying the shortcut learning of background for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 073–13 085, 2021.

[31] A. A. Rusu, D. Rao, J. Sygnowski, *et al.*, "Meta-learning with latent embedding optimization," *arXiv preprint arXiv:1807.05960*, 2018.

[32] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8059–8068.

[33] Z. Chen, J. Ge, H. Zhan, S. Huang, and D. Wang, "Pareto self-supervised training for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 663–13 672.

[34] X. Zhang, D. Meng, H. Gouk, and T. M. Hospedales, "Shallow bayesian meta learning for real-world few-shot recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 651–660.

[35] G. Qi, H. Yu, Z. Lu, and S. Li, "Transductive few-shot classification on the oblique manifold," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8412–8422.

[36] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[37] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," *arXiv preprint arXiv:1805.08136*, 2018.

[38] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *Advances in neural information processing systems*, vol. 31, 2018.