NOISE SUPPRESSION FOR IMPROVED FEW-SHOT LEARNING

*Zhikui Chen** *Tiandong Ji* *Suhua Zhang* *Fangming Zhong**

School of Software Technology, Dalian University of Technology, Dalian, Liaoning 116620, China
{zkchen, fmzhong}@dlut.edu.cn

ABSTRACT

Few-shot learning (FSL) aims to generalize from few labeled samples. Recently, metric-based methods have achieved surprising classification performance on many FSL benchmarks. However, those methods ignore the impact of noise, making the few-shot learning still tricky. In this work, we identify that noise suppression is important to improve the performance of FSL algorithms. Hence, we proposed a novel attention-based contrastive learning model with discrete cosine transform input (ACL-DCT), which can suppress the noise in input images, image labels, and learned features, respectively. ACL-DCT takes the transformed frequency domain representations by DCT as input and removes the high-frequency part to suppress the input noise. Besides, an attention-based alignment of the feature maps and a supervised contrastive loss are used to mitigate the feature and label noise. We evaluate our ACL-DCT by comparing previous methods on two widely used datasets for few-shot classification (i.e., miniImageNet and CUB). The results indicate that our proposed method outperforms the state-of-the-art methods.

***Index Terms*—** Few-shot learning, image classification, noise suppression, contrastive learning

1. INTRODUCTION

Deep learning has achieved encouraging breakthroughs in various recognition tasks. Its success highly relies on sufficient and fully annotated data. However, the huge data annotation is expensive and time-consuming. Moreover, in some fields such as biomedicine and land use, it is not easy to acquire enough labeled samples. Hence, traditional deep learning methods may suffer from the data scarcity problem. Inspired by humans' ability to acquire new concepts with only a few labeled samples, few-shot learning (FSL) has recently drawn considerable research interest.

Several few-shot learning methods have been proposed, which can be briefly organized into two categories, i.e., optimization-based [1] and metric-based [2, 3, 4, 5, 6] methods. Optimization-based approaches such as MAML [1] aim to learn a good initialization and then finetune the parameters on new samples a few gradient steps for effective generalization. Metric-based approaches, on the other hand, attempt to make predictions by comparing distances in a learned embedding space. Here, we mainly focus on the metric-based methods. These methods train a feature extractor to map images to an embedding space, where the similarities between images are calculated. Given a query sample and few support samples, they are first mapped to embeddings. Then, the class label of the support samples which are most similar to the query sample is selected as the prediction result. Besides the conventional distance metrics, including Euclidean distance [2] and cosine [3, 7], several more sophisticated metrics [8, 4] have also been proposed. As for distance computation, the earlier metric-based methods compute distance on the pooled feature vectors. Recently, a lot of methods [5, 9] prefer to compute directly on the feature map. This is because the feature vectors may lose much spatial information after pooling.

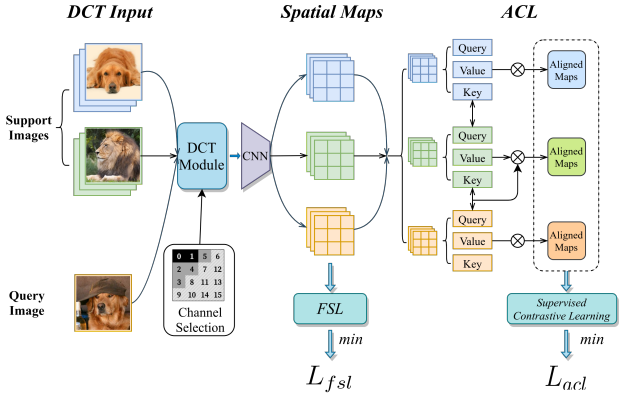

Fig. 1. The framework of our proposed ACL-DCT.

Although many FSL methods have achieved excellent classification performance, we argue that the previous methods ignore the impact of noise. As we know, noise widely exists, and the noise problem is one of the most crucial machine learning problems. In traditional deep learning, models can automatically ignore noise using huge and diverse data. But in FSL, due to the lack of data, it is necessary to introduce some prior knowledge (or inductive bias) to address the noise problems. This requires us to design delicate suppression modules for different noise problems.

In this paper, we summarize three kinds of noise as follows. **(1) Input noise** is a common issue that declines the model's generalizing ability. Several techniques have been used to mitigate the noise in input data. For example, the high-frequency noise channels of the input data are removed in [10] to improve the classification performance. In addition, data augmentation also enables the FSL model to suppress input noise by randomly adding noise while keeping labels the same. **(2) Feature noise** means there exist some noisy features which are irrelevant to the current task. Lacking sufficient guidance from training data, the feature noise problem is particularly severe in FSL. To this end, some recent works [3, 11, 12, 5] have introduced attention mechanisms into FSL, allowing models to focus more on valuable features. The attention mechanism provides an inductive bias to instruct few-shot learning models to suppress the feature noise and therefore leads to better performance. **(3) Label noise** refers to the interference caused by inappropriate labels. Although an

*Corresponding authors. This work is supported by the National Natural Science Foundation of China (62076047, 62006035), the Dalian Science and Technology Innovation Foundation (2021JJ12SN44), and the Fundamental Research Funds for the Central Universities (DUT20LAB136).

image may contain multiple possible labels, most datasets only provide a single label, resulting in label noise. Several works [13, 14] discover that deep neural networks are relatively not robust to label noise due to the utilization of one-hot label and cross-entropy loss. There are many methods used to reduce label noise. For example, Ghosh and Lan [15] identified that pre-training using contrastive learning effectively mitigates the label noise.

To address the noise problems mentioned above, a novel attention-based contrastive learning model with DCT input (ACL-DCT) for FSL is proposed in this work. As shown in Fig. 1, ACL-DCT mainly includes two components: **(1) DCT module** is proposed to reduce the noise in input data. Specifically, a discrete cosine transform is employed to extract the representations of input images from the frequency domain. Then the majority of the high-frequency signals, which are regarded as noise, are filtered out. A recent work [16] shares a similar idea, but it integrates representations from both the spatial and frequency domains. Different from [16], we focus on noise suppression and train the model using only filtered frequency-domain features. As a result, our model is also more computationally efficient. **(2) Attention-based contrastive learning (ACL)** is proposed to mitigate both feature noise and label noise. Firstly, we apply the attention-based alignment between feature maps to suppress the feature noise. This encourages the model to obtain more relevant features to improve classification accuracy. Secondly, in order to mitigate label noise, we adopt supervised contrastive learning in FSL. Contrastive learning has achieved tremendous success in self-supervised learning for computer vision [17]. It does not rely on labels and focuses on instance-level discrimination rather than class-level discrimination. This makes the learned features more semantically rich, thus alleviating the label noise problem. Many previous FSL works [5, 18] have adopted contrastive pre-training using pooled embeddings to improve the performance. Different from them, we compute the contrastive loss directly on the unpooled spatial feature maps to keep spatial information. Moreover, we also align the feature maps based on the attention mechanism to suppress the feature noise. Combining the two components, our proposed ACL-DCT can comprehensively suppress the noise in input, labels, and learned features.

To summarize, our contributions are as follows: (i) To the best of our knowledge, this is the first work that comprehensively introduces the impact of noise in few-shot learning. We argue that suppressing the noise in input images, learned features, and image labels can lead to superior few-shot classification performance. (ii) We propose a new attention-based contrastive learning model with DCT input. ACL-DCT employs a discrete cosine transform to suppress the noise in input data. Then, the noise in learned features and data labels is mitigated by attention-based feature alignment and supervised contrastive learning, respectively. (iii) Extensive experiments are conducted on two widely used datasets, and the results demonstrate that our proposed method is effective in suppressing noise, which further improves the classification performance.

2. PROPOSED APPROACH

2.1. Preliminaries

Here, we briefly describe the formulation of few-shot classification. Given a base set D_b and a novel set D_n, where D_b and D_n are disjoint. The goal is to train a good classifier for novel classes using the base set. Usually, an episodic training mechanism is adopted. In each episodic stage, a support set $\mathcal{S}$ and a query set $\mathcal{Q}$ are randomly sampled. Specifically, the support set contains n classes, where each class contains k samples, also known as the n-way k-shot setting. Formally, the support set is defined as $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{n\times k}$, where x_i^s is the i-th image in D_b (or D_n), y_i^s is the label of x_i^s. Similarly, the query set used for testing can be defined as $\mathcal{Q} = \{(x_i^q, y_i^q)\}_{i=1}^{|\mathcal{Q}|}$. The FSL model is then trained using $\mathcal{S}$ and $\mathcal{Q}$ sampled from D_b, and tested on the $\mathcal{S}$ and $\mathcal{Q}$ sampled from D_n.

2.2. Overview of ACL-DCT

As shown in Fig. 1, our proposed ACL-DCT mainly contains two components. The support and query images are first passed through the DCT module for input noise reduction before being fed into a convolutional neural network (CNN) such as ResNet-12 [19, 20]. The last pooling layer from the CNN is removed so that the spatial feature maps are obtained. Then, the maps are utilized to compute the FSL loss via map reconstruction. Besides, the maps are aligned using the attention mechanism, and a supervised contrastive loss is computed after that to suppress both feature noise and label noise.

2.3. FSL with DCT Module

The proposed DCT module uses DCT to extract the frequency domain representations of the input, from which the input noise is removed. Given an input image, some common augmentation methods such as resizing, cropping, and color jittering are applied first. Following [10], the image is converted to YCbCr color space and then transformed to the frequency domain using DCT. For a specified filter size, each channel of the YCbCr image is patched and then computed as two-dimensional DCT coefficient matrices.

0	**1**	**5**	**6**	14	15	27	28
2	**4**	**7**	**13**	16	26	29	42
3	**8**	**12**	17	25	30	41	43
9	**11**	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Fig. 2. The DCT frequency channels.

Take the 8×8 filter as an example. The input YCbCr image $I \in \mathbb{R}^{H_i \times W_i \times 3}$ is first divided into a total of $H_i/8 \times W_i/8$ patches. For each patch, an 8×8 coefficient matrix is obtained using Eq. (1). As shown in Fig. 2, the serial numbers marked in zigzag order indicate the components from low frequency to high frequency. Usually, the high-frequency part is regarded as noise. Hence, we can keep only the low-frequency part to reduce noise. Since the human vision system (HVS) is more sensitive to luminance (Y) than color (Cr and Cb), we use different selection strategies for these two types of channels. As in Fig. 2, we keep the colored part for the Y channel, while for the Cr and Cb channels, the dark-colored part is kept. Finally, a total of 24 DCT coefficients are kept. Subsequently, the coefficients of the same frequency are combined into one channel to obtain a three-dimensional tensor $D \in \mathbb{R}^{H_i/8 \times W_i/8 \times 24}$. It is then used as input to the CNN for feature extraction.

$$D_{i,j} = \frac{C_i C_j}{4} \sum_{x,y \in [0,7]} I_{x,y} \cos\frac{(2x+1)i\pi}{16} \cos\frac{(2y+1)j\pi}{16} \qquad (1)$$
$$C_u = \begin{cases} \frac{1}{\sqrt{2}} & u = 0 \\ 1 & \text{otherwise} \end{cases}$$

To preserve the spatial information, we remove the last pooling layer of CNN. Thus, the output of the feature extractor is a feature map $F \in \mathbb{R}^{H_f \times W_f \times d}$, where H_f and W_f are the height and width of F, and d is the number of channels. To directly use the spatial maps, we employ the feature reconstruction method [21, 9] to compute the FSL loss. In an n-way k-shot episode, let $Y \in \mathbb{R}^{H_f W_f \times d}$ denote the flattened feature map of the query image. For a class c, features of all k support images compose $X_c \in \mathbb{R}^{kH_f W_f \times d}$. The feature reconstruction method attempts to find a $W \in \mathbb{R}^{H_f W_f \times kH_f W_f}$ to reconstruct Y by X_c so that $WX_c \approx Y$. This can be formulated as solving the following least-squares problem,

$$\bar{W} = \arg\min_{W} \|Y - WX_c\|^2 + e^{\gamma}\|W\|^2 \tag{2}$$

where $\bar{W}$ is the optimal solution and $e^{\gamma} > 0$ is the learnable penalty term. By deriving, the optimal reconstruction $\bar{Y}$ with respect to $\bar{W}$ can be computed directly according to the closed-form solution $\bar{W}$ of Eq. (2):

$$\bar{Y}_c = \bar{W}X_c = YX_c^T \left(X_c X_c^T + e^{\gamma} I\right)^{-1} X_c \tag{3}$$

The closer $\bar{Y}_c$ is to Y, the more likely the query image belongs to class c. We define the distance between $\bar{Y}_c$ and Y as follows:

$$d\left(Y, \bar{Y}_c\right) = \frac{1}{H_f W_f} \left\|Y - \bar{Y}_c\right\|^2 \tag{4}$$

Following [2, 7], the few-shot classification loss over episodes can be formulated as:

$$\mathcal{L}_{\text{fsl}} = \frac{-1}{|\mathcal{T}_Q|} \sum_{(Y_i, y_i) \in \mathcal{T}_Q} \log \frac{\exp\left(-\tau_1 d\left(Y, \bar{Y}_{y_i}\right)\right)}{\sum_{c' \in C} \exp\left(-\tau_1 d\left(Y, \bar{Y}_{c'}\right)\right)} \tag{5}$$

where $\mathcal{T}_Q$ denotes the feature and label pairs of the query set in an episode, τ_1 is the learnable softmax temperature.

2.4. Attention-based Contrastive Learning

The proposed ACL loss is a supervised contrastive loss [22] adapted to few-shot learning. Unlike traditional unsupervised contrastive learning, ACL takes advantage of the label information in each episode to employ supervised contrastive learning using aligned features.

For an n-way k-shot episode, we assume there are q query samples per class, then we have $N = n \times (k+q)$ training data pairs, each containing a feature $F_i \in \mathbb{R}^{H_f W_f \times d}$ and a label y_i. To solve the feature noise problem, we first apply the attention mechanism [23] to align the features. Concretely, given two features F_m and F_n, we use each to compute a triplet $(\mathcal{Q}, \mathcal{K}, \mathcal{V})$, namely the query, key, and value,

$$(\mathcal{Q}_*, \mathcal{K}_*, \mathcal{V}_*) = (F_* W_Q, F_* W_K, F_* W_V) \tag{6}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d'}$ are the parameters of three linear mappings. The similarity between $\mathcal{Q}_n$ and $\mathcal{K}_m$ is calculated as *attention*, which is used to compute the weighted sum of $\mathcal{V}_m$. Finally, the aligned $\hat{\mathcal{V}}_m$ can be denoted as,

$$\hat{\mathcal{V}}_m = \text{softmax}\left(\frac{\mathcal{Q}_n \mathcal{K}_m^{\top}}{\sqrt{d'}}\right) \mathcal{V}_m \tag{7}$$

Let $\mathcal{V}_*^i$ depict the vector of i-th spatial position. After ℓ_2 normalization, the distance between two features is then defined as follows:

$$d\left(F_m, F_n\right) = -\frac{1}{H_f W_f} \sum_{i=1}^{H_f W_f} \left(\mathcal{V}_n^i\right)^{\top} \hat{\mathcal{V}}_m^i \tag{8}$$

Using the similarity function shown in Eq. (8), we can then compute the contrastive loss. First, in each episode, we use the same pipeline proposed in [17] to apply augmentation to obtain an additional set of images. Thus, we have $2N = 2n(k+q)$ data pairs in total. For image x_i, all support and query images and their augmentations from the same class are considered as positive images, and all other images and their augmentations are considered as negative images. The contrastive loss for x_i is then defined as,

$$\mathcal{L}_i = \frac{-1}{2P_i - 1} \sum_{r=1}^{2N} \mathbf{1}_{\substack{[y_r = y_i] \\ r \neq i}} \log \frac{\exp\left(-\tau_2 d\left(F_i, F_r\right)\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(-\tau_2 d\left(F_i, F_k\right)\right)} \tag{9}$$

where $P_i = k + q$ is the number of samples in an episode that have the same label as x_i, F_* is the feature map of x_*, τ_2 is a learnable scalar temperature, and $\mathbf{1}_{[\mathbf{C}]} \in \{0, 1\}$ is an indicator function that equals one if $\mathbf{C}$ is true. Compared to cross-entropy loss, Eq. (9) can model certain discrimination even between samples with the same label. Thanks to it, the model is less sensitive to label noise. Finally, considering all samples, we have the ACL loss as follows:

$$\mathcal{L}_{\text{acl}} = \sum_{i=1}^{2N} \mathcal{L}_i \tag{10}$$

2.5. Learning Objective for ACL-DCT

Combining Eq. (5) and Eq. (10), the learning objective for the proposed attention-based contrastive learning model with DCT input is stated as,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fsl}} + \lambda \mathcal{L}_{\text{acl}} \tag{11}$$

where λ is a hyperparameter used to balance the impact of two losses. During the training phase, the model is trained using episodes sampled from D_b. When the model is well trained, we evaluate our model using episodes sampled from D_n.

Furthermore, previous work [7] has shown that a non-episodic pre-training can significantly improve the performance of few-shot learning models. Therefore, we also try to pre-train during the training phase. Suppose $\mathcal{C}_n$ contains all the classes in $\mathcal{D}_n$, this can be simply done by replacing X_c in Eq. (3) with a learnable matrix $P_c \in \mathbb{R}^{H_f W_f \times d}$, where $c \in \mathcal{C}_n$. After pre-training, all learnable matrices are discarded and the model's weights are retained for further episodic training.

3. EXPERIMENTS

3.1. Datasets

We validate our model using two benchmark datasets, i.e., miniImageNet [24] and CUB [25]. The miniImageNet is a subset of ImageNet, which includes 100 categories, each containing 600 images. We follow the same setting as in [24] that the categories are divided into 64 base classes for training, 16 and 24 classes for validation and testing. The CUB consists of 11,788 images in 200 categories. We follow the splits from [26], where 100 classes are selected for training, 50 for validation and 50 for evaluation. All images are resized to 84×84 resolution before being fed into models.

3.2. Training Details

We train all models using stochastic gradient descent (SGD) with Nesterov momentum 0.9 and weight decay 5e-4. λ is set to 1 in our experiments. γ, τ_1, and τ_2 are initialized to 0, 1, and 10. ResNet-12 is used as the backbone. Note that for models with a DCT module, an 8×8 DCT filter is utilized. And all images are upsampled by 4. We then remove the first pooling layer in the backbone so that the output feature maps are consistent with those using RGB input.

For the miniImageNet dataset, we pre-train 400 epochs with a batch size of 64. The learning rate is 0.1 and reduced by a factor of 10 at epochs 200 and 300. The best model selected by the validation set is used for subsequent episodic finetuning. We finetune a total of 130 epochs following the 5-way 5-shot setting. The learning rate is 1e-3 and reduced by a factor of 10 at epochs 70 and 120. As for the CUB dataset, we conduct episodic training of 800 epochs following the 5-way 5-shot setting. The learning rate is set to 0.1 and divided by 10 at epochs 350 and 700.

Table 1. The mean accuracies (%) with a 95% confidence interval on the miniImageNet. The top two results are bolded and underlined.

Method	Backbone	5-way 1-shot	5-way 5-shot
MatchingNets [24]	Conv4	43.56±0.84	55.31±0.73
MAML [1]	Conv4	48.70±1.75	63.11±0.92
ProtoNet [2]	Conv4	49.42±0.78	68.20±0.72
TADAM [27]	ResNet-12	58.50±0.30	76.70±0.30
MetaOptNet [20]	ResNet-12	62.64±0.61	78.63±0.46
CAN [11]	ResNet-12	63.85±0.48	79.44±0.34
Meta-Baseline [7]	ResNet-12	63.17±0.23	79.26±0.17
FEAT [12]	ResNet-12	66.78±0.20	82.05±0.14
DeepEMD [4]	ResNet-12	65.91±0.82	82.41±0.56
Neg-Cosine [28]	ResNet-12	63.85±0.81	81.57±0.56
RFS-distill [29]	ResNet-12	64.82±0.60	82.14±0.43
DMF [6]	ResNet-12	<u>67.76±0.46</u>	<u>82.71±0.31</u>
ACL-DCT (ours)	ResNet-12	**68.74±0.20**	**84.70±0.12**

Table 2. The mean accuracies (%) with a 95% confidence interval on the CUB. The top two results are bolded and underlined.

Method	Backbone	5-way 1-shot	5-way 5-shot
Baseline [26]	ResNet-18	65.51±0.87	82.85±0.55
Baseline++ [26]	ResNet-18	67.02±0.90	83.58±0.54
MatchingNets [26, 24]	ResNet-18	73.49±0.89	84.45±0.58
ProtoNet [26, 2]	ResNet-18	72.99±0.88	86.64±0.51
MAML [26, 1]	ResNet-18	68.42±1.07	83.47±0.62
RelationNet [26, 30]	ResNet-18	68.58±0.94	84.05±0.56
S2M2 [31]	ResNet-18	71.43±0.28	85.55±0.52
Neg-Cosine [28]	ResNet-18	72.66±0.85	89.40±0.43
CTX [5]	ResNet-12	<u>79.34±0.21</u>	<u>91.42±0.11</u>
ACL-DCT (ours)	ResNet-12	**84.05±0.19**	**93.00±0.10**

Table 3. Performance comparison in the cross-domain setting. The top two results are bolded and underlined.

Method	Backbone	5-way 1-shot	5-way 5-shot
MAML [1, 26]	ResNet-18	-	51.34±0.72
ProtoNet [2, 26]	ResNet-18	-	62.02±0.70
Baseline [26]	ResNet-18	-	<u>65.57±0.70</u>
MetaOptNet [20, 31]	ResNet-12	<u>44.79±0.75</u>	64.98±0.68
RelationNet [32]	ResNet-10	44.07±0.77	59.46±0.71
ACL-DCT (ours)	ResNet-12	**52.14±0.21**	**72.90±0.17**

3.3. Comparison with State-of-the-art

We compare our approach against a number of current FSL methods. Previous studies usually evaluate their models on 600 sampled tasks which may introduce high-variances. Thus, we follow a more credible setting to test our models using 10,000 randomly sampled tasks. The results of miniImageNet and CUB are reported in Table 1 and Table 2. From the tables, we can see that our proposed ACL-DCT model achieves competitive results on both datasets. For results on the miniImageNet, compared with the second-best DMF [6], the performance gains of our ACL-DCT are 1% and 2% on 1-shot and 5-shot tasks, respectively. For CUB, our method achieves performance gains by 4.7% and 1.6% compared to the second-best CTX [5]. Note that our method does not follow any additional settings such as transductive settings or settings with extra training data. The results corroborate that noise suppression is essential to improve the performance of FSL models.

In addition, we evaluate ACL-DCT under the cross-domain setting as in [26]. This is a more challenging setting where the model is trained on the miniImageNet and then tested directly on the CUB dataset. The results are listed in Table 3, and it can be seen that our proposed method exceeds previous methods by a large margin. The results further demonstrate that through noise suppression, our method has better generalization ability.

3.4. Ablation Study

Moreover, we conduct an ablation study to investigate the effectiveness of various components of our ACL-DCT on miniImageNet. For the ACL module, we also test the performance of the model without attention-based alignment. The results are reported in Table 4, where the baseline model uses only the FSL loss in Eq. (5). Compared to the baseline, the DCT and ACL modules both boost the performance. Specifically, each component can bring a performance gain, which indicates their importance. Combining ACL and DCT, the 1-shot accuracy and 5-shot accuracy finally improve by 3.4% and 2.5%, respectively.

Table 4. Ablation study of the components in ACL-DCT. 'AA' denotes attention-based alignment and 'CL' is contrastive learning.

DCT	ACL		*mini*-ImageNet, 5-way	
	AA	CL	1-shot	5-shot
			65.31 ± 0.20	82.22 ± 0.13
✓			67.40 ± 0.19	83.45 ± 0.13
		✓	67.49 ± 0.19	83.35 ± 0.13
	✓	✓	68.21 ± 0.19	84.36 ± 0.12
✓	✓	✓	**68.74 ± 0.20**	**84.70 ± 0.12**

4. CONCLUSION

In this paper, we have presented a novel attention-based contrastive learning model with DCT input (ACL-DCT) to address the noise problem in few-shot learning. Our approach can suppress noise from three different aspects. Extensive experiments have shown that our ACL-DCT achieves the new state-of-the-art.

5. REFERENCES

[1] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks,"

in *ICML*, 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135.

[2] Jake Snell, Kevin Swersky, and Richard S. Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017, pp. 4077–4087.

[3] Spyros Gidaris and Nikos Komodakis, "Dynamic few-shot visual learning without forgetting," in *CVPR*, 2018, pp. 4367–4375.

[4] 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, pp. 12200–12210.

[5] Carl Doersch, Ankush Gupta, and Andrew Zisserman, "Crosstransformers: spatially-aware few-shot transfer," in *NeurIPS*, 2020.

[6] Chengming Xu, Chen Liu, Li Zhang, Chengjie Wang, Jilin Li, Feiyue Huang, Xiangyang Xue, and Yanwei Fu, "Learning dynamic alignment via meta-filter for few-shot learning," in *CVPR*, 2021.

[7] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell, "A new meta-baseline for few-shot learning," *arXiv preprint arXiv:2003.04390*, 2020.

[8] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi, "Adaptive subspaces for few-shot learning," in *CVPR*, 2020, pp. 4135–4144.

[9] Davis Wertheimer, Luming Tang, and Bharath Hariharan, "Few-shot classification with feature map reconstruction networks," in *CVPR*, June 2021, pp. 8012–8021.

[10] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren, "Learning in the frequency domain," in *CVPR*, 2020, pp. 1740–1749.

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen, "Cross attention network for few-shot classification," in *NeurIPS*, 2019, pp. 4005–4016.

[12] 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, pp. 8805–8814.

[13] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus, "Training convolutional networks with noisy labels," in *ICLR*, 2015.

[14] Zhilu Zhang and Mert R Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *NeurIPS*, 2018.

[15] Aritra Ghosh and Andrew Lan, "Contrastive learning improves model robustness under label noise," in *CVPR*, 2021, pp. 2703–2708.

[16] Xiangyu Chen and Guanghui Wang, "Few-shot learning by integrating spatial and frequency representation," *arXiv preprint arXiv:2105.05348*, 2021.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[18] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue, "Self-supervised learning for few-shot image classification," in *ICASSP*. IEEE, 2021, pp. 1745–1749.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, 2019, pp. 10657–10665.

[21] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi, "Meta-learning with differentiable closed-form solvers," in *ICLR*, 2018.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *NeurIPS*, vol. 33, 2020.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[24] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016, pp. 3630–3638.

[25] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

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

[27] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste, "Tadam: task dependent adaptive metric for improved few-shot learning," in *NeurIPS*, 2018, pp. 719–729.

[28] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu, "Negative margin matters: Understanding margin in few-shot classification," in *ECCV*, 2020.

[29] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola, "Rethinking few-shot image classification: a good embedding is all you need?," in *ECCV*, 2020.

[30] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *CVPR*, 2018, pp. 1199–1208.

[31] 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 *WACV*, 2020, pp. 2218–2227.

[32] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang, "Cross-domain few-shot classification via learned feature-wise transformation," in *ICLR*, 2020.