Pattern Recognition 142 (2023) 109702

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Global- and local-aware feature augmentation with semantic orthogonality for few-shot image classification

Boyao Shi [a], Wenbin Li [a,*], Jing Huo [a], Pengfei Zhu [b], Lei Wang [c], Yang Gao [a]

[a] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[b] College of Intelligence and Computing, Tianjin University, Tianjin 300354, China
[c] School of Computing and Information Technology, University of Wollongong, Wollongong 2522, Australia

ARTICLE INFO

Article history:
Received 31 January 2023
Revised 19 April 2023
Accepted 18 May 2023
Available online 23 May 2023

Keywords:
Few-shot image classification
Transfer learning
Feature augmentation
Semantic orthogonal learning

ABSTRACT

As for few-shot image classification, recently, some works revisit the standard transfer learning paradigm, i.e., pre-training and fine-tuning, and have achieved some success. However, we find that this kind of methods heavily relies on a naive image-level data augmentation (e.g., cropping and flipping) at the fine-tuning stage, which will easily suffer from the overfitting problem because of the limited-data regime. To tackle this issue, in this paper, we attempt to perform a novel feature-level semantic augmentation at the fine-tuning stage and propose a *Global- and Local-aware Feature Augmentation method (GLFA)* from both the channel- and spatial-wise perspectives. In addition, at the pre-training stage, we further propose a *Semantic Orthogonal Learning Framework (SOLF)* to make the learned feature channels more independently, orthogonal and diverse. Extensive experiments demonstrate that the proposed method can obtain significant performance improvements over the state of the arts. Code is available at https://github.com/onlyyao/GLFA-SOLF.

1. Introduction

The huge and rich labeled data has tremendously promoted the development of deep learning [1–4], such as ResNet [5], DenseNet [6], MAE (Masked Autoencoders) [7], ViT (Vision Transformer) [8] and CLIP (Contrastive Language-Image Pre-training) [9]. However, in many specific scenarios, the annotation of data is costly and only limited labeled samples are accessible. To overcome this challenge, few-shot learning (FSL) aiming to learn from the limited-data regime has attracted a wide range of interests and attention from the community. Also, a variety of advanced FSL methods has been proposed, including metric-based methods [10–12], optimization-based methods [13,14], fine-tuning-based methods [15,16] and other interdisciplinary fields [17,18], such as Graph Neural Network (GNN). The metric-based methods adopt an episodic-training paradigm to learn a good embedding feature space from a metric-learning perspective. The optimization-based methods employ a meta-learning paradigm to learn how to quickly adapt to new tasks with a good initialization. Both of these two kinds of methods belong to a task (episode) learning paradigm by learning from thousands of mimetic few-shot tasks built on a disjoint auxiliary set.

Different from the above metric- and optimization-based methods, some latest works [15,16] have tried to revisit the traditional transfer learning paradigm, i.e., pre-training and fine-tuning in FSL, showing much promising results. That is to say, at the training stage, a single large-way classification task is pre-trained on the seen base classes to obtain a good feature extractor. At the meta-test stage, fixing the pre-trained feature extractor, only a new linear classifier is fine-tuned for the unseen novel classes with the few labeled training examples. However, we notice that these methods still suffer from the underlying issues of FSL: (1) During the pre-training stage, the feature extractor may overfit to the base seen classes, leading to reduce its generalization ability on novel classes. Overfitting occurs when the feature extractor becomes too specialized in detecting specific features of the base classes, which can make it less adaptable to detecting new features of novel classes. (2) During the fine-tuning stage, the scarcity of support images for unseen novel classes can result in overfitting of the new linear classifier. Overfitting occurs when the model only memorizes the few available samples for novel classes, rather than learning generalizable feature. Morever, data augmentation techniques such as cropping and flipping can help increase the number of training samples, but may not be enough to prevent overfitting on the limited data.

* Corresponding author.
E-mail addresses: boyao@smail.nju.edu.cn (B. Shi), liwenbin@nju.edu.cn (W. Li), huojing@nju.edu.cn (J. Huo), zhupengfei@tju.edu.cn (P. Zhu), leiw@uow.edu.au (L. Wang), gaoy@nju.edu.cn (Y. Gao).

https://doi.org/10.1016/j.patcog.2023.109702
0031-3203/

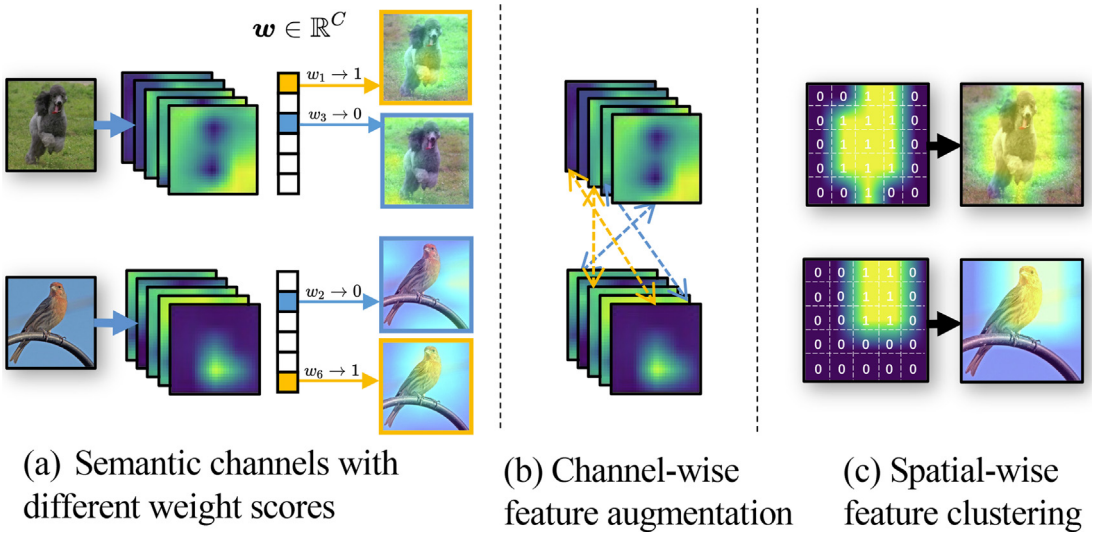


Fig. 1. Our semantic augmentation in a feature level. (a) Feature channels with the higher weight scores contain more meaningful class-specific semantics. (b) Channel-wise feature augmentation can be interpolated between different feature maps, by guaranteeing the orthogonality cross channels. (c) Spatial-wise feature clustering could be performed to cluster the similar semantics together.

In this paper, to tackle the above first issue, we first introduce an additional self-supervised learning (SSL) task (e.g., rotation prediction) at the pre-training stage like the latest works [19] to alleviate the overfitting problem on the base classes, which will be regarded as a strong baseline. In addition, as shown in Fig. 1(a), for a well trained convolutional neural network (CNN), different channels of the feature map generally can generate responses to different semantics. However, the semantic relationships between different feature channels are normally not independently and identically distributed (*non-i.i.d.*). Although these learned *non-i.i.d.* relationships may be suitable for the seen base classes, they may no longer suit the disjoint unseen novel classes. Therefore, to further reduce the risk of overfitting, we propose a novel *semantic orthogonal learning framework (SOLF)* at the pre-training stage to learn more diverse and discriminative features, by making the feature channels independent and orthogonal.

Moreover, to address the above second issue, we present a new *global- and local-aware feature augmentation method (GLFA)* at the fine-tuning stage from both the channel- and spatial-wise perspectives. To learn an effective classifier for the novel few-shot tasks, the existing methods usually rely on an image-level data augmentation, such as cropping or flipping. However, because the training examples are extremely scarce, this kind of image-level data augmentation can not effectively avoid the overfitting problem. In contrast, as shown in Fig. 1(a), because different feature channels are able to learn different global semantics for an input image (e.g., dog's paw and bird's beak in the yellow box), it will be promising to explore the rich and informative feature channels for a feature-level augmentation. Benefiting from the orthogonality created by SOLF, our *global-aware feature augmentation* is to randomly replace the class-independent (small weights) feature channels with class-specific (large weights) channels in both intra- and inter-class, making the augmented features contain more diverse and discriminative semantics. The process is shown in Fig. 1(b). Furthermore, inspired by the idea of Luo et al. [20] that the background information is harmful for FSL, we propose a new *local-aware background smoothing* method to suppress the background perturbations for FSL. Compared to Luo et al. [20], without any additional learning-based operations, our proposed method can recognize the local foreground and background regions (see Fig. 1(c)) respectively in an unsupervised manner and naturally smooth the background noises.

In summary, the contributions of our paper are as follows:

- We propose a *semantic orthogonal learning framework (SOLF)* to obtain orthogonal and diverse feature channels. This framework aims to learn and generate better features for Few-Shot Learning (FSL) in a purely semantic-aware manner. By using SOLF, we can achieve better performance on FSL tasks by obtaining high-quality features.
- We propose a *global- and local-aware feature augmentation (GLFA) method* to augment features in terms of improving the diversity of the augmented samples and alleviating the overfitting problem. GLFA achieves this by incorporating global and local awareness into the augmentation process, which allows for more effective and diverse feature.
- hrough extensive experimentation on four standard benchmarks, we have demonstrated that our proposed method significantly outperforms baseline methods on both 5-way and large-way settings, without introducing excessive parameters. These results indicate the effectiveness of our approach, and suggest that our approach has practical applications for real-world scenarios.

In the remainder of this paper, we first summarize the related work in Section 2. Next, we review the definition of the few-shot learning problem and introduce a strong baseline in Section 3. In the Section 4, we present the technical details of the proposed method, including Global-aware Feature Interpolation with Semantic Orthogonality and Local-aware Background Smoothing. After that, we introduce the details and results of the experiments and ablation study in Section 5. Finally, our work is concluded in the last section.

2. Related works

In this section, we introduce mainstream methods which are relevant to our work including three categories: (1) metric-based methods [21–25], which focus on representing a task-agnostic embedding that can distinguish novel categories under a distance metrics, (2) optimization-based methods [13,14,26], which target at searching for good parameters that can quickly adapt to novel samples, and (3) fine-tuning based methods [16,27–30], which certify that pre-training on the whole base dataset can bring a huge improvement.

2.1. Metric-based FSL methods

Metric-based FSL methods aim to learn an informative embedding space, in which data from different classes can be distinguishable with simple distance metrics. There is a broad range of methods in this direction. For example, MatchingNet [31] applies a new nearest neighbor method with an embedded feature extractor and combines the advantages of both parametric and non-parametric methods. ProtoNet [32] represents the mean vector of samples as its class prototype in a representation space and uses the nearest neighbor classifier between prototypes and query images to make predictions. IMP[33] infinite mixture prototypes to adaptive different data distributions. BD-CSPN [34] believes that there is a significant bias between the prototype generated by ProtoNet and the true prototype and rectifies it from two aspects: the intra-class bias and the cross-class bias. In addition, instead of using the global feature representations in the feature space, some recent methods have tried to apply local feature representations to FSL. For instance, DN4 [11] does not use image-level feature vectors but uses the rich local descriptors and employs the image-to-class measure to perform the final classification. DC-IMP [35] directly studies local activations and fuses these local activations and features to learn task-specific features. CrossTransformers [36] find coarse spatial correspondence between the query and the support images and then calculate distances between their spatially-corresponding features for the final classification. DeepEMD [22] applies the Earth Mover's Distance as a measurement method to find the optimal matching distance between images. Moreover, DeepBDC [37] measures the discrepancy between the product of the marginals of embedded features and the joint characteristic functions. Relational Embedding Network (RENet) [38] combines both a global classifier and a local classifier to learn relational embedding. This method introduces a cross-correlational attention module to learn the self-correlational representation and transferable structure.

2.2. Optimization-based FSL methods

Optimization-based FSL methods aim to learn a good initialization so that the model could rapidly adapt to unseen novel tasks through a sequence of training episodes. As a representative, model-agnostic meta-learning (MAML) [39] follows a pure meta-training paradigm, employs the second-order gradients, and learns to fast adapt to a new task with a small number of gradient updates. Reptile [40] adopts a simpler way to update the slow weight, without dividing the task into a support set and a query set. Specifically, Reptile only uses first-order derivatives for the mate-learning updates. Almost No Inner Loop (ANIL) [41] further explores the effectiveness of MAML [39], and finds that feature reuse is key for learning. ANIL [41] removes all inner loop updates except the head of the network, significantly improving computing efficiency. Latent embedding optimization (LEO) [42] learns a low-dimensional latent embedding space and performs optimization-based adaptation in this space to obtain a better initialization more effectively. MetaOptNet [43] learns better generalization feature embedding under linear classification. To learn feature embedding effectively, MetaOptNet fellows two properties: the implicit differentiation of optimality conditions for convex problems and the dual formula for optimal problems, which can improve computational and memory efficiency. Category Traversal Module (CTM) [14] extracts feature relevance to each task through the context of support samples and uses inter-class uniqueness and intra-class commonality for better classification. This method could identify discriminatively and learning effectively features. BOIL [44] means learning the body of the model only in the inner loop. This method could solve overfitting and improve robustness to hyperparameters change. iMAMl [45] solve to the inner level optimization instead of inner loop optimizer.

2.3. Fine-tuning-based FSL methods

Fine-tuning-based FSL methods apply pre-training on the base classes as a pre-processing for FSL, which brings great improvements. It has been found that a simple pre-training can be helpful for few-shot learning, even without episodic training. These methods have significant effectiveness and received increasing attention. For example, MTL [46] learns to transfer the pre-trained network weights to new tasks by fine-tuning some parameters at the test stage. Specifically, MTL first fixes the parameters of the pre-trained model and then relearns the scale and shift parameters to fine-tune the network. RFS-simple [16] employs the logistic regression instead of the fully connected layer as new classifiers at the fine-tuning stage. In the training stage, a CNN model (extractor) is trained through the entire training set by a common classifier. Then, the pre-trained fixed extractor combines a learnable linear classifier for each task is used in the stage of meta-testing. The work in Gidaris et al. [27] considers predicting the rotation degrees of images as an auxiliary self-supervision task at the pre-training stage to help the model obtain better generalize when facing novel classes. Meta-Baseline [29] firstly discusses why meta-learning is not as good as fine-tuning-based methods and finds that in the process of meta-training, improving the generalization ability of base classes will lead to the deterioration of the generalization ability of the model to new classes. As for transductive fine-tuning [47], this method uses a large number of meta-training classes to pre-train a model in order to obtain high few-shot accuracy and introduces a soft-max classifier during the fine-tuning stage. Neg-Cosine [48] adopts an appropriate negative margin in standard softmax loss during the pre-training stage which could avoid falsely mapping the same class of new samples to other clusters or peaks in the base class and improve the ability to identify new classes.

Following this paradigm, in this paper, we also propose a two-stage model by pre-training on the base classes via standard cross-entropy loss and fine-tuning on the novel classes with the trained embedding model. The main difference is that at the pre-training stage, we orthogonalize the semantic information cross channels to learn more diverse features and at the meta-testing stage we propose a global- and local-aware feature augmentation from both channel- and spatial-wise perspectives. We have sorted out the mathematical notations in Table 1.

3. Preliminaries

This section has three parts. We first review the definition of the few-shot learning problem and then detail how to pre-training and fine-tuning, and finally introduce a strong baseline with self-supervision.

3.1. Problem formulation

Few-shot image classification aims to generalize to new tasks given only a few labeled training examples. Following the common setting of FSL in the literature [22,31,39], a given dataset is generally divided into $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$, where $\mathcal{D}_{base}$ containing $\mathcal{C}_{base}$ base classes is used for training, $\mathcal{D}_{novel}$ consists of $\mathcal{C}_{novel}$ novel classes for testing, and $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. For the episodic-training mechanism, each N-way K-shot task of $\{\mathcal{T} = \langle \mathcal{S}, \mathcal{Q} \rangle\}$ is randomly sampled from the dataset, where the support set $\mathcal{S} = \{(X_i, y_i)\}_{i=1}^{NK}$ includes N classes with K samples per class and the query set $\mathcal{Q} = \{(Q_i, y_i)\}_{i=1}^{NM}$ contains the same N classes with M samples per class.

Table 1
Definition of mathematical notations.

Mathematical notation	Definition
$\mathcal{D}_{base}$, $\mathcal{D}_{novel}$	Dataset for training and testing
$\mathcal{C}_{base}$, $\mathcal{C}_{novel}$	Number of categories for $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$
$\mathcal{T} = \langle \mathcal{S}, \mathcal{Q} \rangle$	Few-shot tasks, support and query set
f_θ	Feature extractor to extract feature representation
g_ϕ	Block to learn a weight of each channel
C_ω	The classifier at pre-training stage
$\mathcal{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_C\}$	Feature map $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{f}_C \in \mathbb{R}^{1\times HW}$
$\boldsymbol{w}_i$	Weight of each channel, $\boldsymbol{w}_i = g_\phi(avg(f_\theta(X_i))) \in \mathbb{R}^C$
$\mathcal{V} = \{\boldsymbol{v}_i\}$	A set of local descriptors, $\boldsymbol{v}_i \in \mathbb{R}^C|_{i=1}^{HW}$
$\mathcal{Z} = \{\mathcal{Z}^i\}$	A set of foreground and background clusters features
$\boldsymbol{c} = \{\boldsymbol{c}_i\}$	The cluster center, $i = 1, 2$
$\boldsymbol{p} = \{\boldsymbol{p}_i\}$	The weighted cluster representations, $i = 1, 2$
$D_{i,j}$	The cosine similarity between channel pair
$\mathcal{L}^{\mathrm{CE}}$	The cross-entropy loss function
Γ_{ce}	The classification loss
Γ_{ss}	The self-supervised loss
Γ_{os}	The proposed semantic orthogonal loss
Γ_{total}	The total loss to optimize network
λ	A weight parameter to balance label
α	A weight parameter to balance different losses

3.2. Pre-training and fine-tuning

Recent works [16,29] have demonstrated that a good feature embedding is beneficial to the generalization on novel classes. This kind of methods generally adopt a two-stage learning paradigm: pre-training (i.e., meta-training) and fine tuning (i.e., meta-testing). In the pre-training phase, there is a CNN as the feature extractor f_θ and a fully-connected (FC) layer as the classifier C_ω. At this stage, the whole $\mathcal{D}_{base}$ is used to train a $\mathcal{C}_{base}$-class classifier by using the standard cross-entropy loss as below:

$$\Gamma_{ce} = \arg\min_{\theta,\omega} \sum_{i=1}^{|\mathcal{D}_{base}|} \mathcal{L}^{\mathrm{CE}}(C_\omega(f_\theta(X_i)), y_i), \tag{1}$$

where $\mathcal{L}^{CE}$ is the cross-entropy loss function, $|\mathcal{D}_{base}|$ is the number of base class samples and C_ω is classifier.

In the fine-tuning phase, a new classifier will be individually learned for each novel few-shot task $\mathcal{T} = \langle \mathcal{S}, \mathcal{Q} \rangle$ sampled from $\mathcal{D}_{novel}$ and the parameters of the feature extractor are normally fixed. In general, a logistic regression or a FC layer will be taken as the new classifier to obtain the corresponding logits and further calculate the cross entropy loss $\Gamma_{new} = -\sum_{i=1}^{NK} y_i log(p(y_i|X_i))$, where $(X_i, y_i) \in \mathcal{S}$. After that, the learned classifier is used to predict the labels of the samples in $\mathcal{Q}$. In this paper, we will follow this pre-training and fine-tuning paradigm and employ logistic regression as the new classifier for the novel classes.

3.3. A Strong Baseline with self supervision

Recent studies [28,49] have attempted to enhance few-shot learning by introducing self-supervised learning (SSL) to improve the transferability of feature extraction. These studies combine the few-shot classification task with a self-supervised learning task, sharing a feature extraction network. The self-supervised learning task is employed to enhance the feature extraction network's ability, thus improving the effectiveness of the few-shot classification task.

To make a fair comparison with these methods, we follow them and introduce rotation prediction [50] as an auxiliary task at the pre-training stage to make a strong baseline. On the other hand, the SSL-based auxiliary task [51] can also somewhat alleviate the overfitting problem on the base classes. Specifically, given an image X_i, we first rotate it by r degrees to create four copies $\{X_i^r | r \in \mathcal{R}\}$, where $\mathcal{R} = \{0^o, 90^o, 180^o, 270^o\}$. For each rotated image, we use the same feature extractor f_θ to extract its feature representation and then perform a 4-class classification task with an additional rotation-angle prediction classification head R_γ to predict its corresponding angle. To be specific, the self-supervised loss of this rotation task can be defined as:

$$\Gamma_{ss} = \arg\min_{\gamma,\theta} \sum_{i=1}^{|\mathcal{D}_{base}|} \sum_{r\in\mathcal{R}} \mathcal{L}^{\mathrm{CE}}\left(R_\gamma\left(f_\theta\left(X_i^r\right)\right), r\right), \tag{2}$$

where $\mathcal{L}^{CE}$ is the cross-entropy loss function, X_i^r is an image rotated by r degrees with the original images X_i. Therefore, the overall objective function of our strong baseline is $\Gamma_{total} = \Gamma_{ce} + \alpha\Gamma_{ss}$, where α is a balance weight parameter. This method can improve the generalization ability to adapt to new classes with few training data. Note that we do not take the strong baseline as our contribution.

4. Method

In this section, we will present the proposed *semantic orthogonal learning framework (SOLF)* and *global- and local-aware feature augmentation method (GLFA)* in detail. The overview of the proposed method is shown in Fig. 2.

4.1. Global-aware feature interpolation with semantic orthogonality

CNN has been known to be good at extracting the abstract high-level feature representations in a deep feature space, where different channels of the feature representations can generate response to different semantics. That is to say, each channel can be seen as an individual global view (semantics) of one input image [52]. Also, channels with larger weights have higher response to the class-specific semantics (e.g., dog's paw and bird's beak in the yellow box in Fig. 1), while the channels with smaller weights are more focused on unimportant semantics (e.g., grass and sky in the blue box in Fig. 1). However, the existing fine-tuning based FSL methods [53] mainly focus on the summarized global feature vector after the global average pooling (GAP) layer, which do not make full use of the informative feature channels. This will loses some important and discriminative semantics information. Therefore, in this paper, we focus on the feature representations before the Global Average Pooling (GAP) layer, and offer a novel channel-wise and spatial-wise perspective on improving few-shot learning.

Semantic orthogonal learning framework (SOLF) Unfortunately, we can not straightforward perform the channel-wise feature augmentation with the standard CNN. This is because the relationships between different feature channels are normally not independently and identically distributed (*non-i.i.d.*), as stated in Zeiler and Fergus [52]. On the other hand, there is generally a feature redundancy in the learned feature channels. Several studies [54] have shown that the diversity of features learned by a convolutional neural network (CNN) can significantly improve classification accuracy. Therefore, it is beneficial to enable the CNN to learn diverse feature channels, which has the potential to lead to improved few-shot Learning performance. To this end, many existing methods [55] apply orthogonality regularization during training to solve this problem. However, previous works have either used complex structures that increase model complexity or approximated the full-rank identity matrix with a non-full rank Gram matrix. In contrast, our proposed method is much simpler yet highly effective in enforcing orthogonality in the feature channels.

Given an image X, we feed it to a CNN-based feature extractor to obtain its feature map $\mathcal{F} \in \mathbb{R}^{C\times H\times W} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_C | \boldsymbol{f}_C \in \mathbb{R}^{1\times HW}\}$. Meanwhile, we apply a two-layer linear block g_ϕ to learn a weight of each channel like the squeeze-and-excitation network [56] (which will be used in the following module M_g). After

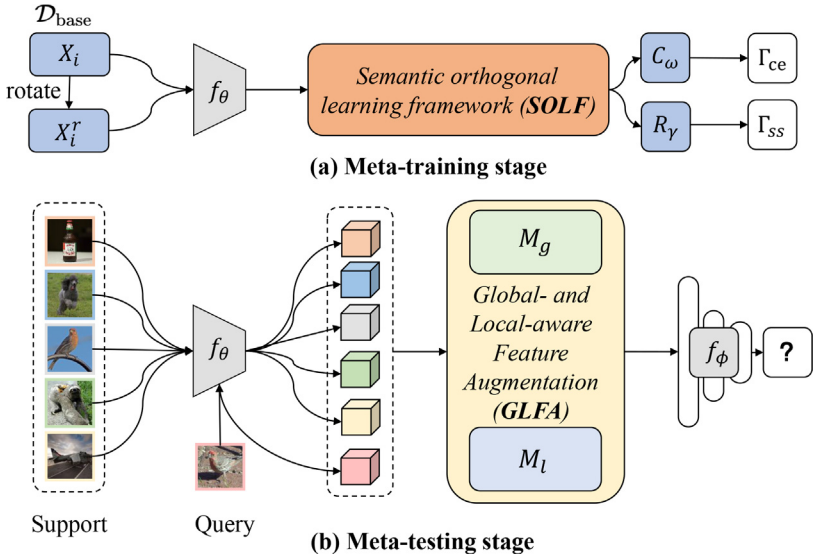


Fig. 2. The overview of proposed method for a 5-way 1-shot task. The above line indicates the meta-training flow, and the below line indicates the meta-testing flow. (a) In meta-training, the proposed *SOLF* enforces the channel orthogonality of features extracted from f_θ. (b) In meta-testing, with f_θ fixed, a new classifier f_ϕ is fine-tuned using the novel task $\mathcal{T}$. In *GLFA*, the global-aware module M_g generates diverse features based on *i.i.d.* channels. The local-aware module M_l is adopted for spatial-wise background smoothing.

that, we can easily calculate the cosine similarity $D_{i,j}$ between any channel pair of $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$:

$$D_{i,j} = \frac{\boldsymbol{f}_i \times \boldsymbol{f}_j^T}{||\boldsymbol{f}_i|| \times ||\boldsymbol{f}_j^T||}, \tag{3}$$

where $\boldsymbol{f}_j^T \in \mathbb{R}^{HW\times 1}$ is the transpose of $\boldsymbol{f}_j$, and thus we can obtain a $C \times C$ square matrix D. Next, we can make this similarity matrix D to be close to the identity matrix I:

$$\Gamma_{os} = \arg\min_\theta ||D - I||_F^2, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

As seen, our orthogonalization operation can ensure the similarity between channel and channel itself (diagonal elements) tends to 1, and the similarity between channel and other channels (non-diagonal elements) tends to 0, so that it can not only encourage the channel's diversity, but also guarantee the independence between channels. In addition, our proposed semantic orthogonal learning framework enhances the orthogonality between different feature channels. leading to reduced disturbances in our subsequent global-aware feature augmentation process. Additionally, we conduct experimrnts to feature explore the benefits of channel orthogonalization, providing deeper insights into the effectiveness of our proposed approach.

Global-aware feature augmentation (M_g) As seen in Fig. 3, after uncoupling the *non-i.i.d.* relationships between different feature channels, we are able to conduct the global-aware (channel-wise) feature augmentation. For each support image X_i in the support set $\mathcal{S}$, to further enhance the diversity of feature channels, we propose a feature interpolation-based approach that selects unimportant feature channels, such as background information, from the same or other classes, and generates multiple diverse feature augmentations via interpolation. By doing so, we can obtain a larger set of diverse feature representations. Specifically, each image X_i will be represented as a three-dimensional (3D) tensor $\mathcal{F}_i = f_\theta(X_i) \in \mathbb{R}^{C\times H\times W}$. In order to selectively perturb the channel's semantics, we first apply g_ϕ learned in SOLF module to acquire a weight of each channel, i.e., $\boldsymbol{w}_i = g_\phi(avg(f_\theta(X_i))) \in \mathbb{R}^C$, where avg indicates the global average pooling operation. $\boldsymbol{w}_i$ is adapted to modulate $\mathcal{F}_i$ and produces $\mathcal{F}_i' = \boldsymbol{w}_i \otimes \mathcal{F}_i$. Next, we choose the top-k

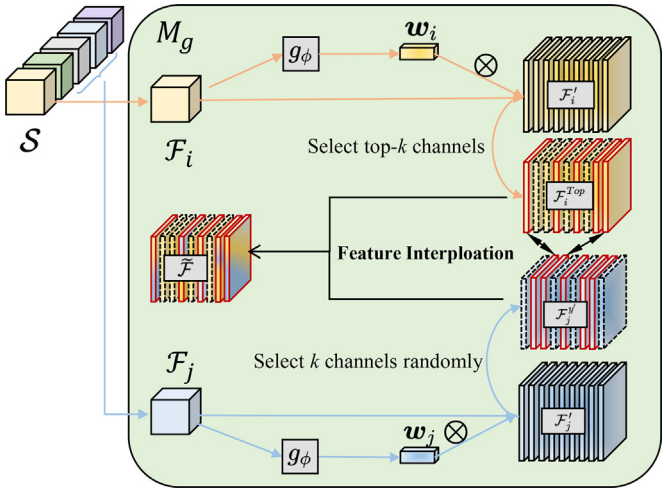


Fig. 3. The framework of the proposed *global-aware (channel-wise) feature augmentation* module M_g for a 5-way 1-shot task.

channels $\mathcal{F}_i^{Top} = \{\boldsymbol{f}_{i,j}^{Top}|_{j=1}^k\}$ from $\mathcal{F}_i'$. For each instance $(\boldsymbol{f}_{i,j}^{Top}, y)$, we randomly select another instance $(\boldsymbol{f}_j^{y'}, y')$ from the feature maps of other images $\mathcal{F}_j^{y'} = \{\boldsymbol{f}_{j,m}^{y'}|_{m=1}^k\}$ (see Fig. 3 for intuitive details). Eventually, a new feature map $\tilde{\mathcal{F}} = \{\tilde{\boldsymbol{f}}_i\}$ is synthesized as follows:

$$\tilde{\boldsymbol{f}}_i = \begin{cases} \lambda \boldsymbol{f}_i^{Top} + (1-\lambda)\boldsymbol{f}_j^{y'} & \text{if } \boldsymbol{f}_i \in \mathcal{F}^{Top} \\ \boldsymbol{f}_i & \text{others} \end{cases} \tag{5}$$

where $\lambda \in [0.5, 1.0]$ is a tradeoff between the selected class label and the original class label. Specifically, we calculate the similarity between $\boldsymbol{f}_i^{Top}$ and $\boldsymbol{f}_j^{y'}$ as λ's selection measure. If the similarity is greater than a threshold, it indicates that the semantics are similar between $\boldsymbol{f}_i^{Top}$ and $\boldsymbol{f}_j^{y'}$. In such cases, we use a relatively smaller λ to obtain larger disturbances from other channel's semantics, which helps to maximize the diversity of the augmented samples while ensuring class identification. Conversely, if the simi-

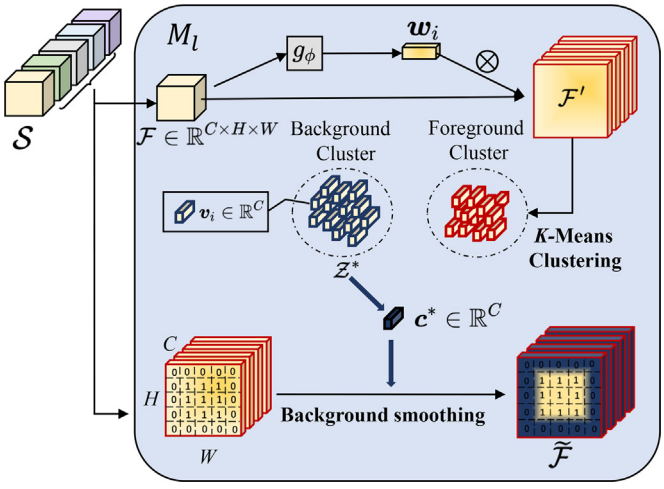


Fig. 4. The framework of the proposed *local-aware (spatial-wise) feature augmentation* module M_l for a 5-way 1-shot task.

larity is lower, a larger λ should be allocated between the channel pair to preserve the class label.

4.2. Local-aware background smoothing (M_l)

In general, different spatial positions of the feature map have different responses to the foreground or background. For an image, not all spatial positions are contributing to the classification equally and sometimes the background information is harmful to the final classification, as described in Luo et al. [20]. To reduce the classification error resulted from background noises, as shown in Fig. 4, we propose a spatial-wise background smoothing method ($\mathcal{M}_l$), which is regarded as another augmentation module, generating a *background smoothed (augmented) version* of each samples in $\mathcal{S}$. Specifically, $\mathcal{M}_l$ consists of two steps:

Local-aware clustering Firstly, given the feature map $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$ of an image, it can be further flatted as a set of local descriptors $\mathcal{V} = \{\boldsymbol{v}_i \in \mathbb{R}^C|_{i=1}^{HW}\} \in \mathbb{R}^{C\times HW}$. Because the local descriptors with similar semantics trends to be clustered together, we directly apply the K-means clustering on $\mathcal{V}$, where $K = 2$, and obtain clusters $\mathcal{Z} = \{\mathcal{Z}^i \in \mathbb{R}^{L_i\times C}|i = 1, 2\}$ centered on $\boldsymbol{c} = \{\boldsymbol{c}_i \in \mathbb{R}^C|i = 1, 2\}$. After that, the weighted cluster representations could be represented as $\boldsymbol{p} = \{\boldsymbol{p}_i|\boldsymbol{p}_i \in \mathbb{R}, i = 1, 2\}$, and $\boldsymbol{p}_i$ is calculated as

$$\boldsymbol{p}_i = \frac{\sum_{j=1}^{L_i}\sum_{m=1}^{C}\mathcal{Z}^i_{j,m}}{L_i \times C}, \tag{6}$$

where L_i denotes the number of local descriptors in the ith cluster, and $\mathcal{Z}^i_{j,m}$ represents the mth response of the jth local descriptor in the ith cluster. [57] believes that the background response is relatively smaller than the class-related foreground, we select the cluster with smaller $p^* \in \boldsymbol{p}$ as the background, which is marked as $\mathcal{Z}^* \in \mathcal{Z}$ centered on $\boldsymbol{c}^* \in \boldsymbol{c}$.

Background smoothing Mean filtering [58] is an effective image smoothing algorithm, which simply replaces the center value with the average of all the pixel values in a window. Analogously, given the background cluster $\mathcal{Z}^*$ and the weighted center $\boldsymbol{c}^*$, our method set the each local descriptor in $\mathcal{Z}^*$ to $\boldsymbol{c}^*$ in a feature level. The augmented feature could be formulated as $\widetilde{\mathcal{F}} = \widetilde{\mathcal{F}}_b \cup \widetilde{\mathcal{F}}_f \in \mathbb{R}^{C\times H\times W}$, where $\widetilde{\mathcal{F}}_b = \{\boldsymbol{f}_i = \boldsymbol{c}^*|\boldsymbol{f}_i \in \mathcal{Z}^*, i = 1, 2, \ldots, H_bW_b\}$, H_bW_b represents the number of local descriptors in the background cluster and $\widetilde{\mathcal{F}}_f$ represents smoothed background and foreground, respectively.

Note that the clustering algorithm is not perfectly accurate, especially when the inter-cluster discrepancy is too small and thus the foreground may be misclassified as background or vice versa. For example, as shown in Fig. 1, the bird's paw and the pole contains less discriminative color semantics. In this paper, the contributions of the foreground and background for classification are modulated by the weight scores $\boldsymbol{w}$ learned by g_ϕ. Therefore, our proposed method could be more robust and balanced to distinct semantics, producing discriminative and diverse augmented features, *i.e.*, $\widetilde{\mathcal{F}}$.

5. Experiments

In this section, we introduce the details and results of the experiments. Firstly, we present dataset information and important implementation details in our design. In addition, we also explored evaluation metrics for few-shot image classification accuracy. Next, we compare our model with the state-of-the-art methods on all benchmark datasets and conduct various ablation studies to verify that each component in our method can effectively boost the classification performance. Finally, we visually verify the effect of channel de-correlation and experimental results of our model on large-way 1-shot classification.

5.1. Datasets

We evaluate the proposed method on four popular FSL benchmark datasets following RENet [38], namely miniImageNet [31], tieredImageNet [59], CIFAR-FewShot (CIFAR-FS for shot) [60] and CUB-200-2011 (CUB for short) [61].

miniImageNet. The miniImageNet is the subset of ImageNet [62] and is the most popular benchmark datasets in few-shot classification task. It contains 100 classes in total, with 600 samples in each class. Following [63], we split all classes into 64 classes for training, 16 classes for validation and 20 classes for testing. The images in miniImageNet are resized to 84×84.

tieredImageNet. The tieredImageNet is also the subset of ImageNet [62]. It contains 608 classes from 34 super-classes, with 779,165 images in total. The dataset is partitioned into 20, 6 and 8 disjoint sets of meta-training, meta-validation and meta-testing according to the super-classes. This split could lessen the similarity between training samples and testing samples. The images in tieredImageNet are resized to 84×84.

CIFAR-FewShot. The CIFAR-FS is randomly sampled from CIFAR-100 [72]. It contains 64, 15, and 20 classes for training, validation, and testing, respectively. The average inter-class similarity is satisfactory high bringing a challenge. Specially, The images in CIFAR-FS have limited original resolution of 32×32. Therefore, we resize them to 32×32.

CUB-200-2011. The CUB is the most popular dataset for fine-grained classification task. It consists of 11,788 images from 200 bird classes. Following [15], we randomly split all classes into 100 classes for training, 50 classes for validation and 50 classes for evaluation. The images in CUB are resized to 84×84.

5.2. Implementation

Network Following the literature [73], we adopt ResNet12 as the embedding backbone, consisting of four residual blocks along with a skip connection layer, in which the numbers of filters of these blocks are $\{64, 160, 320, 640\}$, respectively. In the pre-training stage, for the classification and the rotation prediction heads, we apply a fully-connected (FC) layer. For our proposed *SOLF*, the block g_ϕ is implemented by two linear FC layers. Specially, for the *GLFA* module, it doesn't introduce any additional parameters and we use logistic regression as the new classifier at meta-testing stage.

Training For training, the 5-way 1-shot task has $5 \times 1 = 5$ support images and $5 \times 15 = 75$ query images and the 5-way 5-shot

Table 2
Performance comparison of both 5-way 1-shot and 5-shot tasks in terms of top-1 mean accuracy (%) with 95% confidence intervals on miniImageNet and tieredImageNet datasets. *Results from original papers. †Results from [64] with pre-training.

Method	Backbone	miniImageNet		tieredImageNet	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
R2D2* [60]	96-192-384-512	51.20 ± 0.60	68.80 ± 0.10	–	–
PPA* [65]	WRN-28-10	59.60 ± 0.41	73.74 ± 0.19	65.65 ± 0.92	83.40 ± 0.65
LEO* [42]	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09
SimpleShot* [66]	ResNet18	62.85 ± 0.20	80.02 ± 0.14	69.09 ± 0.22	84.58 ± 0.16
CTM* [14]	ResNet18	64.12 ± 0.82	80.51 ± 0.13	68.41 ± 0.39	84.28 ± 1.73
S2M2* [19]	ResNet18	64.06 ± 0.18	80.58 ± 0.11	–	–
TADAM* [25]	ResNet12	58.50 ± 0.30	76.70 ± 0.30	–	–
MTL* [46]	ResNet12	61.20 ± 1.80	75.50 ± 0.80	–	–
RFS-simple* [16]	ResNet12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55
ProtoNet† [32]	ResNet12	62.39 ± 0.21	80.53 ± 0.14	68.23 ± 0.23	84.03 ± 0.16
MetaOptNet* [43]	ResNet12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
MatchingNet† [31]	ResNet12	65.64 ± 0.20	78.72 ± 0.15	68.50 ± 0.92	80.60 ± 0.71
CAN* [67]	ResNet12	63.85 ± 0.48	79.44 ± 0.34	69.89 ± 0.51	84.23 ± 0.37
DeepEMD* [22]	ResNet12	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
FEAT* [24]	ResNet12	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
ArL* [68]	ResNet12	65.21 ± 0.58	80.41 ± 0.49	–	–
MCL* [69]	ResNet12	64.40	78.60	70.62	83.84
P-Transfer* [70]	ResNet12	64.21 ± 0.77	80.38 ± 0.59	–	–
UAFS* [71]	ResNet12	64.22 ± 0.67	79.99 ± 0.49	69.13 ± 0.84	84.33 ± 0.59
FRN* [64]	ResNet12	66.45 ± 0.19	**82.83 ± 0.13**	71.16 ± 0.22	86.01 ± 0.15
Strong Baseline	ResNet12	65.56 ± 0.36	82.05 ± 0.30	69.92 ± 0.39	83.93 ± 0.28
GLFA (Ours)	ResNet12	**67.25 ± 0.36**	82.80 ± 0.30	**72.25 ± 0.40**	**86.37 ± 0.27**

Table 3
Performance comparison in terms of top-1 mean accuracy (%) with 95% confidence intervals on CUB dataset. *Results from original papers. †Results from [22] with pre-training.

Method	Backbone	5-way 1-shot	5-way 5-shot
RelationNet† [75]	ResNet34	66.20 ± 0.99	82.30 ± 0.58
S2M2* [19]	ResNet34	72.92 ± 0.83	86.55 ± 0.51
MAML† [39]	ResNet34	67.28 ± 1.08	83.47 ± 0.59
S2M2* [19]	ResNet18	71.81 ± 0.43	86.22 ± 0.53
RAP* [76]	ResNet18	74.09 ± 0.60	89.23 ± 0.31
ProtoNet† [32]	ResNet12	66.09 ± 0.92	82.50 ± 0.58
MatchingNet† [31]	ResNet12	71.87 ± 0.85	85.08 ± 0.57
FEAT* [24]	ResNet12	73.27 ± 0.22	85.77 ± 0.14
P-Transfer* [70]	ResNet12	73.88 ± 0.87	87.81 ± 0.48
DeepEMD* [22]	ResNet12	75.65 ± 0.83	88.69 ± 0.50
Strong Baseline	ResNet12	74.93 ± 0.36	89.36 ± 0.35
GLFA (Ours)	ResNet12	**76.52 ± 0.37**	**90.27 ± 0.38**

Table 4
Performance comparison in terms of top-1 mean accuracy (%) with 95% confidence intervals on CIFAR-FS dataset. *Results from original papers. †Results from [74] with pre-training.

Method	Backbone	5-way 1-shot	5-way 5-shot
R2D2* [60]	96-192-384-512	65.30 ± 0.20	79.40 ± 0.10
Boosting* [27]	WRN-28-10	73.62 ± 0.31	86.05 ± 0.22
S2M2* [19]	ResNet18	63.66 ± 0.17	76.07 ± 0.19
Shot-Free* [77]	ResNet12	69.15	84.70
RFS-simple* [16]	ResNet12	71.50 ± 0.80	86.00 ± 0.50
ProtoNet† [32]	ResNet12	72.2 ± 0.70	83.5 ± 0.50
MetaOptNet* [43]	ResNet12	72.0 ± 0.70	84.2 ± 0.50
MABAS* [74]	ResNet12	73.51 ± 0.92	85.49 ± 0.68
Strong Baseline	ResNet12	72.57 ± 0.40	86.67 ± 0.27
GLFA (Ours)	ResNet12	**74.01 ± 0.40**	**87.02 ± 0.27**

task has $5 \times 5 = 25$ support images and $5 \times 15 = 75$ query images. We use stochastic gradient descent (SGD) optimizer with the momentum of 0.9, the weight decay of 5e−4 and the initial learning rate is set to 0.05 and decreased by a factor of 10 every 30 epochs. The total training epochs are 100 and each epoch include 6000 episodes.

Evaluation metrics We focus on the few-shot learning task for the evaluation of model performance based on the classification accuracy. More specifically, a large number of N-way K-shot tasks are sampled from the novel test classes, as shown in 3.1. We evaluate the models on 5*600 sampled tasks to avoid introducing high variance and report the mean accuracy (in %) as well as the 95% confidence interval.

5.3. Comparison with the state of the arts

The results on four benchmark datasets are reported in Tables 2, 3 and 4, respectively. Because the superiority of pre-training on base classes has been demonstrated in many recent FSL works, e.g., SimpleShot [66], RFS-simple [16] and FEAT [24], in this paper, most of the compared methods have employed the pre-training, except MetaOptNet [43], Boosting [27], ArL [68] and MABAS. However, note that both Boosting and ArL use an additional SSL task during training. MABAS [74] also generates samples at the test-time by using adversarial samples, which is closely related to our work. In addition, although CAN [67] and TADAM [25] do not use pre-training, they leverage a global classification as an auxiliary training task by using the global labels of the base classes. S2M2 [19] also applies an SSL auxiliary loss at the pre-training stage, which is closely related to our work. RFS-simple [16] firstly proposes to learn a extractor on mete-training stage by supervised or self-supervised ways and relearn a classifier on meta-testing stage. SimpleShot [66] uses a nearest-neighbor classifier conbination with mean-subtraction and L2-normalization. FEAT [24] applies a embedding adaptation to customize task-specific embedding spaces on the meta-testing satge. We propose corresponding methods in the meta-training and meta-testing stages, and demonstrated excellent results in comparison with the above methods.

First, from all the tables, we can see that the presented *Strong Baseline* have already achieved very competitive results with the comparison methods, which effectively demonstrates the promising potential of the pre-training and fine-tuning paradigm. Second, the proposed GLFA method can further consistently improve the performance over the Strong Baseline on all the datasets. For example, under the 5-way 1-shot setting, GLFA obtains 1.69%, 2.33%, 0.59% and 1.44% improvements over the Strong Baseline on four datasets, respectively. This confidently verifies the effectiveness of

Table 5
Ablation studies on miniImageNet, tieredImageNet and CUB on 5-way 1-shot tasks. GA Augmentation: global-aware feature augmentation; LA Smoothing: local-aware background smoothing; SOLF: semantic orthogonal learning framework.

SOLF	GA Augmentation	LA Smoothing	miniImageNet	tieredImageNet	CUB
✗	✗	✗	65.56 ± 0.35	69.92 ± 0.39	74.93 ± 0.37
√	✗	✗	66.69 ± 0.37	72.14 ± 0.41	75.93 ± 0.36
✗	√	✗	66.43 ± 0.36	71.46 ± 0.40	75.32 ± 0.37
✗	✗	√	65.78 ± 0.37	70.08 ± 0.40	75.11 ± 0.37
√	√	✗	67.03 ± 0.36	72.37 ± 0.40	76.34 ± 0.36
√	√	√	$\mathbf{67.25 \pm 0.36}$	$\mathbf{72.54 \pm 0.40}$	$\mathbf{76.52 \pm 0.37}$

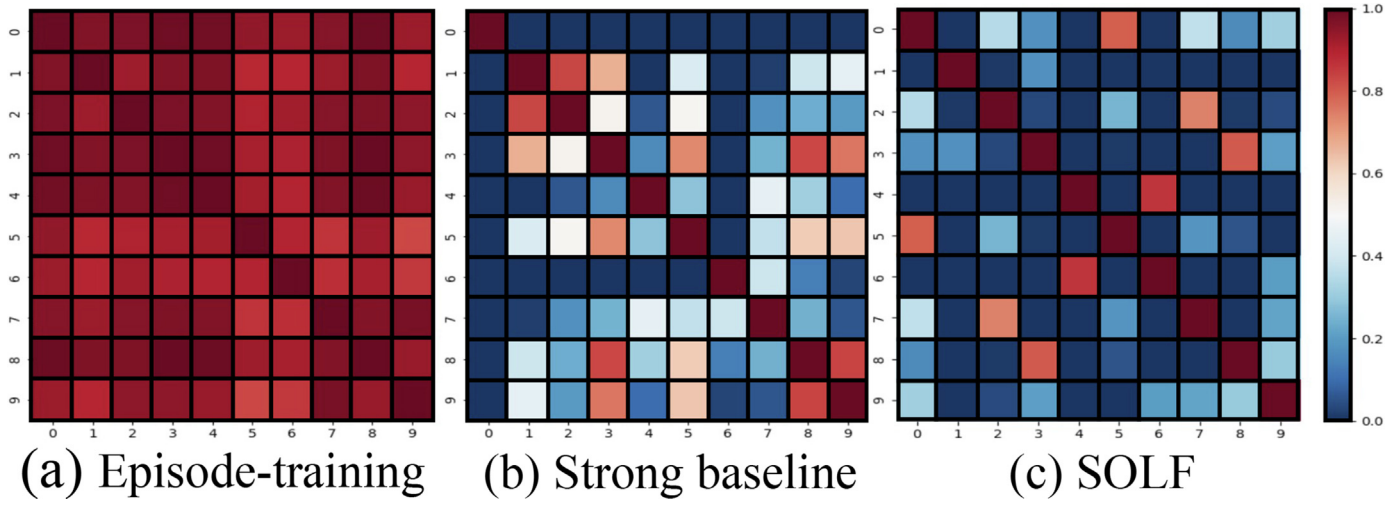


Fig. 5. Visualization of channel correlation matrices from the same local view. The horizontal axis and vertical axis denote the channels. The colors give the cross-channel similarities. The darker red color means the greater similarity between channels. In contrast, the darker blue color means the smaller similarity between channels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the proposed SOLF framework and GLFA method. Finally, compared with the latest FSL methods, no matter pre-training based methods, e.g., DeepEMD [22], FEAT [24] and FRN [64], nor methods using SSL auxiliary tasks, e.g., ArL [68] and Boosting [27], our proposed GLFA consistently outperforms these comparison FSL methods and could achieve new state of the arts on all datasets under both the 5-way 1-shot and 5-way 5-shot settings.

In summary, according to the results and analyses, we could conclude that (1) the pre-training paradigm is indeed effective in the field of FSL; (2) the proposed global- and local-aware feature augmentation module as well as the semantic orthogonal learning framework (SOLF) are indeed effective owing to the ability of alleviating the over-fitting issue in the meta-testing phase.

5.4. Ablation study

To investigate the effects of the core components in our method, we conduct ablation studies on miniImageNet, tieredImageNet and CUB under the 5-way 1-shot setting. Note that there are three core components in our proposed methods, i.e., *semantic orthogonal learning framework (SOLF), global-aware feature augmentation (GA Augmentation)* and *local-aware background smoothing (LA Smoothing)*. We verify the role of different components individually and together.

The experimental results are shown in Table 5, here the results in the first row are the results of the strong baseline. As seen, compared with the strong baseline, all components are able to further improve the classification accuracy on all three datasets. Moreover, we observe that both the GA augmentation and SOLF alone could obtain significant improvements over the baseline. Integrating the LA smoothing with GA augmentation together can also clearly boost the performance. These results successfully demonstrate that the operation of feature de-correlation (orthogonality), i.e., SOLF, is beneficial to the classification. Also, it shows that the GA augmentation is effective to alleviate the overfitting problem at the fine-tuning stage, obtaining a better classification performance. In addition, the background smoothing operation for the spatial regions, i.e., LA smoothing, can also effectively reduces the interference of the background noises to further improves the classification performance. The combination of different components in our method will further improve the model performance.

5.5. Visualization of the channels de-correlation

The fine-tuning based methods have been fruitful in FSL with a much simpler pre-training methodology. Compared to the traditional meta- or episodic-training, the pre-training mechanism seems to be able to obtain better representations. Why is the pre-training mechanism so much better than episodeic-training on the few-shot problems? To explore this point, we visualize the correlation matrices of feature channels learned by different learning paradigms. As seen in Fig. 5, where (a) represents the learned correlation matrix by episodic-training, (b) shows the result learned by the strong baseline mentioned in this paper, and (c) is the result when the proposed SOLF is further applied to the strong baseline. Compare (a) with (b), we can see that the pre-training based model have a better ability on channel de-correlation than the episodic-training based model. This may be attributed to the fact that the features extracted after pre-training will greatly reduce the correlations between channels, so as to extract more discriminative features for classification. The orthogonal and diverse features may boost performance. For this reason, similarly, when our proposed semantic orthogonal learning operation, i.e., SOLF, is added, the correlation between channels is further weakened, the

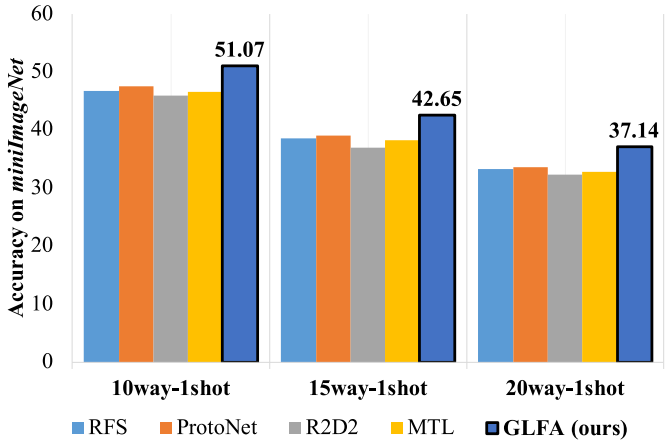


Fig. 6. Classification results under the large-way 1-shot setting on miniImageNet.

discriminative features are further enriched, and the performance of the classification is further improved.

5.6. Large-way 1-shot classification

To further verify the superiority of our proposed method, we compare the proposed method with other closely related fine-tuning based methods, such as RFS [16] and MTL [46], on the large-way 1-shot tasks. In addition, ProtoNet [32] and R2D2 [60] are also re-implemented with pre-training for comparison. From the results in Fig. 6, Our proposed method is shown to be significantly superior to other methods across all settings of 10-way 1-shot, 15-way 1-shot, and 20-way 1-shot tasks. This further highlights the exceptional generalization ability of our method, enabling it to perform well even in the face of challenging tasks. One of the reasons behind the superior performance of our method is the proposed SOLF technique. The SOLF method enhances feature extraction by making feature channels de-correlated, which leads to the extraction of more discriminative features. Moreover, our proposed channel-wise and spatial-wise feature augmentations further contribute to the improved performance of our method. These augmentations make the augmented features more diverse, which is beneficial to the classification results. By combining the SOLF technique with the proposed feature augmentations, we obtain a powerful method for one-shot learning, which outperforms other state-of-the-art methods in the field.

6. Conclusion

In this paper, we follow the pre-training and fine-tuning paradigm to tackle the FSL problem. At the pre-training stage, we propose a *semantic orthogonal learning framework (SOLF)* to make the learned feature channels semantically diverse and orthogonal. At the meta-test stage, we propose a *global- and local-aware feature augmentation method (GLFA)* from both channel and spatial perspectives. Extensive experiments demonstrate that (1) SOLF can efficiently remove the correlations between feature channels, which can learn more diverse and discriminative features; (2) GLFA performs augmentation in a feature level, effectively alleviating the overfitting problem at the fine-tuning stage. We expect that our study can benefit the field of few-shot learning by inspiring new feature augmentation methods. Our approach can address the inadequacy of Generative Adversarial Networks (GANs) that are highly unstable during training. Moever, this methods can serve as a plug-and-play module to match any feature extractor and classifier. This flexibility makes the proposed method easily implementable in existing systems and enhances the potential for wider adoption in the field. Overall, our study can contribute to the development of more effective and efficient few-shot learning algorithms. Morever, we can investigate the effectiveness of our methods on other related tasks such as few-shot object detection or segmentation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (62106100, 62192783, 62276128), Jiangsu Natural Science Foundation (BK20221441), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and Jiangsu Provincial Double-Innovation Doctor Program (JSSCBS20210021).

References

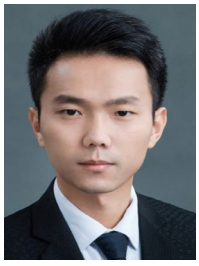
[1] L. Zhang, J. Shen, J. Zhang, J. Xu, Z. Li, Y. Yao, L. Yu, Multimodal marketing intent analysis for effective targeted advertising, IEEE Transactions on Multimedia, 2022.
[2] J. Shen, N. Robertson, BBAS: towards large scale effective ensemble adversarial attacks against deep neural network learning, Information Sciences, 2021.
[3] J. Zeng, J. Zhou, T. Liu, Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities, IEEE Transactions on Multimedia, 2022.
[4] M. Zhang, S. Huang, W. Li, D. Wang, Tree structure-aware few-shot image classification via hierarchical aggregation, in: European Conference onComputer Vision (ECCV), 2022.
[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015.
[6] G. Huang, Z. Liu, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: transformers for image recognition at scale, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[9] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conferenceon Machine Learning (ICML), 2021.
[10] H. Xilang, C. Seon Han, Sapenet: self-attention based prototype enhancement network for few-shot learning, Pattern Recognition, 2022.
[11] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, J. Luo, Revisiting local descriptor based image-to-class measure for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[12] B. Zhang, H. Ling, P. Li, Q. Wang, Y. Shi, L. Wu, R. Wang, J. Shen, Multi-head attention graph network for few shot learning, CMC-COMPUTERS MATERIALS & CONTINUA, 2021.
[13] L. Zijun, H. Zhengping, L. Weiwei, XiaoHua, Sabernet: self-attention based effective relation network for few-shot learning, Pattern Recognition, 2022.
[14] H. Li, D. Eigen, S. Dodge, M. Zeiler, X. Wang, Finding task-relevant features for few-shot learning by category traversal, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[15] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C.F. Wang, J.-B. Huang, A closer look at few-shot classification, in: International Conference on Learning Representations (ICLR), 2018.
[16] Y. Tian, Y. Wang, D. Krishnan, J.B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need? in: European Conference on Computer Vision (ECCV), 2020.
[17] B. Zhang, H. Ling, J. Shen, Q. Wang, J. Lei, Y. Shi, L. Wu, P. Li, Mixture distribution graph network for few shot learning, IEEE Transactions on Cognitive and Developmental Systems, 2022.
[18] T. Yu, S. He, Y.-Z. Song, T. Xiang, Hybrid graph neural networks for few-shot learning, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2022.

[19] P. Mangla, M. Singh, A. Sinha, N. Kumari, V.N. Balasubramanian, B. Krishnamurthy, Charting the right manifold: manifold mixup for few-shot learning, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2020.
[20] X. Luo, L. Wei, L. Wen, J. Yang, L. Xie, Z. Xu, Q. Tian, Rectifying the shortcut learning of background: shared object concentration for few-shot image recognition, Neural Information Processing Systems (NeurIPS), 2021.
[21] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: a review of recent developments, Pattern Recognition, 2023.
[22] C. Zhang, Y. Cai, G. Lin, C. Shen, DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[23] Z. Ji, X. Chai, Y. Yu, Y. Pang, Z. Zhang, Improved prototypical networks for few-shot learning, Pattern Recognition, 2020.
[24] H.-J. Ye, H. Hu, D.-C. Zhan, F. Sha, Few-shot learning via embedding adaptation with set-to-set functions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[25] B.N. Oreshkin, P.R. López, A. Lacoste, TADAM: task dependent adaptive metric for improved few-shot learning, Neural Information Processing Systems (NeurIPS), 2018.
[26] Z. Lei, Z. Fei, W. Wei, Z. Yanning, Meta-hallucinating prototype for few-shot learning promotion, Pattern Recognition, 2022.
[27] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, M. Cord, Boosting few-shot visual learning with self-supervision, in: IEEE International Conference on Computer Vision (ICCV), 2019.
[28] J. Rajasegaran, S.H. Khan, M. Hayat, F.S. Khan, M. Shah, Self-supervised knowledge distillation for few-shot learning, 2020.
[29] Y. Chen, Z. Liu, H. Xu, T. Darrell, X. Wang, Meta-baseline: exploring simple meta-learning for few-shot learning, in: IEEE International Conference on Computer Vision (ICCV), 2021.
[30] H.-J. Ye, L. Ming, D.-C. Zhan, W.-L. Chao, Few-shot learning with a strong teacher, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
[31] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, Neural Information Processing Systems (NeurIPS), 2016.
[32] J. Snell, K. Swersky, R.S. Zemel, Prototypical networks for few-shot learning, Neural Information Processing Systems (NeurIPS), 2017.
[33] K.R. Allen, E. Shelhamer, H. Shin, J.B. Tenenbaum, Infinite mixture prototypes for few-shot learning, in: International Conferenceon Machine Learning (ICML), 2019.
[34] J. Liu, L. Song, Y. Qin, Prototype rectification for few-shot learning, in: European Conference on Computer Vision (ECCV), 2020.
[35] Y. Lifchitz, Y. Avrithis, S. Picard, A. Bursuc, Dense classification and implanting for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[36] C. Doersch, A. Gupta, A. Zisserman, Crosstransformers: spatially-aware few-shot transfer, Neural Information Processing Systems (NeurIPS), 2020.
[37] J. Xie, F. Long, J. Lv, Q. Wang, P. Li, Joint distribution matters: deep brownian distance covariance for few-shot classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[38] D. Kang, H. Kwon, J. Min, M. Cho, Relational embedding for few-shot classification, in: IEEE International Conference on Computer Vision (ICCV), 2021.
[39] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conferenceon Machine Learning (ICML), 2017.
[40] A. Nichol, J. Schulman, Reptile: a scalable metalearning algorithm, 2018.
[41] S.B. Aniruddh Raghu, M. Raghu, O. Vinyals, Reptile: a scalable metalearning algorithm, in: International Conference on Learning Representations (ICLR), 2020.
[42] A.A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, R. Hadsell, Meta-learning with latent embedding optimization, in: International Conference on Learning Representations (ICLR), 2019.
[43] K. Lee, S. Maji, A. Ravichandran, S. Soatto, Meta-learning with differentiable convex optimization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[44] J. Oh, H. Yoo, C. Kim, S.-Y. Yun, Boil: towards representation change for few-shot learning, in: International Conference on Learning Representations (ICLR), 2021.
[45] A. Rajeswaran, C. Finn, S.M. Kakade, S. Levine, Meta-learning with implicit gradients, Neural Information Processing Systems (NeurIPS), 2019.
[46] Q. Sun, Y. Liu, T.-S. Chua, B. Schiele, Meta-transfer learning for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[47] G.S. Dhillon, P. Chaudhari, A. Ravichandran, S. Soatto, A baseline for few-shot image classification, in: International Conference on Learning Representations (ICLR), 2020.
[48] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, H. Hu, Negative margin matters: understanding margin in few-shot classification, in: European Conference on Computer Vision (ECCV), 2020.
[49] J.-C. Su, S. Maji, B. Hariharan, When does self-supervision improve few-shot learning? in: International Conference on Learning Representations (ICLR), 2020.
[50] N. Komodakis, S. Gidaris, Unsupervised representation learning by predicting image rotations, in: International Conference on Learning Representations (ICLR), 2018.
[51] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[52] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, 2013.
[53] M. Chen, Y. Fang, X. Wang, H. Luo, Y. Geng, X. Zhang, C. Huang, W. Liu, B. Wang, Diversity transfer network for few-shot learning, Association for the Advancement of Artificial Intelligence (AAAI), 2020.
[54] L. Huang, X. Liu, B. Lang, A.W. Yu, Y. Wang, B. Li, Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks, Association for the Advancement of Artificial Intelligence (AAAI), 2018.
[55] H. Xu, Z. Wang, H. Yang, D. Liu, J. Liu, Learning simple thresholded features with sparse support recovery, 2019.
[56] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[57] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad–cam: visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision (ICCV), 2017.
[58] T.S. Huang, Two-dimensional digital signal processing II. Transforms and median filters, Two-Dimensional Digital Signal Processing II. Transforms and Median Filters, 1981.
[59] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J.B. Tenenbaum, H. Larochelle, R.S. Zemel, Meta-learning for semi-supervised few-shot classification, in: International Conference on Learning Representations (ICLR), 2018.
[60] L. Bertinetto, J.a.F. Henriques, P.H.S. Torr, A. Vedaldi, Meta-learning with differentiable closed-form solvers, in: International Conference on Learning Representations (ICLR), 2019.
[61] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, 2011.
[62] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision, 2015.
[63] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: International Conference on Learning Representations (ICLR), 2017.
[64] W.D.T. Luming, Few-shot classification with feature map reconstruction networks, in: IEEE International Conference on Computer Vision (ICCV), 2021.
[65] S. Qiao, C. Liu, W. Shen, A.L. Yuille, Few-shot image recognition by predicting parameters from activations, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[66] Y. Wang, W.-L. Chao, K.Q. Weinberger, L. van der Maaten, Simpleshot: revisiting nearest-neighbor classification for few-shot learning, 2019.
[67] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, Neural Information Processing Systems (NeurIPS), 2019.
[68] H. Zhang, P. Koniusz, S. Jian, H. Li, H.S. Torr Philip, Rethinking class relations: absolute-relative supervised and unsupervised few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[69] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, X. He, Learning to affiliate: mutual centralized learning for few-shot classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[70] S. Zhiqiang, L. Zechun, Q. Jie, S. Marios, C. Kwang-Ting, Partial is better than all: revisiting fine-tuning strategy for few-shot learning, Association for the Advancement of Artificial Intelligence (AAAI), 2021.
[71] Z. Zhang, C. Lan, W. Zeng, Z. Chen, S.-F. Chang, Uncertainty-aware few-shot image classification (2020).
[72] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images, 2009.
[73] W. Li, C. Dong, P. Tian, T. Qin, X. Yang, Z. Wang, J. Huo, Y. Shi, L. Wang, Y. Gao, et al., Libfewshot: a comprehensive library for few-shot learning, 2021.
[74] J. Kim, H. Kim, G. Kim, Model-agnostic boundary-adversarial sampling for test–time generalization in few-shot learning, in: European Conference on Computer Vision (ECCV), 2020.
[75] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H.S. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[76] J. Hong, P. Fang, W. Li, T. Zhang, C. Simon, M. Harandi, L. Petersson, Reinforced attention for few-shot learning and beyond, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[77] A. Ravichandran, R. Bhotika, S. Soatto, Few-shot learning with embedded class models and shot-free meta training, in: IEEE International Conference on Computer Vision (ICCV), 2019.

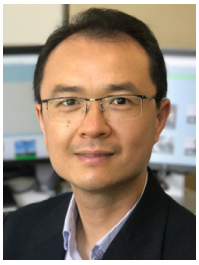
Boyao Shi received the B.Sc. degree from Nanjing University of Finance & Economics in 2020. She is currently working towards the M.Sc. degree with the Department of Computer Science and Technology, Nanjing University, Nanjing, China. She is a member of R&L Group, which is led by Professor Yang Gao. Her research interests include machine learning and computer vision, with a focus on few shot learning.

Wenbin Li received his Ph.D. degree from the Department of Computer Science and Technology at Nanjing University in 2019. He is currently an assistant researcher in the Department of Computer Science and Technology at Nanjing University, China. His research interests include machine learning and computer vision, particularly in metric learning, fewshot learning and their applications to image classification and image generation.

Jing Huo received the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2017. She is currently an Associate Researcher with the Department of Computer Science and Technology, Nanjing University. Her current research interests include machine learning and computer vision, with a focus on subspace learning, adversarial learning and their applications to heterogeneous face recognition and cross-modal face generation.



Pengfei Zhu received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin, China, in 2009 and 2011, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2015. He is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests are focused on machine learning and computer vision.



Lei Wang received his Ph.D. degree from Nanyang Technological University, Singapore. He is now Professor at School of Computing and Information Technology of University of Wollongong, Australia. His research interests include machine learning, pattern recognition, and computer vision. Lei Wang has published more than 190 peer-reviewed papers, including those in highly regarded journals and conferences such as IEEE TPAMI, IJCV, CVPR, ICCV and ECCV, etc. He was awarded Early Career Researcher Award by Australian Academy of Science and Australian Research Council. He served as General Co-Chair of DICTA 2014 and Area Chair of ICIP2019, and will serve as Program Co-Chair of ACCV2022 in Macau. Lei Wang is senior member of IEEE.

Yang Gao received the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, China, in 2000. He is currently a Professor and also the Deputy Director of the Department of Computer Science and Technology, Nanjing University, where he is also directing the Reasoning and Learning Research Group. He has published more than 100 papers in top-tier conferences and journals. His current research interests include artificial intelligence and machine learning. He also serves as the program chair and area chair for many international conferences.