Pattern Recognition 159 (2025) 111089

Contents lists available at ScienceDirect

Pattern Recognition

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

Exploring sample relationship for few-shot classification

Xingye Chen [a,1], Wenxiao Wu [b,1], Li Ma [c], Xinge You [a], Changxin Gao [b], Nong Sang [b], Yuanjie Shao [a,*]

[a] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, 430074, China
[b] National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074, China
[c] Faculty of Mechanical and Electronic Information, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Dataset link: https://github.com/Chenguoz/SRE

Keywords:
Few-shot classification
Sample Relationship Exploration
Meta-learning
Transfer learning

ABSTRACT

Few-shot classification (FSC) is a challenging problem, which aims to identify novel classes with limited samples. Most existing methods employ vanilla transfer learning or episodic meta-training to learn a feature extractor, and then measure the similarity between the query image and the few support examples of novel classes. However, these approaches merely learn feature representations from individual images, overlooking the exploration of the interrelationships among images. This neglect can hinder the attainment of more discriminative feature representations, thus limiting the potential improvement of few-shot classification performance. To address this issue, we propose a Sample Relationship Exploration (SRE) module comprising the Sample-level Attention (SA), Explicit Guidance (EG) and Channel-wise Adaptive Fusion (CAF) components, to learn discriminative category-related features. Specifically, we first employ the SA component to explore the similarity relationships among samples and obtain aggregated features of similar samples. Furthermore, to enhance the robustness of these features, we introduce the EG component to explicitly guide the learning of sample relationships by providing an ideal affinity map among samples. Finally, the CAF component is adopted to perform weighted fusion of the original features and the aggregated features, yielding category-related embeddings. The proposed method is a plug-and-play module which can be embedded into both transfer learning and meta-learning based few-shot classification frameworks. Extensive experiments on benchmark datasets show that the proposed module can effectively improve the performance over baseline models, and also perform competitively against the state-of-the-art algorithms. The source code is available at https://github.com/Chenguoz/SRE.

1. Introduction

In recent years, deep learning methods have achieved significant advances in various visual recognition tasks. However, the training of a deep learning model significantly relies on abundant labeled training to achieve a good performance. In contrast, the human visual system can recognize new objects with only a few labeled examples available. Inspired by human capabilities of learning from a small number of samples, the concept of few-shot classification (FSC) [1–3] has emerged at the forefront of deep learning to quickly adapt to new tasks with limited annotated data.

Due to the limitation of training data, there exist many issues like overfitting and poor generalization in FSC. In order to solve these problems, various methods have been explored, among which meta-learning based and transfer learning based methods account for the vast majority. In terms of transfer learning based methods, many works [4–6] encode sufficient prior information to initialized parameters with various pre-training strategies and some others [5,7,8] improve the loss function during the stage of fine-tuning. Unlike transfer learning based approaches aimed at obtaining better pre-trained models and fine-tuning loss functions, meta-learning based methods [9–12] attempt to capture the ability to quickly adapt to new tasks.

Although a variety of advancements have been made by both kinds of methods to enhance loss functions or the overall structural aspects of few-shot models, the majority of them tend to simply learn feature embeddings from an individual image, overlooking the huge potential of relationships among images. In fact, in FSC tasks, due to the small number of samples in each class, there may be significant variation

* Corresponding author.
E-mail addresses: chenxingye@hust.edu.cn (X. Chen), wenxiaowu@hust.edu.cn (W. Wu), mali@cug.edu.cn (L. Ma), youxg@hust.edu.cn (X. You), cgao@hust.edu.cn (C. Gao), nsang@hust.edu.cn (N. Sang), shaoyuanjie@hust.edu.cn (Y. Shao).
[1] Equal contributions.

https://doi.org/10.1016/j.patcog.2024.111089
Received 24 March 2024; Received in revised form 15 October 2024; Accepted 16 October 2024
Available online 28 October 2024

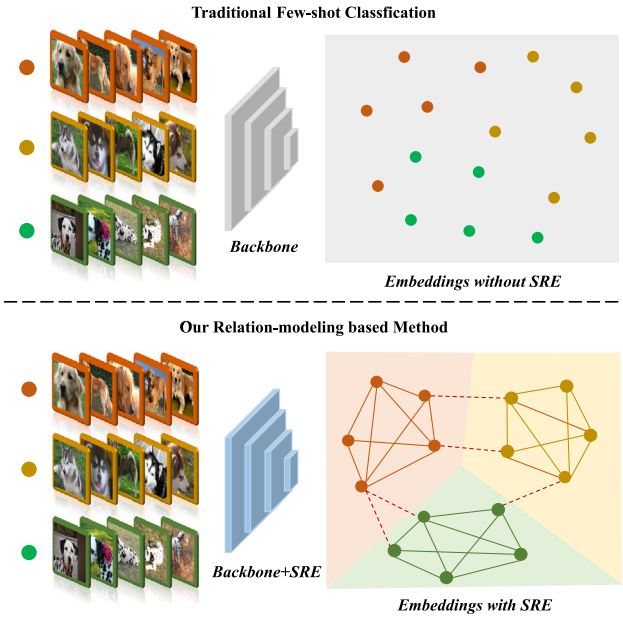

Fig. 1. Comparison between traditional few-shot classification methods and our relation-modeling based method. The proposed method aims to obtain more robust and discriminative sample features by aggregating category information across samples.

within the class. Extracting features from a single image may lead to poor discriminability, which in turn affects the improvement of few-shot classification performance. Effectively exploring the class similarity relationship among images during the feature extraction process and aggregating similar sample features can help reduce intra-class variation and increase the discriminability of features.

Inspired by above observation, in this paper, we propose a Sample Relation Exploration (SRE) module for both transfer learning and meta-learning frameworks in FSC. Specifically, inspired by the role of attention mechanism [13,14] in exploring relationships between patches, a Sample-level Attention (SA) component is employed to model similarity relations among samples. In this way, a task-related similarity map and aggregated features of similar samples can be obtained. In addition, we introduce the Explicit Guidance (EG) component to explicitly supervise the task-related similarity map by an ideal affinity map constructed by class information. We hope that such an explicit constraint can help the model to achieve more compact intra-class clusters in feature space from a global perspective, so as to enhance the robustness of features. Last but not least, we also propose a Channel-wise Adaptive Fusion (CAF) component to dynamically and adaptively fuse both the original and aggregated features. By doing so, a category-related feature with information extracted both individually and jointly from one task can be obtained. As illustrated in Fig. 1, by exploring sample relationships with the proposed SRE module, we can integrate category-specific information into features and enhance the connection between similar images, thereby obtaining a more discriminative embedding space.

We summarize our contributions as three-folds:

- In this paper, we propose a Sample Relation Exploration (SRE) module to explore category information from similar samples and construct a more discriminative feature space. Three components in SRE module are employed to model similarity relationships, guide task-related similarity map learning and fuse category information, respectively.
- Our SRE module is a plug-and-play module, which can be seamlessly integrated with any feature extractor of both meta-learning based and transfer learning based methods for FSC.
- Extensive experiments are conducted on several few-shot learning benchmarks. The experimental results show that our SRE module can effectively improve the performance of both meta-learning and transfer learning based methods.

2. Related work

2.1. Few-shot classification

Among various few-shot classification methods, meta-learning based approaches [9,15–17] have attracted many follow-up works [4,18–20] due to their elegant formulation and suitability for few-shot learning scenarios. Wertheimer et al. [11] introduce a novel mechanism for few-shot classification by reconstructing each query sample from the support set in a closed form, without introducing any new modules or large-scale learnable parameters. Xie et al. [6] propose a Deep Brownian Distance Covariance for few-shot classification to replace the conventional features, applying it to both meta-learning and transfer learning frameworks. These works have designed different learning mechanisms to address the few-shot classification problem. Recently, a surprising finding is that a simple transfer learning baseline [5,6], which involves training a supervised model on the training set followed by fine-tuning with a simple adaptation algorithm (such as logistic regression), actually outperforms many methods based on meta-learning [9,16,21]. Since simple supervised training is not specifically designed for few-shot classification, this observation suggests that the training algorithm can be designed without considering the choice of adaptation algorithm while still achieving satisfactory performance. It is noteworthy that above methods do not fully consider the relationships among samples during feature extraction. To address this issue, we propose a sample relationship exploration module which can be seamlessly integrated into both the meta-learning and transfer learning based FSC frameworks.

2.2. Attention-based method for few-shot classification

The fundamental idea behind attention-based methods [11,22,23] is to incorporate attention mechanisms into the model, which can suppress the influence of irrelevant features and thus improve classification performance in few-shot settings. Liu et al. [24] propose to exploit universal features for few-shot classification by dynamically re-weighting and composing the most appropriate domain-specific representations. Lai et al. [12] propose a novel Clustered-patch Element Connection layer to correct the mismatch problem of semantic information between local patches. However, we observe that most of these methods do not explicitly guide the learning of attention models, making it difficult to learn discriminative sample features and unable to analyze the benefits of attention models for few-shot scenarios. Therefore, we propose a novel Sample-level Attention component, which can be explicitly guided by label information to learn category relationships between samples and obtain more robust features.

2.3. Transductive few-shot classification

There are also some works [25–27] introduce the transductive FSC method, building upon the foundation of inductive FSL to achieve more robust performance. Simon et al. [25] introduce an approach called Adaptive Subspace for few-shot learning, which learns a set of subspaces to adapt to different tasks, resulting in improved generalization capability. Chen et al. [28] propose ECKPN, a network that leverages inter-class knowledge for label propagation from query to support set, enhancing generalization. Recently, methods based on Label Propagation (LP) [8,29,30] have gained prominence in transductive few-shot learning. These methods construct a graph from the support set and the entire query set, propagating labels within the graph, and have become the mainstream in transductive few-shot learning due to their

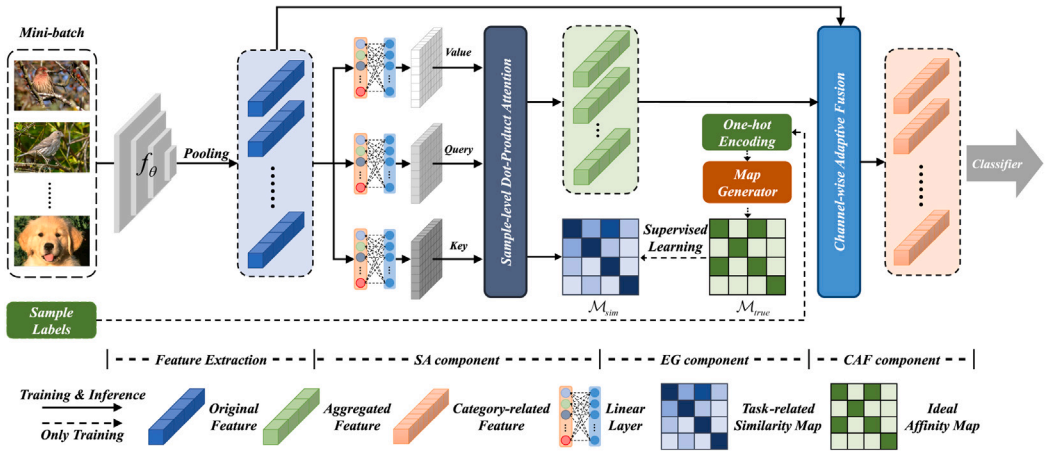

Fig. 2. Overview of the proposed method. We employ either episodic meta-training or vanilla transfer learning from scratch to train our SRE module. Firstly, we feed the original features extracted from the shared backbone to the SA component, model the task-related similarity map of the samples, and obtain aggregated features containing interactions among similar samples. Then, to further improve the reliability of task-related similarity map, the EG component is employed to explicitly guide the similarity learning via an ideal affinity map based on the one-hot encoding of labels. Finally, the CAF component is used to achieve a balance between aggregated features and original features, generating category-related features that incorporate sample relationships.

significant performance improvements. In contrast, our method focuses on utilizing all available information from the input samples during the feature extraction stage of the model, by constructing a task-related similarity map among the samples to address the issue of fragile features learned by the model in few-shot scenario.

3. Method

In this section, we first introduce the proposed Sample Relationship Exploration (SRE) module in detail. Then we embed our SRE module into two few-shot classification frameworks. Finally, we discuss the differences between our SRE module and previous relationship modeling approaches.

3.1. Sample Relationship Exploration (SRE) module

As illustrated in Fig. 2, the proposed SRE module consists of three main components: the Sample-level Attention (SA), Explicit Guidance (EG) and Channel-wise Adaptive Fusion (CAF) component. Given a mini-batch of input images $\{x_i\}_{i=1}^{B}$, d-dimensional feature vectors $z_i \in \mathbb{R}^{1\times d}$ are obtained through the backbone feature extractor $f_\theta(\cdot)$. Then, the feature vectors from the mini-batch are concatenated as original features $\mathbf{Z} \in \mathbb{R}^{B\times d}$ and fed into the SA component to estimate the task-related similarity map $\mathcal{M}_{sim} \in \mathbb{R}^{B\times B}$ of the samples, and produce the aggregated features $\mathbf{Z}_c \in \mathbb{R}^{B\times d}$ containing correlation information between similar samples. Additionally, the one-hot encoding of the sample label is fed into the EG component to construct the ideal affinity map $\mathcal{M}_{true} \in \mathbb{R}^{B\times B}$ for explicit supervision of similarity learning, thereby further improving the reliability of the task-related similarity map. Finally, $\mathbf{Z}$ and $\mathbf{Z}_c$ are adaptively fused through the CAF component to finally generate category-related features that contain sample relationships.

3.1.1. Sample-level attention component

To achieve interaction between feature information among samples and explicitly model the intra-class relationships among samples, this method is inspired by self-attention mechanisms [22] and their extension to the sample dimension [13] for modeling the similarity relations among samples.

Specifically, for the original features $\mathbf{Z} \in \mathbb{R}^{B\times d}$ of input samples, this method employs an sample-level dot-product attention mechanism to transform the feature embeddings of input samples, thereby measuring inter-sample category similarities. Here, the query, key and value vector $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times d}$ are all derived from the original features $\mathbf{Z}$ through three linear transformations, as shown in following Eq. (1).

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = Linear_{q,k,v}(\mathbf{Z}). \tag{1}$$

Unlike traditional attention mechanisms, we use the sigmoid function $\psi(\cdot)$ instead of softmax function to better model the similarity between sample categories as follows:

$$\mathcal{M}_{sim} = \psi(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d}}), s_{ij} \in [0, 1]. \tag{2}$$

The element s_{ij} in category attention map $\mathcal{M}_{sim} = \{s_{ij} | i, j \in 1, 2, \ldots, B\}$ indicates the similarity between the ith and jth image, and d is the scaling factor in attention mechanism. This task-related similarity map $\mathcal{M}_{sim}$ can be supervised by the ideal affinity map constructed by Explicit Guidance component in Section 3.1.2.

Then, we reweight the value vector $\mathbf{V}$ through the task-related similarity map $\mathcal{M}_{sim}$ to obtain corresponding category aggregated features $\mathbf{Z}_c \in \mathbb{R}^{B\times d}$ of each sample:

$$\mathbf{Z}_c = \phi(\mathcal{M}_{sim}) \times \mathbf{V}, \tag{3}$$

where $\times$ represents the operation of matrix products, and the softmax function $\phi(\cdot)$ ensures that the weighted aggregated features are consistent with the original feature scale. In this way, a task-related similarity map $\mathcal{M}_{sim}$ and aggregated features $\mathbf{Z}_c$ of similar samples can be obtained.

3.1.2. Explicit guidance component

In FSC tasks, each sample is associated with a true label during training. It is challenging for the network to model the relationship information between isolated samples. In this section, we calculate the ideal affinity map of input samples based on their true labels to explicitly guide the network in learning the correct relationships between samples.

For a single mini-batch or episode of input samples with a total of B samples and L classes, the one-hot encoding matrix is denoted as $Onehot \in \mathbb{R}^{B\times L}$. As shown in the map generator of Fig. 3, the category relationship can be explicitly modeled as:

$$\mathcal{M}_{true} = Onehot \times Onehot^{T}, \tag{4}$$

where $\mathcal{M}_{true} = \{t_{ij} | i, j = 1, \ldots, B\} \in \mathbb{R}^{B\times B}$ represents the ideal affinity map between samples.

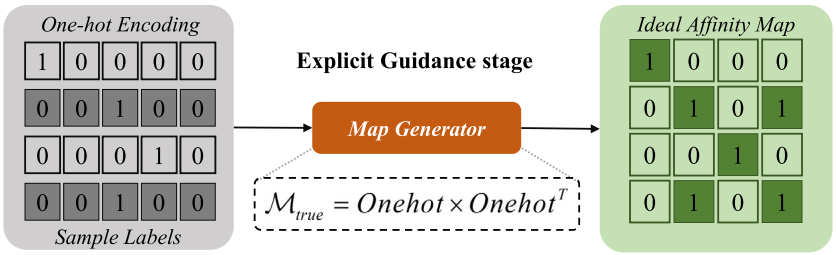

Fig. 3. Example of the map generator. In training phase, the one-hot encodings of input samples are packed into mini-batch input and passed through the map generator to obtain the ideal affinity map.

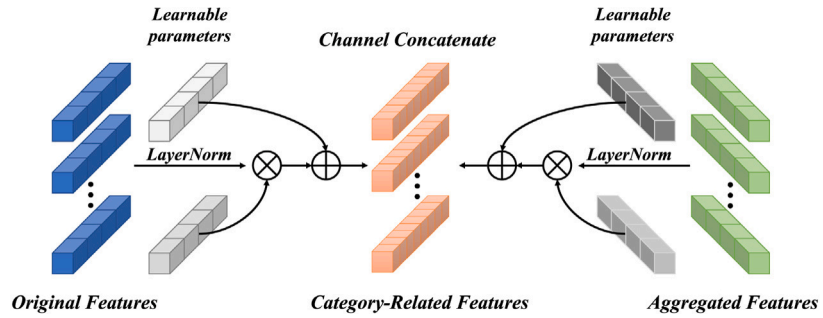

Fig. 4. The structure of Channel-wise Adaptive Fusion component. The original features and aggregated features are first normalized, then weighted by learnable parameters and finally concatenated to obtain category-related features.

As a result, the ideal affinity map $\mathcal{M}_{true}$ can act as the additional supervision for task-related similarity map $\mathcal{M}_{sim}$ in Eq. (2), enabling the network to capture the real relationships between samples.

For each element in the task-related similarity map $\mathcal{M}_{sim}$ and ideal affinity map $\mathcal{M}_{true}$, a binary cross-entropy loss function can be denoted as follows:

$$\mathcal{L}_u = -\frac{1}{B^2}\sum_{j=1}^{B}\sum_{i=1}^{B}\Big(s_{ij}\log t_{ij} + (1-s_{ij})\log\Big(1-t_{ij}\Big)\Big), \tag{5}$$

where $s_{ij} \in \mathcal{M}_{sim}$ and $t_{ij} \in \mathcal{M}_{true}$.

By constructing such a class relation loss to explicitly guide the learning of task-related similarity map $\mathcal{M}_{sim}$, it helps the model achieve more compact intra-class clusters in the feature space from a global perspective, thus enhancing the generalization ability of models.

3.1.3. Channel-wise adaptive fusion component

As illustrated in Fig. 4, to dynamically and adaptively integrate the original features of input samples with their aggregated features based on class information, we propose the Channel-wise Adaptive Fusion component.

Due to the different focuses of original features and aggregated features, learnable feature weighting parameters should be applied to each channel. Therefore, we design two learnable parameters for original features and aggregated features of input samples: channel weight parameters $\gamma \in \mathbb{R}^{1\times d}$ and channel bias parameters $\beta \in \mathbb{R}^{1\times d}$. For each feature, a linear transformation function $\varphi(\cdot)$ parameterized by γ and β, is denoted as follows:

$$\varphi(z) = \frac{z - E[z]}{\sqrt{Var[z]+\epsilon}} \times \gamma + \beta, z \in \mathbb{R}^{1\times d}, \tag{6}$$

where $E(\cdot)$ represents the mean operation, $Var(\cdot)$ represents the variance operation, and ϵ is a small constant. The learnable parameter weights of the original features and aggregated features are independent of each other.

By linearly weighting the original features and aggregated features, and then concatenating them along the channel dimension, we obtain the category-related features $\mathbf{Z}_f$ containing information separately and jointly extracted from one task, as follows:

$$\mathbf{Z}_f = \varphi(\mathbf{Z}) \oplus \varphi(\mathbf{Z}_c), \mathbf{Z}_f \in \mathbb{R}^{B\times 2d}, \tag{7}$$

where $\oplus$ represents the operation of concatenation.

Finally, we use the category-related features for subsequent few-shot classification tasks.

3.2. SRE for few-shot classification

The proposed method is a plug-and-play module that can be embedded in both meta-learning and vanilla transfer learning frameworks such as ProtoNet [9] and EASY [5]. In this section, we will introduce the two methods and present how to integrate our SRE module into them.

3.2.1. Problem definition

We consider a standard FSC problem, where a training set D_{train} and a test set D_{test} are provided. The FSC model f_θ is trained on a series of tasks randomly sampled from D_{train} and is subsequently tested on tasks randomly sampled from D_{test}. Each task consists of two disjoint sets: the support set S and the query set Q. Following the "N-way K-shot" setting, $S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s=N\times K}$ is composed of n_s annotated images from N classes, with K images for each class; and $Q = \{(x_i^q, y_i^q)\}_{i=1}^{n_q}$ shares the same label space with S, containing n_q images to be classified. During the training stage, the FSC model learns to generalize on the support set and then performs evaluation on the query set sampled from D_{train}. This training strategy emulates the model's behavior in dealing with a small number of samples, thereby enhancing the model's robustness.

3.2.2. SRE-ProtoNet

Recent meta-learning based methods, such as ProtoNet [9], have shown promising results in few-shot learning tasks [6,21,31]. ProtoNet utilizes a feature extractor to compute class prototypes:

$$\overline{\mathbf{c}_k} = \frac{1}{|S_k|}\sum_{(x_i^s, y_i^s)\sim S_k} f_\theta(x_i), \tag{8}$$

where S_k represents the samples from class k, and f_θ represents the feature extractor. The class probability for a query sample x_q is then computed using softmax:

$$p_\theta(y_i = k|x_q) = \frac{\exp(-d(f_\theta(x_q), \overline{\mathbf{c}_k}))}{\sum_{k'} \exp(-d(f_\theta(x_q), \overline{\mathbf{c}_{k'}}))}, \tag{9}$$

where $d(\cdot)$ is the Euclidean distance between query and prototype.

To address the limitation of traditional prototype networks in capturing inter-class relationships, we propose integrating the SRE module into the feature extractor f_θ. This enhancement allows for more robust and discriminative feature representations. We adopt the cross-entropy loss $\mathcal{L}_s$ to minimize the prediction error of query samples for each task as follows:

$$\mathcal{L}_s = -\frac{1}{n_q} \sum_{q=1}^{n_q} \sum_{k=1}^{N} \mathbf{x}_q \log(p_\theta(y_i = k|x_q)) \tag{10}$$

Finally, based on explicit guidance loss $\mathcal{L}_u$ in Eq. (5) for category relation learning and the cross-entropy loss $\mathcal{L}_s$ for classification tasks, the overall loss is defined as:

$$\mathcal{L}_p = \lambda_s \mathcal{L}_s + \lambda_u \mathcal{L}_u, \tag{11}$$

where λ_s and λ_u are hyperparameters controlling the effect of each loss term.

3.2.3. SRE-EASY

Similarly, we embed the SRE module to the end of the feature extractor f_θ in the EASY [5] framework to explore the potential relationships between samples. Unlike ProtoNet, EASY uses mini-batch training instead of episodic training and incorporates a fully connected layer as a classifier. During testing, this classifier is replaced by a soft K-means algorithm for iterative computation of class centers, leveraging unlabeled query samples:

$$\forall i, t : \begin{cases} \overline{\mathbf{c}}_i^0 & = & \overline{\mathbf{c}}_i \,, \\ \overline{\mathbf{c}}_i^{t+1} & = & \sum_{\mathbf{z} \in S_i \cup Q} \frac{w(\mathbf{z}, \overline{\mathbf{c}}_i^t)}{\sum_{\mathbf{z}' \in S_i \cup Q} w(\mathbf{z}', \overline{\mathbf{c}}_i^t)} \mathbf{z} \,, \end{cases} \tag{12}$$

where S_i represents the ith class of the support set, and Q represents the query set samples. $w(\mathbf{z}, \overline{\mathbf{c}}_i^t)$ is a weighting function on $\mathbf{z}$, which gives it a probability of being associated with class center $\overline{\mathbf{c}}_i^t$, obtained by calculating the Euclidean distance from the class center.

In addition to the classifier used to identify the category of the input samples, a new logistic regression classifier is branched out in the penultimate layer of the backbone to retrieve the four possible rotations corresponding to the rotation data augmentation used. Finally, based on the rotation loss and the loss mentioned in Eq. (11), the overall loss is defined as:

$$\mathcal{L}_p = \lambda_r \mathcal{L}_r + \lambda_s \mathcal{L}_s + \lambda_u \mathcal{L}_u, \tag{13}$$

where $\mathcal{L}_r$ represents the rotation loss used to supervise the network in learning the four possible rotations of the image and λ_r is hyperparameter controlling the effect of each loss term.

3.3. Relation with previous relationship exploration methods

In this subsection, we discuss the connections between several existing relationship exploration methods and our proposed approach. Let $X \in \mathbb{R}^{B \times C}$ represent the feature vectors and $X' \in \mathbb{R}^{B \times C \times H \times W}$ represent the feature maps of input samples, where B is the number of samples, and C denotes the dimensionality of input features.

BatchFormer [13] introduces a batch converter module that enables deep neural networks to learn relationships between samples from each mini-batch. Different from the typical usage of transformer layers, the input samples in BatchFormer are simply reshaped into individual patches, enabling the Transformer layers to operate on the batch dimension of the input data.

NFormer [14] models interactions between input images by computing an approximate affinity matrix $\widetilde{A}$ that represents the relations between individual representations:

$$\widetilde{A} = \frac{(\mathbf{Q}\widetilde{\mathbf{K}_l^{\mathbf{T}}})(\mathbf{K}\widetilde{\mathbf{Q}_l^{\mathbf{T}}})}{\sqrt{d}} \tag{14}$$

where $\widetilde{K_l}$ and $\widetilde{Q_l}$ represents landmark agents sampled from the original Q, K to map high-dimensional vector X to a low-dimensional encoding space.

RENet [32] extracts structural patterns in images by learning auto-correlation and cross-correlation patterns and acquires reliable image-wide attention $\widetilde{A}_\mathrm{q}$ through convolutional filtering:

$$\widetilde{A}_\mathrm{q}(\mathbf{x}_\mathrm{q}) = \frac{1}{HW} \sum_{\mathbf{x}_s} \frac{\exp\left(\hat{\mathbf{C}}(\mathbf{x}_\mathrm{q}, \mathbf{x}_s)/\boldsymbol{\gamma}\right)}{\sum_{\mathbf{x}'_\mathrm{q}} \exp\left(\hat{\mathbf{C}}(\mathbf{x}'_\mathrm{q}, \mathbf{x}_s)/\boldsymbol{\gamma}\right)} \tag{15}$$

where $\mathbf{x}$ is a position at the feature map X' and $\boldsymbol{\gamma}$ is a temperature factor. $\hat{\mathbf{C}}(\mathbf{x}_\mathrm{q}, \mathbf{x}_s)$ is a matching score between the positions $\mathbf{x}_\mathrm{q}$ and $\mathbf{x}_s$.

FEAT [18] employs a set-to-set transformation, specifically a self-attention mechanism, to transform the embedding of each training instance while simultaneously considering its contextual instances:

$$A = \phi(\frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{d}}), \mathbf{Q}, \mathbf{K} \in D_{train}, \tag{16}$$

where **Q,K** is obtained by linear mapping through D_{train} and $\phi(\cdot)$ is the softmax function.

The above four relationship exploration methods either simply apply attention to the sample level without explicit constraints or fail to fully explore the relationships among all associated samples. On the contrary, the proposed SRE module enable interactions among all samples, and provide explicitly guidance for the learning of similarity relationships by constructing an ideal affinity map, thus can obtain more accurate sample relationships. In addition, through adaptive feature fusion, we obtain both individually extracted and jointly extracted sample embeddings tailored for the current task, making our method more suitable for few-shot scenarios.

4. Experiments

In this section, we first provide experimental settings about implementation details and datasets evaluated on. Then, we insert our proposed SRE module into two few-shot frameworks, ProtoNet [9] and EASY [5], and compare these improved versions with recent state-of-the-art approaches.

4.1. Experimental settings

For fair comparisons with previous methods, we adopt ResNet12 [33] as the backbone network. The input resolution of images is 84×84 and features with 640 dimensions are obtained by the feature extractor after the global pooling operation. Our SRE-ProtoNet is based on ProtoNet [9]. Each task follows the standard "5-way 1/5-shot" FSC setting, sampled uniformly from the meta-training or meta-testing sets. The 5-way 1-shot accuracy measures the model's performance when classifying samples from 5 novel classes given only 1 labeled example per class, while 5-way 5-shot accuracy uses 5 labeled examples per class. Following the previous protocol used in [6,34], before the meta-training stage, we employ a pre-trained model with its weights used as initialization. In SRE-ProtoNet, the values of λ_s and λ_u are set to 0.75 and 0.25 respectively. Our SRE-EASY is based on a simple transfer learning framework EASY [5] and the values of λ_s, λ_r and λ_u are set to 0.375, 0.375 and 0.25 respectively. Methods with an asterisk use random resized crops for feature enhancement during testing stage. We evaluate on four few-shot classification benchmarks: *mini*ImageNet [3,35], *tiered*ImageNet [36], CUB-200-2011 [37], and CIFAR-FS [38]. Results are reported with 95% confidence intervals to indicate the statistical reliability of our findings.

Table 1
Comparison with the state-of-the-art 5-way 1-shot and 5-way 5-shot performance (%) with 95% confidence intervals on *mini*ImageNet and *tiered*ImageNet.

Method	Setting	Backbone	*mini*ImageNet		*tiered*ImageNet	
			1-shot	5-shot	1-shot	5-shot
Good-Embed [4]	Inductive	ResNet12	64.82 ± 0.60	82.14 ± 0.43	71.52 ± 0.69	86.03 ± 0.58
DeepEMD [31]	Inductive	ResNet12	65.91 ± 0.82	82.41 ± 0.56	71.16 ± 0.87	86.03 ± 0.58
FEAT [18]	Inductive	ResNet12	66.78 ± 0.20	82.05 ± 0.14	70.80 ± 0.23	84.79 ± 0.16
RENet [32]	Inductive	ResNet12	67.60 ± 0.44	82.58 ± 0.30	71.61 ± 0.51	85.28 ± 0.35
FRN [11]	Inductive	ResNet12	66.45 ± 0.19	82.83 ± 0.13	72.06 ± 0.22	86.89 ± 0.14
BLC-MAML [39]	Inductive	ResNet12	66.26 ± 0.32	82.92 ± 0.30	–	–
ProtoNet+CL [40]	Inductive	ResNet12	66.17 ± 0.46	81.73 ± 0.30	69.37 ± 0.48	85.73 ± 0.37
STL DeepBDC [6]	Inductive	ResNet12	67.83 ± 0.43	85.45 ± 0.29	73.82 ± 0.47	89.00 ± 0.30
CECNet [12]	Inductive	ResNet12	69.32 ± 0.46	84.65 ± 0.32	73.14 ± 0.50	86.88 ± 0.36
SAPENet[41]	Inductive	ResNet12	66.41 ± 0.20	82.76 ± 0.14	68.63 ± 0.23	84.30 ± 0.16
GLFA[42]	Inductive	ResNet12	67.25 ± 0.36	82.80 ± 0.30	72.25 ± 0.40	86.37 ± 0.27
LeadNet [43]	Inductive	ResNet12	67.32 ± 0.20	83.21 ± 0.14	72.42 ± 0.23	86.60 ± 0.16
SCL [44]	Inductive	ResNet12	66.79 ± 0.43	83.39 ± 0.29	70.58 ± 0.50	85.39 ± 0.35
UniSiam [45]	Inductive	ResNet34	64.77 ± 0.37	81.75 ± 0.26	67.67 ± 0.39	84.12 ± 0.28
cluster-FSL [46]	Transductive	ResNet12	77.81 ± 0.81	85.55 ± 0.41	83.89 ± 0.81	89.94 ± 0.46
EPNet-SSL [47]	Transductive	ResNet12	75.36 ± 1.01	84.07 ± 0.60	81.79 ± 0.97	88.45 ± 0.61
MetaOptNet [48]	Transductive	ResNet12	62.64 ± 0.82	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
BatchFormer† [13]	Transductive	ResNet12	62.59 ± 0.40	80.20 ± 0.32	66.17 ± 0.46	83.81 ± 0.36
NFormer† [14]	Transductive	ResNet12	64.42 ± 0.45	81.52 ± 0.30	67.81 ± 0.53	85.14 ± 0.35
CAN+T [10]	Transductive	ResNet12	67.19 ± 0.55	80.64 ± 0.35	73.21 ± 0.58	84.93 ± 0.38
DSN-MR [25]	Transductive	ResNet12	64.60 ± 0.72	79.51 ± 0.50	67.39 ± 0.82	82.85 ± 0.56
COSOC* [49]	Transductive	ResNet12	69.28 ± 0.49	85.16 ± 0.42	73.57 ± 0.43	87.57 ± 0.10
MCT* [50]	Transductive	ResNet12	78.55 ± 0.86	86.03 ± 0.42	82.32 ± 0.81	87.36 ± 0.50
ODC* [51]	Transductive	ResNet18	77.20 ± 0.36	87.11 ± 0.42	83.73 ± 0.36	90.46 ± 0.46
PDN-PAS[52]	Transductive	ResNet50	66.38 ± 0.82	84.29 ± 0.56	71.13 ± 0.97	85.75 ± 0.66
ProtoNet [9]	Inductive	ResNet12	62.39 ± 0.21	80.53 ± 0.14	68.23 ± 0.23	84.03 ± 0.16
ProtoNet† [9]	Transductive	ResNet12	68.31 ± 0.44	82.18 ± 0.30	68.81 ± 0.52	84.50 ± 0.35
SRE-ProtoNet	Transductive	ResNet12	**68.73 ± 0.43**	**84.31 ± 0.29**	**70.73 ± 0.50**	**86.40 ± 0.32**
SRE-ProtoNet*	Transductive	ResNet12	**70.86 ± 0.44**	**87.01 ± 0.27**	**73.02 ± 0.50**	**87.78 ± 0.31**
Meta DeepBDC [6]	Inductive	ResNet12	67.34 ± 0.43	84.46 ± 0.28	72.34 ± 0.49	87.31 ± 0.32
SRE-Meta DeepBDC	Transductive	ResNet12	**68.91 ± 0.46**	**85.07 ± 0.29**	**71.46 ± 0.52**	**87.31 ± 0.31**
SRE-Meta DeepBDC*	Transductive	ResNet12	**71.85 ± 0.44**	**87.50 ± 0.28**	**74.22 ± 0.50**	**87.95 ± 0.32**
EASY*† [5]	Transductive	ResNet12	81.89 ± 0.24	88.95 ± 0.12	82.98 ± 0.26	88.10 ± 0.15
SRE-EASY*	Transductive	ResNet12	**84.07 ± 0.22**	**89.87 ± 0.11**	**84.45 ± 0.24**	**89.66 ± 0.14**

† Reproduced under our settings.
* Data augmentation with random resized crops.

Table 2
Comparison with the state-of-the-art 5-way 1-shot and 5-way 5-shot performance (%) with 95% confidence intervals on CUB.

Method	Reference	Backbone	CUB	
			1-shot	5-shot
MatchNet [15]	NIPS'16	ResNet12	71.87 ± 0.85	85.08 ± 0.57
MAML [53]	ICML'17	ResNet18	68.42 ± 1.07	83.47 ± 0.62
Baseline++ [21]	ICLR'19	ResNet18	67.02 ± 0.90	83.58 ± 0.54
ADM [54]	IJCAI'20	ResNet18	79.31 ± 0.43	90.69 ± 0.21
FRN [11]	CVPR'21	ResNet18	82.55 ± 0.19	92.98 ± 0.10
EASY* [5]	ArXiv'22	ResNet12	90.56 ± 0.19	93.79 ± 0.10
Meta DeepBDC [6]	CVPR'22	ResNet18	83.55 ± 0.40	93.82 ± 0.17
STL DeepBDC [6]	CVPR'22	ResNet18	84.01 ± 0.42	94.02 ± 0.24
FGM [55]	ICCV'23	ResNet12	80.77 ± 0.90	92.01 ± 0.71
SAPENet[41]	PR'23	Conv4-64	70.38 ± 0.23	84.47 ± 0.14
COML [56]	PR'23	ResNet12	83.93 ± 0.66	93.95 ± 0.30
GLFA[42]	PR'23	ResNet12	76.52 ± 0.37	90.27 ± 0.38
LEADNET[43]	TPAMI'23	ResNet12	79.05 ± 0.20	90.85 ± 0.11
ProtoNet† [9]	NIPS'17	ResNet12	80.73 ± 0.43	92.17 ± 0.21
SRE-ProtoNet	-	ResNet12	**84.34 ± 0.40**	**92.42 ± 0.20**
SRE-ProtoNet*	-	ResNet12	**86.71 ± 0.38**	**93.19 ± 0.20**
EASY*† [5]	ArXiv'22	ResNet12	89.81 ± 0.19	93.48 ± 0.09
SRE-EASY*	-	ResNet12	**92.10 ± 0.17**	**94.26 ± 0.08**

† Reproduced under our settings.
* Data augmentation with random resized crops.

4.2. Comparison with state-of-the-art methods

As shown in Table 1, we compare our approach with state-of-the-art few-shot methods on *mini*ImageNet and *tiered*ImageNet. It indicates that our approach outperforms the existing state-of-the-art methods, demonstrating the effectiveness and advantages of our SRE module. Here are several observations. First of all, all baseline methods achieve significant performance gains on *mini*ImageNet and *tiered*ImageNet under 5-way-1-shot and 5-way-5-shot settings with our proposed SRE module inserted, which demonstrates our method obtains more discriminative representations by exploring potential category information among samples. For example, under the 5-shot setting, the performance of ProtoNet is improved from 82.18% to 84.31% by inserting our method on *mini*ImageNet, and nearly 5% improvement is achieved with the strong data augmentation.

Secondly, compared with the meta-learning based method like Meta DeepBDC [6] and the transfer-based method STL-DeepBDC, SRE-ProtoNet* and SRE-EASY* achieved 2.55% and 4.42% accuracy superiority under the 5-shot setting on *mini*ImageNet, which also illustrates the effectiveness of our SRE module. In addition, when compared with some methods using attention mechanisms, such as COSOC [49] and CECNet [12], our SRE-EASY* outperform them by 2.09% and 2.78% on *tiered*ImageNet, respectively. From Table 1, we can observe that our method outperforms other relationship exploration methods, such as RENet [32] and FEAT [18]. These existing approaches often apply attention at the sample level without explicit constraints or fail to fully explore inter-sample relationships. In contrast, our proposed SRE module enables interactions among all samples and provides explicit guidance for learning similarity relationships by constructing an ideal affinity map. This approach leads to more accurate sample relationship exploration and improved feature aggregation. Finally, please note that our SRE module achieves greater improvements in conjunction with baseline methods employing random resized crops enhancement.

Table 3
Comparison with the state-of-the-art 5-way 1-shot and 5-way 5-shot performance (%) with 95% confidence intervals on CIFAR-FS.

Method	Reference	Backbone	CIFAR-FS	
			1-shot	5-shot
ProtoNet [9]	NIPS'17	ResNet12	72.20 ± 0.70	83.50 ± 0.50
Cosine [21]	ICLR'19	ResNet34	60.39 ± 0.28	72.85 ± 0.65
MetaOptNet [48]	CVPR'19	ResNet12	72.80 ± 0.70	85.00 ± 0.50
Boosting [57]	CVPR'19	WRN28	73.60 ± 0.30	86.00 ± 0.20
RFS[4]	ECCV'20	ResNet12	73.90 ± 0.80	86.90 ± 0.50
S2M2 [58]	WACV'20	ResNet18	63.66 ± 0.17	76.07 ± 0.19
RENet [32]	CVPR'21	ResNet12	74.51 ± 0.46	86.60 ± 0.32
EASY* [5]	ArXiv'22	ResNet12	87.16 ± 0.23	90.63 ± 0.15
CORL[59]	WACV'23	ResNet12	74.13 ± 0.71	87.54 ± 0.51
GLFA[42]	PR'23	ResNet12	74.01 ± 0.40	87.02 ± 0.27
ProtoNet† [9]	NIPS'17	ResNet12	73.51 ± 0.54	84.32 ± 0.16
SRE-ProtoNet	-	ResNet12	**77.03 ± 0.48**	**87.55 ± 0.32**
SRE-ProtoNet*	-	ResNet12	**78.81 ± 0.47**	**89.11 ± 0.31**
EASY*† [5]	ArXiv'22	ResNet12	86.95 ± 0.22	89.85 ± 0.15
SRE-EASY*	-	ResNet12	**87.52 ± 0.21**	**90.62 ± 0.15**

† Reproduced under our settings.
* Data augmentation with random resized crops.

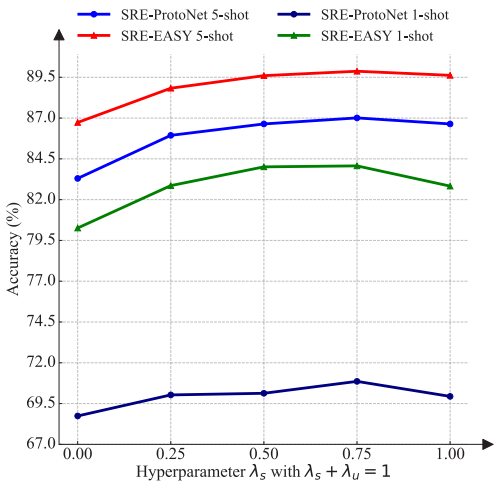

Fig. 5. The impact of different hyperparameters scales of EG component on *mini*ImageNet. Triangular and circular data points represent the EASY and ProtoNet in conjunction with our SRE module respectively.

This result shows that our method can be combined with a variety of advanced feature extractors or feature enhancement methods to perform better, demonstrating its scalability and generalization.

Moreover, we present experimental results on two fine-grained datasets. As shown in Tables 2 and 3, even when evaluated on fine-grained classification tasks, our method can also exhibit significant improvements on ProtoNet and EASY. For example, under 1-shot setting, our SRE module improves the performance of ProtoNet from 80.73% to 84.34% on CUB and achieves a nearly 5% accuracy increase on CIFAR-FS. Meanwhile, it can also be seen that the improvements on EASY brought by our SRE module range from 1% to 3%, which is not as significant as those on ProtoNet. Besides the influence of marginal effects, the reason behind this may be that the way EASY calculates prototypes by combining all samples is similar to the calculation of the task-related similarity map in our SRE module.

4.3. Ablation study

4.3.1. Comparison of latency of few-shot classification task

In this section, we compare the latency during the meta-training and meta-testing phases with single RTX 3090 in Table 4. All the experiments are conducted on *mini*ImageNet with ResNet-12 as backbone and we calculate the relative value of the latency based on the experiment results given in [6]. In the stage of meta-training, ProtoNet exhibit the lowest latency as the simplest metric learning based methods, while our SRE-ProtoNet is just a little slower. However, our SRE-ProtoNet with only a small number of learnable parameters added outperforms almost all the ProtoNet-based methods. Though the meta-training and meta-testing speed of FRN is comparable to that of our SRE-ProtoNet, its accuracy is much lower. Compared with STL DeepBDC, our SRE-EASY* method substantially outperforms it by 16.24% on the 1-shot setting while using shorter time. These results demonstrate that our SRE module efficiently balances the trade-off between performance and computational overhead, making it a viable solution for practical applications in few-shot learning scenarios.

4.3.2. Effect of SRE module during the training stage

Following previous works [6,18,34], we analyze our SRE module's effectiveness by incorporating it into pre-training, meta-training, or both phases. As shown in Table 5, integrating SRE in either phase alone improves performance by 3.7% and 1.2% on average in 1-shot and 5-shot settings, respectively. Notably, incorporating SRE in both stages yields the most significant improvement for ProtoNet. This indicates that the two stages with SRE are complementary, and capturing category information in both phases enhances feature robustness, improving the model's classification performance.

4.3.3. Revisiting on different guidance methods in EG component

Here, we first demonstrate the impact of the EG component on performance under different loss scales. As shown in Fig. 5, it can be observed that the network achieves the best performance when the hyperparameters of classification loss λ_s and explicit guidance loss λ_u are set to 0.75 and 0.25 respectively. It is worth noting that when $\lambda_u = 0$, the learning of task-related similarity map transits from applying additional guidance to completely implicit learning. This results in a significant drop in classification performance in the 1-shot setting, confirming the effectiveness of the proposed EG component. Additionally, we conduct an interesting experiment to investigate the use of CLIP [60] for feature extraction and similarity calculation in the SRE module. Table 6 presents two approaches: "CLIP-supervised", where M_{sim} learning is supervised by CLIP-generated M_{true}, and "CLIP-measured", where CLIP-computed sample distance scores directly replace M_{sim}. As we expected, since the scores computed by the CLIP model are merely estimates of the true class relationship of the samples, the performance of both CLIP-based methods is slightly inferior to our method.

4.3.4. Comparison of different feature fusion methods

In addition to using proposed CAF component to fuse the original features and aggregated features, we also explore the usage of traditional pointwise-addition operation and channel-concatenation operation for feature fusion. As illustrated in Table 7, the performance improvement of channel-concatenation operation is slightly larger than that of pointwise-addition operation. Moreover, our CAF component shows a 1% to 2% improvement compared to the above two methods in both 1-shot and 5-shot settings. Consequently, the experimental results show that our CAF component is able to better fuse the two features by learning the scales of different feature channels and normalizing them in order to achieve better performance improvement.

4.3.5. Evaluation on the contribution of three proposed component

As shown in Table 8, we conduct comprehensive ablation studies to evaluate the contribution of each proposed component in our SRE module. The experiments are performed on the *mini*ImageNet dataset using both meta-learning (ProtoNet [61]) and transfer learning (EASY [5]) frameworks, under vanilla and strong data augmentation settings. Our

Table 4
Comparison of FLOPs, parameters and latency (ms) for 5-way classification on *mini*ImageNet.

Method	FLOPS/Params	Latency					
		Meta-training		Meta-testing		Accuracy	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [9]	3.5×10^9/12.5M	304	365	115	143	62.11	80.77
ADM [54]	2.2×10^8/7.64M	908	967	199	221	65.87	82.05
DeepEMD [31]	3.5×10^9/12.5M	>80K	$>10^6$	457	12,617	65.91	82.41
FRN [11]	3.5×10^9/12.5M	327	427	159	197	66.45	82.83
Meta DeepBDC [6]	3.5×10^9/12.5M	505	623	161	198	67.34	84.46
STL DeepBDC [6]	3.5×10^9/12.5M	–	–	184	245	67.83	85.45
EASY [5]†	3.5×10^9/12.5M	–	–	134	157	78.49	86.71
SRE-ProtoNet°	3.5×10^9/13.7M	371	467	152	184	68.73	84.31
SRE-ProtoNet*	3.5×10^9/13.7M	373	463	316	335	**70.86**	**87.01**
SRE-EASY°	3.5×10^9/13.7M	–	–	170	206	78.86	87.32
SRE-EASY*	3.5×10^9/13.7M	–	–	439	532	**84.07**	**89.87**

† Reproduced under our settings.
∘ Vanilla data augmentation.
* Strong data augmentation.

Table 5
The effect of the SRE module employed in different training stages, P represents the pre-training stage, and M represents the meta-training stage. ✓ and ✗ represent respectively using or not using our SRE module.

(a) Vanilla data augmentation.

Method	P	M	*miniImageNet*	
			1-shot	5-shot
ProtoNet°	✗	✗	62.39 ± 0.21	80.53 ± 0.14
SRE-ProtoNet°	✓	✗	67.37 ± 0.45	82.66 ± 0.28
	✗	✓	67.84 ± 0.46	83.73 ± 0.30
	✓	✓	**68.73 ± 0.43**	**84.31 ± 0.29**

(b) Strong data augmentation.

Method	P	M	*miniImageNet*	
			1-shot	5-shot
ProtoNet*	✗	✗	65.53 ± 0.36	83.59 ± 0.30
SRE-ProtoNet*	✓	✗	68.58 ± 0.46	84.72 ± 0.28
	✗	✓	69.25 ± 0.46	84.78 ± 0.28
	✓	✓	**70.86 ± 0.44**	**87.01 ± 0.27**

Table 6
The comparison results of using different ways to guide or replace the task-related similarity map learning in SRE module on *mini*ImageNet.

Method	SRE-ProtoNet		SRE-EASY	
	1-shot	5-shot	1-shot	5-shot
CLIP-supervised°	67.26 ± 0.44	83.47 ± 0.29	77.95 ± 0.23	86.96 ± 0.12
CLIP-supervised*	70.26 ± 0.43	86.37 ± 0.27	83.52 ± 0.22	88.83 ± 0.12
CLIP-measured°	66.98 ± 0.45	83.26 ± 0.30	77.61 ± 0.22	86.57 ± 0.13
CLIP-measured*	70.23 ± 0.44	86.67 ± 0.27	83.75 ± 0.20	88.94 ± 0.12
Ours (SRE)°	68.73 ± 0.43	84.32 ± 0.29	78.86 ± 0.23	87.32 ± 0.12
Ours (SRE)*	**70.86 ± 0.44**	**87.01 ± 0.27**	**84.07 ± 0.22**	**89.87 ± 0.11**

∘ Vanilla data augmentation.
* Strong data augmentation.

Table 7
The comparison results of replacing CAF component with pointwise-addition (Addition) or channel-concatenation (Concatenation) operation on *mini*ImageNet.

Method	SRE-ProtoNet		SRE-EASY	
	1-shot	5-shot	1-shot	5-shot
Addition°	68.16 ± 0.45	82.51 ± 0.31	76.75 ± 0.23	86.64 ± 0.13
Addition*	69.60 ± 0.46	85.61 ± 0.31	82.32 ± 0.22	88.58 ± 0.12
Concatenation°	68.59 ± 0.46	83.11 ± 0.29	77.80 ± 0.24	87.34 ± 0.12
Concatenation*	69.84 ± 0.46	85.96 ± 0.27	83.67 ± 0.23	89.20 ± 0.11
Ours°(CAF)	68.73 ± 0.43	84.31 ± 0.29	78.86 ± 0.23	87.32 ± 0.12
Ours*(CAF)	**70.86 ± 0.44**	**87.01 ± 0.27**	**84.07 ± 0.22**	**89.87 ± 0.11**

∘ Vanilla data augmentation.
* Strong data augmentation.

Table 8
Evaluation on the contribution of three proposed component in our SRE module on *mini*ImageNet. ✓ and ✗ represent respectively adopting or not adopting the module.

(a) Vanilla data augmentation.

SA	CAF	EG	SRE-ProtoNet°		SRE-EASY°	
			1-shot	5-shot	1-shot	5-shot
✗	✗	✗	62.39 ± 0.21	80.53 ± 0.14	75.92 ± 0.23	84.31 ± 0.13
✓	✗	✗	66.54 ± 0.44	82.47 ± 0.30	75.68 ± 0.23	83.92 ± 0.13
✓	✓	✗	68.15 ± 0.45	83.23 ± 0.28	77.30 ± 0.23	86.84 ± 0.13
✓	✗	✓	68.59 ± 0.46	83.11 ± 0.29	77.80 ± 0.24	87.34 ± 0.12
✓	✓	✓	**68.73 ± 0.43**	**84.31 ± 0.29**	**78.86 ± 0.23**	**87.32 ± 0.12**

(b) Strong data augmentation.

SA	CAF	EG	SRE-ProtoNet*		SRE-EASY*	
			1-shot	5-shot	1-shot	5-shot
✗	✗	✗	65.53 ± 0.36	83.59 ± 0.30	81.89 ± 0.25	88.29 ± 0.13
✓	✗	✗	68.94 ± 0.45	85.37 ± 0.29	81.21 ± 0.23	88.17 ± 0.11
✓	✓	✗	69.94 ± 0.44	86.64 ± 0.26	82.83 ± 0.23	89.62 ± 0.11
✓	✗	✓	69.60 ± 0.46	85.61 ± 0.31	82.32 ± 0.22	88.58 ± 0.12
✓	✓	✓	**70.86 ± 0.44**	**87.01 ± 0.27**	**84.07 ± 0.22**	**89.87 ± 0.11**

results demonstrate that each component contributes significantly to the overall performance improvement. With the SA component as the core module, we analyze the impact of CAF and EG. In the vanilla data augmentation setting, incorporating SA alone improve the 1-shot accuracy by 4.15% for ProtoNet. Adding CAF and EG further increase the accuracy by 1.61% and 0.44%, respectively. The combination of all three components yield the best performance, with a total improvement of 6.34% in 1-shot accuracy for ProtoNet and 2.94% for EASY. These consistent improvements across various frameworks and data augmentation settings confirm the generalizability of our approach and demonstrate that all three components complement each other in enhancing network performance.

4.3.6. Qualitative results

In Fig. 6, we present the class activation maps visualized with Grad-CAM++ [62] for EASY and the improved version with our SRE module inserted on *mini*ImageNet. It is evident that our approach enables the model to focus more on the finer details of objects. For example, in

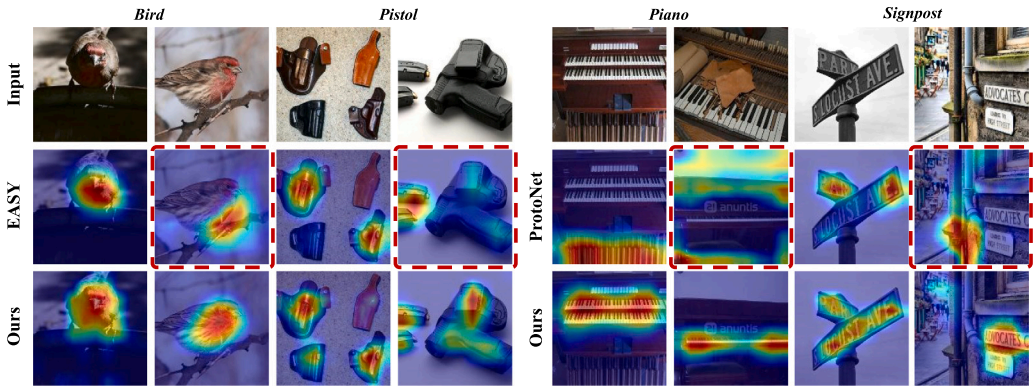

Fig. 6. The Grad-CAM++ [62] visualization on *mini*ImageNet. The first row illustrates the original images, the second row displays the class activation maps of the original model, and the third row represents the results after improvements with our module. The images with red boxes indicate results that are misclassified by the original model.

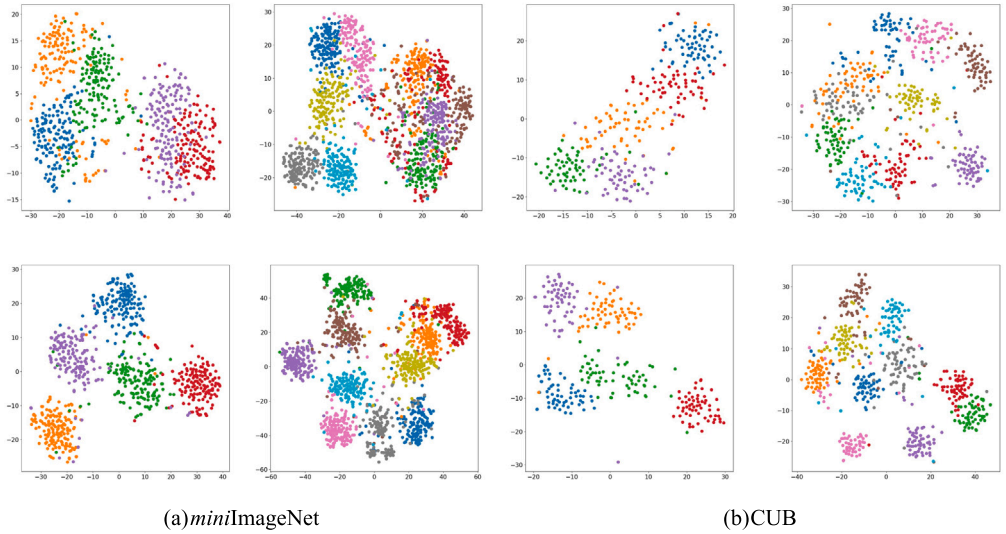

Fig. 7. Visualization using t-SNE [63]. The first row represents the basic features after global pooling by the backbone, while the second row represents the features extracted by our method. The first and second columns respectively represent features randomly selected 5/10 classes on the test sets of *mini*ImageNet and CUB.

the first column of class activation maps for birds, our model pays more attention to the entire bird rather than a specific part. In the first column of class activation maps for handguns, our model effectively attends to various components, whereas the baseline method overlooks the gunstock. These incorrect attentions lead to misclassifications by the original model when dealing with confusing images in the second column for each class, whereas our method correctly classifies the corresponding images.

Moreover, we demonstrate the effectiveness of our method by visualizing the feature distributions of test tasks. In Fig. 7, we showcase the t-SNE [63] visualization of our method compared to baseline methods on the 5-way 5-shot classification tasks sampled from the *mini*ImageNet and CUB datasets. We can observe that our proposed method, through explicit guidance on learning sample relationships, extracts features with better discriminability.

In Fig. 8, we demonstrate a visualization of the task-related similarity map predicted by the sample association information extraction module. This experiment follows "5-way 5-shot" setting and is conducted on the *mini*ImageNet dataset. In this ablation study, six images are randomly selected from each meta-task, and their category association scores are provided. The depth of colors in the figure represents the degree of category relevance. From Fig. 8, it can be observed that the method proposed in this paper correctly learns the category relationships of input samples, and the concept of "which samples belong to the same category" is accurately reflected in the visualized task-related similarity map.

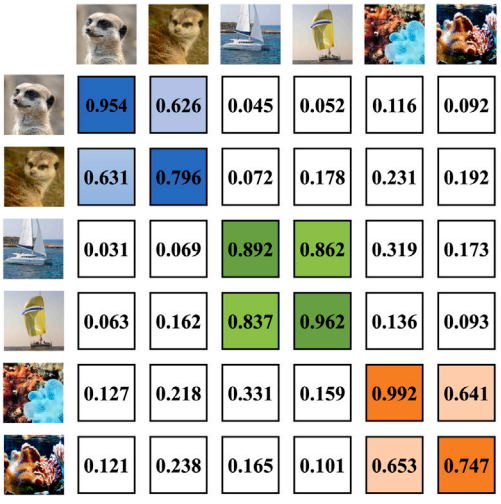

Fig. 8. Visualizations of category association scores for three randomly selected categories with two images on *mini*ImageNet. Darker colors in the visualization correspond to higher values of association, indicating stronger category correlations.

5. Conclusion

In this paper, we propose a Sample Relation Exploration (SRE) module for few-shot classification, which effectively aggregates features from the same class to learn discriminative class-related features. Our approach differs from existing methods by modeling relationships among all input images, rather than focusing on individual image representations. The SRE module comprises three key components: Sample-level Attention (SA) for exploring similarity relationships between samples, Explicit Guidance (EG) for supervising the learning of task-related similarity maps, and Channel-wise Adaptive Fusion (CAF) for dynamically fusing original and aggregated features. Extensive experiments demonstrate that our SRE method can learn robust and discriminative feature representations while significantly enhancing the performance of baseline models across different few-shot learning frameworks. The plug-and-play nature of our SRE module allows for flexible integration into both transfer learning and meta-learning based FSC frameworks. However, we acknowledge certain limitations of our approach. The effectiveness of SRE may vary depending on the specific dataset and backbone architecture used. Additionally, the optimal values for hyperparameters controlling the weights of explicit guidance loss and classification loss may require tuning for different scenarios. Future work could explore the adaptability of our method across a wider range of datasets and architectures, as well as investigate automatic hyperparameter optimization techniques.

CRediT authorship contribution statement

Xingye Chen: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wenxiao Wu:** Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Li Ma:** Writing – review & editing, Supervision, Investigation. **Xinge You:** Writing – review & editing, Supervision, Investigation. **Changxin Gao:** Writing – review & editing, Supervision, Investigation. **Nong Sang:** Writing – review & editing, Supervision, Investigation. **Yuanjie Shao:** Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition.

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.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (No. U22B2053) and the Knowledge Innovation Program of Wuhan-Shuguang, China (No. 2023010201020226).

Data availability

The data that support the findings of this study are available at the following URL: https://github.com/Chenguoz/SRE.

References

[1] B. Lake, R. Salakhutdinov, J. Gross, J. Tenenbaum, One shot learning of simple visual concepts, in: Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 33, (33) 2011.
[2] G. Koch, R. Zemel, R. Salakhutdinov, et al., Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop, Vol. 2, (1) Lille, 2015.
[3] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in: International Conference on Learning Representations, 2016.
[4] Y. Tian, Y. Wang, D. Krishnan, J.B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need? in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, 2020, pp. 266–282.
[5] Y. Bendou, Y. Hu, R. Lafargue, G. Lioi, B. Pasdeloup, S. Pateux, V. Gripon, Easy—ensemble augmented-shot-y-shaped learning: State-of-the-art few-shot classification with simple components, J. Imaging 8 (7) (2022) 179.
[6] 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, 2022, pp. 7972–7981.
[7] R. Das, Y.-X. Wang, J.M. Moura, On the importance of distractors for few-shot classification, in: IEEE International Conference on Computer Vision, 2021, pp. 9030–9040.
[8] H. Zhu, P. Koniusz, Transductive few-shot learning with prototype-based label propagation by iterative graph refinement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23996–24006.
[9] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Neural Information Processing Systems, Vol. 30, 2017.
[10] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, in: Neural Information Processing Systems, Vol. 32, 2019.
[11] D. Wertheimer, L. Tang, B. Hariharan, Few-shot classification with feature map reconstruction networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 8012–8021.
[12] J. Lai, S. Yang, J. Zhou, W. Wu, X. Chen, J. Liu, B.-B. Gao, C. Wang, Clustered-patch element connection for few-shot learning, 2023, arXiv preprint arXiv:2304.10093.
[13] Z. Hou, B. Yu, D. Tao, Batchformer: Learning to explore sample relationships for robust representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7256–7266.
[14] H. Wang, J. Shen, Y. Liu, Y. Gao, E. Gavves, Nformer: Robust person re-identification with neighbor transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7297–7307.
[15] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, in: Neural Information Processing Systems, Vol. 29, 2016.
[16] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[17] F. Zhou, P. Wang, L. Zhang, W. Wei, Y. Zhang, Revisiting prototypical network for cross domain few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 20061–20070.
[18] 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, 2020, pp. 8808–8817.
[19] X. Wang, S. Zhang, Z. Qing, Z. Zuo, C. Gao, R. Jin, N. Sang, HyRSM++: Hybrid relation guided temporal set matching for few-shot action recognition, Pattern Recognit. 147 (2024) 110110.
[20] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, N. Sang, CLIP-guided prototype modulating for few-shot action recognition, Int. J. Comput. Vis. 132 (6) (2024) 1899–1912.
[21] 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, 2019.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[23] R. Wang, H. Zheng, X. Duan, J. Liu, Y. Lu, T. Wang, S. Xu, B. Zhang, Few-shot learning with visual distribution calibration and cross-modal distribution alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23445–23454.
[24] L. Liu, W. Hamilton, G. Long, J. Jiang, H. Larochelle, A universal representation transformer layer for few-shot image classification, 2020, arXiv preprint arXiv:2006.11702.
[25] C. Simon, P. Koniusz, R. Nock, M. Harandi, Adaptive subspaces for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4136–4145.
[26] Y. Ma, S. Bai, S. An, W. Liu, A. Liu, X. Zhen, X. Liu, Transductive relation-propagation network for few-shot learning, in: IJCAI, Vol. 20, 2020, pp. 804–810.
[27] M. Boudiaf, E. Bennequin, M. Tami, A. Toubhans, P. Piantanida, C. Hudelot, I. Ben Ayed, Open-set likelihood maximization for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24007–24016.
[28] C. Chen, X. Yang, C. Xu, X. Huang, Z. Ma, Eckpn: Explicit class knowledge propagation network for transductive few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6596–6605.
[29] I. Ziko, J. Dolz, E. Granger, I.B. Ayed, Laplacian regularized few-shot learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 11660–11670.
[30] H. Zhu, P. Koniusz, EASE: Unsupervised discriminant subspace learning for transductive few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9078–9088.

[31] 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, 2020, pp. 12203–12213.
[32] D. Kang, H. Kwon, J. Min, M. Cho, Relational embedding for few-shot classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8822–8833.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[34] Y. Chen, Z. Liu, H. Xu, T. Darrell, X. Wang, Meta-baseline: Exploring simple meta-learning for few-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9062–9071.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[36] 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, 2018.
[37] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200–2011 dataset, 2011, California Institute of Technology.
[38] L. Bertinetto, J.F. Henriques, P.H. Torr, A. Vedaldi, Meta-learning with differentiable closed-form solvers, 2018, arXiv preprint arXiv:1805.08136.
[39] Y. Shao, W. Wu, X. You, C. Gao, N. Sang, Improving the generalization of MAML in few-shot classification via bi-level constraint, IEEE Trans. Circuits Syst. Video Technol. (2022).
[40] Z. Yang, J. Wang, Y. Zhu, Few-shot classification with contrastive learning, in: European Conference on Computer Vision, Springer, 2022, pp. 293–309.
[41] X. Huang, S.H. Choi, Sapenet: self-attention based prototype enhancement network for few-shot learning, Pattern Recognit. 135 (2023) 109170.
[42] B. Shi, W. Li, J. Huo, P. Zhu, L. Wang, Y. Gao, Global-and local-aware feature augmentation with semantic orthogonality for few-shot image classification, Pattern Recognit. 142 (2023) 109702.
[43] H.-J. Ye, D.-W. Zhou, L. Hong, Z. Li, X.-S. Wei, D.-C. Zhan, Contextualizing meta-learning via learning to decompose, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[44] J.Y. Lim, K.M. Lim, C.P. Lee, Y.X. Tan, SCL: Self-supervised contrastive learning for few-shot image classification, Neural Netw. 165 (2023) 19–30.
[45] Y. Lu, L. Wen, J. Liu, Y. Liu, X. Tian, Self-supervision can be a good few-shot learner, in: European Conference on Computer Vision, Springer, 2022, pp. 740–758.
[46] J. Ling, L. Liao, M. Yang, J. Shuai, Semi-supervised few-shot learning via multi-factor clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14564–14573.
[47] T. Munkhdalai, H. Yu, Meta networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 2554–2563.
[48] K. Lee, S. Maji, A. Ravichandran, S. Soatto, Meta-learning with differentiable convex optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10657–10665.
[49] X. Luo, L. Wei, L. Wen, J. Yang, L. Xie, Z. Xu, Q. Tian, Rectifying the shortcut learning of background for few-shot learning, Adv. Neural Inf. Process. Syst. 34 (2021) 13073–13085.
[50] S.M. Kye, H.B. Lee, H. Kim, S.J. Hwang, Meta-learned confidence for few-shot learning, 2020, arXiv preprint arXiv:2002.12017.
[51] G. Qi, H. Yu, Z. Lu, S. Li, Transductive few-shot classification on the oblique manifold, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8412–8422.
[52] W. Chen, Z. Zhang, W. Wang, L. Wang, Z. Wang, T. Tan, Few-shot learning with unsupervised part discovery and part-aligned similarity, Pattern Recognit. 133 (2023) 108986.
[53] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International Conference on Machine Learning, 2017, pp. 1126–1135.
[54] W. Li, L. Wang, J. Huo, Y. Shi, Y. Gao, J. Luo, Asymmetric distribution measure for few-shot learning, in: International Joint Conference on Artificial Intelligence, 2021, pp. 2957–2963.
[55] H. Cheng, S. Yang, J.T. Zhou, L. Guo, B. Wen, Frequency guidance matters in few-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11814–11824.
[56] Q. Liu, W. Cao, Z. He, Cycle optimization metric learning for few-shot classification, Pattern Recognit. 139 (2023) 109468.
[57] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, M. Cord, Boosting few-shot visual learning with self-supervision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8059–8068.
[58] P. Mangla, N. Kumari, A. Sinha, M. Singh, B. Krishnamurthy, V.N. Balasubramanian, Charting the right manifold: Manifold mixup for few-shot learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2218–2227.
[59] J. He, A. Kortylewski, A. Yuille, CORL: Compositional representation learning for few-shot classification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3890–3899.
[60] 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 Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[61] J. Liu, L. Song, Y. Qin, Prototype rectification for few-shot learning, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, Springer, 2020, pp. 741–756.
[62] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018, pp. 839–847.
[63] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

Xingye Chen received the B.E. degree in School of Computer Science, China University of Geosciences, Wuhan, China, in 2023. He is currently pursuing the M.E. degree with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. His research interests include deep learning and computer vision.

Wenxiao Wu received the B.E. degree in School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, in 2021. He is currently pursuing the M.E. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests include deep learning and computer vision.

Li Ma (Member, IEEE) received the B.S. and M.S. degrees from Shandong University, Jinan, China, in 2004 and 2006, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2011. From 2008 to 2010, she was a Visiting Scholar at Purdue University, West Lafayette, IN, USA. She also visited Mississippi State University, Starkville, MS, USA, for five months in 2018. She is currently an Associate Professor with the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan. Her research interests include hyperspectral data analysis, pattern recognition, and remote sensing applications.

Xinge You is currently a Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan. He received the B.S. and M.S. degrees in mathematics from Hubei University, Wuhan, China, in 1990 and 2000, respectively, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2004. His current research interests include image processing, wavelet analysis and its applications, pattern recognition, machine earning, and computer vision.

Changxin Gao received the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 2010. He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests are pattern recognition and surveillance video analysis.

Nong Sang received the B.E. degree in computer science and engineering, the M.S. degree in pattern recognition and intelligent control, and the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 1990, 1993, and 2000, respectively. He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests include object detection, object tracking, image/video segmentation, and analysis of surveillance videos.

Yuanjie Shao received the B.S. and M.S degree in college of mechanical and electronic information, China University of Geosciences in 2010 and 2013, Wuhan, China, and Ph.D. degree in Control science and Engineering from Huazhong University of Science and Technology in 2018. He is currently a lecturer with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. His research interests include pattern recognition, computer vision.