Attention-Based Multi-View Feature Collaboration for Decoupled Few-Shot Learning

Shuai Shao, Lei Xing, Yanjiang Wang, Baodi Liu, *Member, IEEE*, Weifeng Liu, *Senior Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— Decoupled Few-shot learning (FSL) is an effective methodology that deals with the problem of data-scarce. Its standard paradigm includes two phases: (1) Pre-train. Generating a CNN-based feature extraction model (FEM) via base data. (2) Meta-test. Employing the frozen FEM to obtain the novel data features, then classifying them. Obviously, one crucial factor, the category gap, prevents the development of FSL, i.e., it is challenging for the pre-trained FEM to adapt to the novel class flawlessly. Inspired by a common-sense theory: the FEMs based on different strategies focus on different priorities, we attempt to address this problem from the multi-view feature collaboration (MVFC) perspective. Specifically, we first denoise the multi-view features by subspace learning method, then design three attention blocks (loss-attention block, self-attention block and graph-attention block) to balance the representation between different views. The proposed method is evaluated on four benchmark datasets and achieves significant improvements of 0.9%-5.6% compared with SOTAs.**

***Index Terms*— Decoupled few-shot learning, feature extraction model, loss-attention block, self-attention block.**

I. Introduction

RECENTLY, machine learning has achieved satisfactory results and shown great potential in addressing computer vision tasks, whether in the field of person re-identification [1], [2], [3], or image classification [4], [5], [6], [7], [8], image segmentation [9], [10], [11], [12], [13] it has reached or even exceeded the level of human beings. Its success attributes to many factors, and the most indispensable one is a general assumption: we have enough labeled training data. Without it, the performance will drastically decline. However, this assumption often goes against the real application scenario, that is, collecting a large number of labeled samples is time-consuming or even impossible. Therefore, few-shot learning (FSL) [14], [15], [16], [17], [18], [19], [20] and zero-shot learning (ZSL) [21], [22], [23], [24], [25], [26], [27] as the pioneer methods to address the lack of labeled samples for each category have aroused widespread concerns. In this paper, we mainly focus on the few-shot learning.

In popular FSL-based classification tasks, the framework consists of two phases: (1) Pre-train. Using the base data to train a convolutional neural network (CNN) based feature extraction model (FEM). (2) Meta-test. Applying the FEM to extract the feature of novel data (with extremely limited labeled samples), then designing a classifier for recognizing them. It is worth noting that novel data categories are entirely different from the base data.

Through the above description, we conclude that one of the crucial problems preventing the development of FSL is the category gap between base and novel data. It seems that fine-tuning the FEM in the meta-test phase will solve the problem. But in fact, due to the scarcity of labeled data, fine-tuning can negatively affect the results (demonstrated in [28] and [29]). Therefore instead of the fine-tuning strategy, researchers prefer decoupling the FSL framework, that is, freezing the parameters of FEM after pre-training and directly using it to extract novel features in the meta-test phase. How to improve the cross-category representation ability in decoupled FSL (under decoupling constraints) is an urgent problem that needs to be solved at present.

Recent efforts on addressing this challenge focus on constructing a more robust and adaptive FEM that can directly generate better features for novel data, including self-supervision based FEMs [30], [31]; knowledge distillation based FEMs [32], [33]; meta-learning based FEMs [34], [35], etc. While the approaches mentioned above, merely weaken the negative influence to a certain extent, and the FSL community requires some dedicated methodologies for this specific problem.

We believe that: different feature extraction models (FEMs) possess distinct feature priorities, leading to that the feature distribution of the same data under different views has a certain

Manuscript received 28 July 2022; revised 16 October 2022 and 25 October 2022; accepted 16 November 2022. Date of publication 21 November 2022; date of current version 5 May 2023. This work was supported by the National Natural Science Foundation of China under Grant 62072468; in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2019MF073; in part by the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China), under Grant 20CX05001A; in part by the Major Scientific and Technological Projects of CNPC under Grant ZD2019-183-008; in part by the Creative Research Team of Young Scholars at Universities in Shandong Province under Grant 2019KJN019; and in part by the State Key Laboratory of Shale Oil and Gas Enrichment Mechanisms and Effective Development under Grant 33550000-22-ZC0613-0243. This article was recommended by Associates Editor Y. Wu. *(Shuai Shao and Lei Xing are co-first authors.) (Corresponding authors: Yanjiang Wang; Baodi Liu.)*

Shuai Shao is with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China, and also with the Zhejiang Laboratory, Research Center for Applied Mathematics and Machine Intelligence, Research Institute of Basic Theories, Hangzhou 311100, China.

Lei Xing is with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China.

Yanjiang Wang, Baodi Liu, and Weifeng Liu are with the College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China (e-mail: yjwang@upc.edu.cn; thu.liubaodi@gmail.com).

Yicong Zhou is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2022.3224003.

Digital Object Identifier 10.1109/TCSVT.2022.3224003

TABLE I
SOME IMPORTANT ABBREVIATIONS AND NOTATIONS

Abbreviation and Notation	Definition
FSL	few-shot learning
FEM	feature extraction model
LA-MVFC	loss-attention based multi-view feature collaboration
SA-MVFC	self-attention based multi-view feature collaboration
$\mathcal{D}_{base}$, $\mathcal{D}_{novel}$	base data, novel data
$\mathcal{S}$, $\mathcal{Q}$, $\mathcal{U}$	support set, query set, unlabeled set
$\mathcal{M}_{\theta}^{(v)}(\cdot)$	CNN-based FEM on the v_{th} view
$\mathbf{X}^{(v)}$, $\mathbf{x}_{ts}^{(v)}$	features of training and testing data on the v_{th} view
$\mathbf{X}_s^{(v)}$, $\mathbf{X}_u^{(v)}$, $\mathbf{X}_q^{(v)}$	features of support, unlabeled, query data on the v_{th} view
$\mathcal{J}(\cdot)$	conventional subspace learning method
$\mathbf{P}^{(v)}$	aligned features of training data on the v_{th} view
$\mathbf{P}_{sa}^{(v)}$	self-attention feature matrix on the v_{th} view
$\mathbf{P}_s^{(v)}$, $\mathbf{P}_u^{(v)}$, $\mathbf{P}_q^{(v)}$	aligned features of support, unlabeled, query data on the v_{th} view
$\mathbf{W}^{(v)}$	classifier on the v_{th} view
$\mathbf{W}_z$	final collaborative classifier
$\mathbf{Y}$	label matrices of training data
$\mathbf{\Omega} = [\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(V)}]^T$	loss-attention weights
$\mathbf{T}$	self-attention weight matrix
$\mathbf{Z}$, $\mathbf{z}_{ts}$	collaboration features of training and testing data

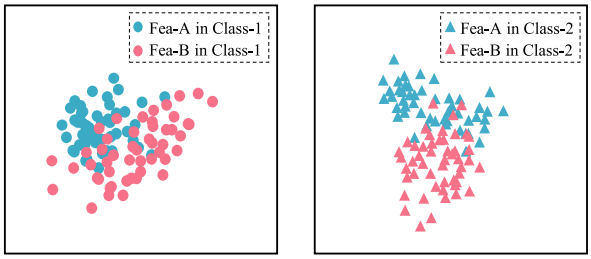


Fig. 1. Feature distribution of the same data under different views.

deviation, as shown in Figure 1. It's not hard to infer that we can obtain a better representation if we conduct multiple views of features collaboration. An example is illustrated in Figure 3. However, two main obstacles hinder the idea of multi-view feature collaboration (MVFC): (1) Considering that the multi-view features extracted from different independent FEMs are in separated spaces, they have noises when fusing these features. Therefore, the first obstacle is how to find an appropriate way to denoise the multi-view features. (2) During the fusion process, the importance of different features will be different. Therefore, the second obstacle lies in how to reasonably assign effective joint weights to multi-view features to maximize their representation capabilities.

To alleviate the first obstacle, we introduce a subspace learning strategy, which can transform the initial multi-view features into an integrated space to reconstruct the denoised and low-dimensional representation (see Section IV-B). To address the second obstacle, we design three different attention blocks for each view of the feature, which could automatically update the weights of the combination. To be more specific, the first one is the Loss-Attention block, which uses each view's objective function loss to achieve the corresponding weights (see Section IV-C.1), and we dub the Loss-Attention based Multi-View Feature Collaboration as LA-MVFC; The second one is the Self-Attention block, referring to [36] to get the weights through finding the views' relations (see Section IV-C.2), and the corresponding Self-Attention based Multi-View Feature Collaboration is called SA-MVFC. And the third one is the Graph-Attention block, inspired by [37], we regard the different views as different connected nodes and compute their weights (see Section IV-C.3), which is called Graph-Attention based Multi-View Feature Collaboration as GA-MVFC; The flowchart is illustrated in Figure 2.

Additionally, in the classifier designing process, researchers classify the FSL-based algorithms to two categories according to the adoption of data: (1) supervised few-shot learning and (2) semi-supervised few-shot learning. This paper extends the proposed method to the two settings (see Section IV-E). For convenience, we list some critical abbreviations and notations in Table I.

A. Contributions

We summarize our contributions as: (1) We propose three kinds of multi-view feature collaboration methods for solving the cross-category challenge in decoupled few-shot learning (FSL). They are separately called Loss-Attention based Multi-View Feature Collaboration (LA-MVFC),Self-Attention based Multi-View Feature Collaboration (SA-MVFC), and Graph-Attention based Multi-View Feature Collaboration (GA-MVFC).

(2) Compared with traditional FSL methods that focus on tuning the network, our proposed strategy is more straightforward and effective, which can directly fuse multi-view features extracted from the existing FEMs (for more details, please see Section IV-F). Moreover, benefiting from the robust feature representation, this method has shown satisfactory performance in dealing with extremely few sample situations. Therefore, we consider that this paper has important practical significance.

(3) The proposed method is evaluated on four benchmark datasets, including mini-ImageNet, tiered-ImageNet, CIFAR-FS, and FC100. Compared with other SOTAs, our methods achieve significant 0.9%-5.6% improvements.

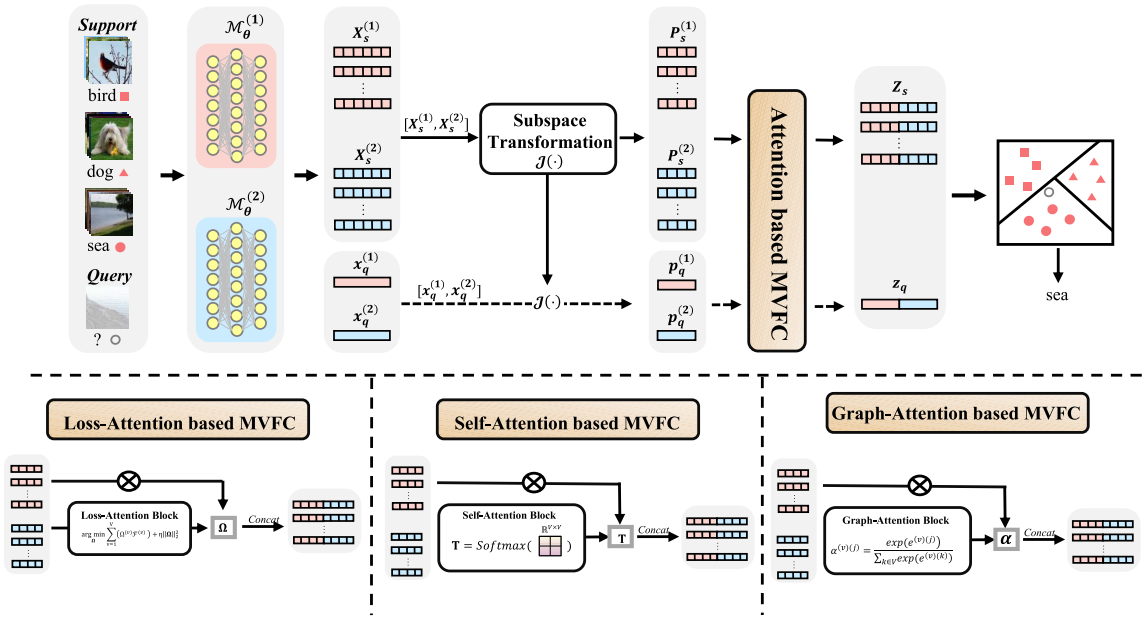


Fig. 2. The structure of Multi-View Feature Collaboration (MVFC). There are two views of feature extraction models (FEMs), *i.e.*, $\mathcal{M}_\theta^{(v)}$, where $v = [1, 2]$ indicates the v_{th} view. The various features correspond to different colors. There are 4 steps in total. **(1)** Inputting images to FEMs and obtain the support features $\mathbf{X}_s^{(v)}$ and query feature $\mathbf{x}_q^{(v)}$. **(2)** Transferring multi-view features to a unified space to obtain the aligned and low-dimensional features $\mathbf{P}_s^{(v)}$, $\mathbf{p}_q^{(v)}$. **(3)** Completing feature collaboration through different attention mechanisms. We dub the Loss-Attention based Multi-View Feature Collaboration as **LA-MVFC**, the Self-Attention based Multi-View Feature Collaboration as **SA-MVFC**, and the Graph-Attention based Multi-View Feature Collaboration as **GA-MVFC**. **(4)** Constructing a classifier by the collaborative support features and recognizing query samples. For more details, please see Section IV.

B. Extensions

A preliminary version of this work was published in the 29^{th} ACM International Conference on Multimedia (ACMMM) [38] in 2021. This paper extends the conference version as follows:

(1) In the theoretical perspective, [38] only proposed the LA-MVFC, while this article provides SA-MVFC and GA-MVFC, which provides more opportunities for applying this *multi-view feature collaboration* thought to reality.

(2) From the experimental perspective, this article supplements a large number of experiments. Specifically, besides comparing with more SOTAs, we decompose the complete job into 5 different components (baseline, self-train, self-supervision, denoising, attention block) and carefully analyze their efficiencies.

II. Related Work

A. Few-Shot Learning

To solve the problem of insufficient samples, FSL [14], [15], [16], [17], [18], [19], [20] and ZSL [21], [22], [23], [24], [25], [26], [27] paradigms are proposed. The goal of the FSL hopes that the machine can achieve effective computer vision tasks when we only have a small number of labeled samples. The ZSL is a more difficult task than the FSL, and hopes that the machine can achieve the above target without labeled samples. This paper only focuses on the FSL paradigm. There are various classical algorithms proposed to deal with this problem. We introduce the two most representative methods. The first is the meta-learning based methods, which focus on obtaining a unified model that rapidly adapts to new tasks. The several popular methods are as follow, [18], [19], [34], [39], [40], and [41]. The second is Metric learning-based methods, which concentrate on searching an ideal distance metrics to enhance the robustness of the model, containing [35], [42], [43] et al. Furthermore, these approaches can be categorized according to other classification criteria, i.e., supervised FSL, and semi-supervised FSL. For example, [30], [34], [43] et al. follow the supervised setting; and LST [31], [44], [45], [46] et al. are explored in the semi-supervised setting.

B. Multi-View Few-Shot Learning

As there are two sides to every coin, it is boundedness to define objects from a single point of view. Multi-view learning as an effective strategy has attracted extensive attention in the past decade. In FSL, some similar methods have been proposed, such as: DenseCls [47] divide the feature map into various blocks, and predict the corresponding labels; MDFM [5] integrates multi-view classifiers and comprehensively considers the final decisions; DivCoop [28] trains the FEMs on various datasets and integrates them into a multi-domain representation; URT [48] is an improved method compared with DivCoop [28], which proposes a transformer layer to help the network employ various datasets; DWC [29] introduces a cooperate strategy on a designed ensemble model to integrate multiple information. Although the above-mentioned approaches are based on multi-view learning, they are limited by the fused FEMs and classifiers. In other words, these methods lose scalability, but ours is not. This article conducts multi-view learning from the perspective of features, aiming to obtain better representations for solving the cross-category problem in decoupled FSL.

C. Subspace Learning

Subspace learning is a transfer strategy that can transform the samples into the other expression. It is an effective dimensionality reduction methodology. Here, we illustrate several classical subspace learning approaches employed in our method. The first is the Locally Linear Embedding (LLE) [49]. It can maintain local distance by referring to searching the projection of data in low-dimensional. The next is the Laplacian Eigenmap (LE) [50], which employs the spectral decomposition of the graph Laplacian method to map the initial data to a low-dimensional portrayal. The last is the Principal Component Analysis (PCA) [51]. It reduces the initial features to lower-dimensional space through singular value decomposition. All of the above-mentioned methods contribute to our approach, and we will illustrate the experimental results in Section V-E.3.

III. Problem Formulation

This section introduces our procedure thoroughly. It contains two stages which are pre-train and meta-test. **(1)** In pre-train stage, we define the base data as $\mathcal{D}_{base} = \{(x_{(i)}, y_{(i)}) | \, y_{(i)} \in \mathcal{C}_{base}\}_{i=1}^{N_{base}}$, where x represents the sample and y denotes its label. N_{base} indicates the number of base data. $\mathcal{C}_{base}$ denotes the base category set. We train the CNN-based FEM $\mathcal{M}_\theta(\cdot)$ on $\mathcal{D}_{base}$, where θ illustrates the parameters in CNN. We employ various FEMs to extract features from different perspectives in this paper, and we define the FEM on the v_{th} view as $\mathcal{M}_\theta^{(v)}(\cdot)$, where $v = 1, 2, \cdots, V$. **(2)** In the meta-test stage, we define the novel data as $\mathcal{D}_{novel} = \{(x_{(j)}, y_{(j)}) | \, y_{(j)} \in \mathcal{C}_{novel}\}_{j=1}^{N_{novel}}$, where $\mathcal{C}_{novel}$ indicates the novel category set, N_{novel} is the number of novel data. $C_{base} \cap C_{novel} = \emptyset$. $\mathcal{D}_{novel}$ consists of three components, e.g., $\mathcal{D}_{novel} = \{\mathcal{S}, \mathcal{U}, \mathcal{Q}\}$, where $\mathcal{S}, \mathcal{U}$ and $\mathcal{Q}$ illustrate support, unlabeled and query sets, repectively. $\mathcal{S} \cap \mathcal{U} = \emptyset$, $\mathcal{S} \cap \mathcal{Q} = \emptyset$, $\mathcal{Q} \cap \mathcal{U} = \emptyset$. In this process, we freeze the pre-trained FEM and use it to extract the feature of $\mathcal{D}_{novel}$. After that, we design the classifier to classify $\mathcal{Q}$.

In our paper, according to whether using the unlabeled data, we split the FSL into two settings: supervised setting and semi-supervised setting. Specifically, defining the feature of $\mathcal{D}_{novel}$ on the v_{th} view as $\mathbf{X}_{novel}^{(v)} = [\mathbf{X}_s^{(v)}, \mathbf{X}_u^{(v)}, \mathbf{X}_q^{(v)}]$, where $\mathbf{X}_s^{(v)} = \mathcal{M}_\theta^{(v)}(\mathcal{S})$, $\mathbf{X}_u^{(v)} = \mathcal{M}_\theta^{(v)}(\mathcal{U})$, and $\mathbf{X}_q^{(v)} = \mathcal{M}_\theta^{(v)}(\mathcal{Q})$ represent the features of support, unlabeled, and query data on the v_{th} view, respectively. Besides, we follow the standared *C-way-N-shot* per episode as [45] for classification task, where *C-way* indicates C classes, and *N-shot* denotes N samples per class.

IV. Methodology

In this section, the linear regression classifier is briefly reviewed first. Then, we align the multi-view features and propose the Multi-View Feature Collaboration (MVFC) method. To balance the multi-view features, we design the Loss-Attention block and Self-Attention block. Next, we introduce how to use the re-constructed features to design classifiers and extend our method to different FSL settings. Finally, we introduce the employed multi-view FEMs. The complete flowchart is shown in Figure 2.

A. Review of Linear Regression Classifier

In decoupled FSL, researchers usually re-construct machine learning based classifiers in meta-test phase, *e.g.*, logistic regression, linear regression, support vector machine. Here, we employ the regularized linear regression method as the example to introduce our complete model in detail. Given labeled samples' feature matrix $\mathbf{X}$ and their one-hot label matrix $\mathbf{Y}$, the objective function can be formulated as:

$$\arg\min_{\mathbf{W}} \mathcal{F} = \|\mathbf{Y} - \mathbf{W}\mathbf{X}\|_F^2 + \mu \|\mathbf{W}\|_F^2 \tag{1}$$

where $\|\cdot\|_F$ is $(\cdot)$'s Frobenius-norm. μ represents the hyper-parameter. $\mathbf{X} \in \mathbb{R}^{dim1 \times N}$, $\mathbf{Y} \in \mathbb{R}^{C \times N}$; $dim1$ denotes labeled samples dimension, and N denotes the number of labeled samples. C indicates the number of categories. $\mathbf{W} \in \mathbb{R}^{C \times dim1}$ denotes the to-be-learned classifier. After simply optimization, we achieve the solution as:

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T + \mu \mathbf{I}\right)^{-1} \tag{2}$$

where $\mathbf{I}$ denotes the identity matrix. And now, given a testing sample embedding $\mathbf{x}_{ts} \in \mathbb{R}^{dim1}$, we can classify it by:

$$\mathcal{A}(\mathbf{x}_{ts}) = id_{max}\{\mathbf{W}\mathbf{x}_{ts}\} \tag{3}$$

where id_{max} represents an operator that obtains the index of the max value.

B. Multi-View Feature Denoising

This section attempts to denoise the to-be-fused featuresthrough the conventional subspace learning algorithms (indicated as $\mathcal{J}(\cdot)$), such as LE [50] PCA [51], LLE [49]. These strategies can transfer the initial features to a unified space with reconstructed low-dimensional representation. Specifically, assume that there are V views in total. Each view corresponds to one kind of feature $\mathbf{X}^{(v)}$, where $v = 1, 2, \cdots, V$. We treat *one-sample's-V-views-features* as *V-samples'-features*, and represent the features of the expanded dataset as $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(V)}] \in \mathbb{R}^{dim1 \times (N \times V)}$. We design subspace learning operation as $\mathcal{J}(\mathbf{X})$ and obtain the novel features $\mathbf{P} = [\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \cdots, \mathbf{P}^{(V)}] \in \mathbb{R}^{dim2 \times (N \times V)}$, where $\mathbf{P}^{(v)} \in \mathbb{R}^{dim2 \times N}$ denotes the feature on the v_{th} view after feature denoising. $dim2$ denotes the novel dimension.

C. Multi-View Feature Collaboration

To balance the representation between different views, we design the Loss-Attention block, the Self-Attention block and the Graph-Attention block, to automatically update combination weights and encourage these features to have different effects on the final decision. An algorithm is conducted in Algorithm 1.

1) Loss-Attention Based Multi-View Feature Collaboration: We first introduce our Loss-Attention based Multi-View Feature Collaboration (LA-MVFC) method. This strategy computes the combination weights through each view's loss function. Specifically, defining the combination weights as $\mathbf{\Omega} = [\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(V)}]^T$, where $\mathbf{\Omega}$ denotes a weight vector, $\Omega^{(v)} (v = 1, 2, \cdots, V)$ is the v_{th} element in $\mathbf{\Omega}$. Our loss-attention block includes 3 steps:

Algorithm 1 Attention-Based Multi-View Feature Collaboration

Input: Base set $\mathcal{D}_{base}$, Novel set $\mathcal{D}_{novel}$
Output: Query label

```
Designing the multi-view feature extraction model M_θ^v(·) through D_base, and obtaining novel data' embedding by M_θ^v(D_novel).
Transforming the novel data's embedding to an unified space and obtaining aligned features.
if Loss-Attention then
    Training a basic classifier W^(v) by Equation (4).
    Calculating the objective function's loss by Equation (5).
    Computing the combination weights by Equation (6),(7),(8).
    Obtaining the collaborative feature by Equation (9).
else if Self-Attention then
    Computing the self-attention weight matrix by Equation (10).
    Getting self-attention feature matrix by Equation (11).
    Obtaining the collaborative feature by Equation (12).
else if Graph-Attention then
    Taking the v_th view as the central node, and calculate its weight with other views by Equation (13).
    Achieving the unified weight for each view by Equation (14).
    Obtaining the collaborative feature by Equation (15).
```

(i) Employing the aligned feature $\mathbf{P}^v$ to substitute $\mathbf{X}^v$ and obtain a classifier $\mathbf{W}^{(v)} \in \mathbb{R}^{C\times dim2}$ by Equation (2), where $\mathbf{W}^{(v)}$ represents the trained novel classifier on the v_{th} view:

$$\mathbf{W}^{(v)} = \mathbf{Y}\mathbf{P}^{(v)^T}\left(\mathbf{P}^{(v)}\mathbf{P}^{(v)^T} + \mu\mathbf{I}\right)^{-1} \tag{4}$$

(ii) Using $\mathbf{P}^{(v)}$ and $\mathbf{W}^{(v)}$ to re-calculate the loss of objective function on the v_{th} view $\mathcal{F}^{(v)}$ by Equation (1):

$$\mathcal{F}^{(v)} = \left\|\mathbf{Y} - \mathbf{W}^{(v)}\mathbf{P}^{(v)}\right\|_F^2 + \mu\left\|\mathbf{W}^{(v)}\right\|_F^2 \tag{5}$$

(iii) Exploiting the $\mathcal{F}^{(v)}$ to compute the combination weights. The objective function is:

$$\begin{aligned}\arg\min_{\Omega}\mathcal{G} &= \sum_{v=1}^{V}\left(\Omega^{(v)}\mathcal{F}^{(v)}\right) + \eta\left\|\Omega\right\|_2^2 \\ \text{s.t. } &\sum_{v=1}^{V}\Omega^{(v)} = 1,\ \ \Omega^{(v)} \geq 0\end{aligned} \tag{6}$$

where $\Omega^{(v)}$ represents the weight of v_{th} view. $\|\cdot\|_2$ denotes $(\cdot)$'s ℓ_2-norm. η as the parameter. The Lagrangian is applied to deal with the problem, the Equation (6) can be rewritten as:

$$\begin{aligned}\arg\min_{\Omega,\zeta,\Lambda}\mathcal{G} = &\sum_{v=1}^{V}\left(\Omega^{(v)}\mathcal{F}^{(v)}\right) + \eta\left\|\Omega\right\|_2^2 \\ &-\zeta\left(\sum_{v=1}^{V}\Omega^{(v)} - 1\right) - \Lambda^T\Omega\end{aligned} \tag{7}$$

where ζ denotes a constant, $\Lambda = [\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(V)}]^T$ indicates a vector. Assuming that $\hat{\Omega}$, $\hat{\zeta}$, $\hat{\Lambda}$ are the optimal solutions, this problem can be solved as:

$$\hat{\Omega}^{(v)} = \frac{1}{2\eta}max\left\{\frac{\sum_{v=1}^{V}\mathcal{F}^{(v)}}{V} + \frac{2\eta}{V} - \mathcal{F}^{(v)} - \hat{\Lambda}_{avg}, 0\right\} \tag{8}$$

where $\hat{\Lambda}_{avg}$ is a constant, indicates the average of $\hat{\Lambda}$. For the detailed optimization procedure, please refer to **Appendix A**.

After completing the loss-attention process, we give multi-view features the combination weights and obtain the final collaborative feature $\mathbf{Z} = [\mathbf{z}_{(1)}, \mathbf{z}_{(2)}, \cdots, \mathbf{z}_{(N)}] \in \mathbb{R}^{dim3\times N}$ by:

$$\mathbf{Z} \leftarrow \mathbf{z}_{(n)} = Concat\left(\hat{\Omega}^{(1)}\mathbf{p}_{(n)}^{(1)}, \hat{\Omega}^{(2)}\mathbf{p}_{(n)}^{(2)}, \cdots, \hat{\Omega}^{(V)}\mathbf{p}_{(n)}^{(V)}\right) \tag{9}$$

where $\mathbf{p}_{(n)}^{(v)}, \mathbf{z}_{(n)}(n = 1, 2, \cdots, N)$ denote the n_{th} vector of $\mathbf{P}^{(v)}$ and $\mathbf{Z}$.

2) Self-Attention Based Multi-View Feature Collaboration: In this section, we introduce our Self-Attention based Multi-View Feature Collaboration (SA-MVFC). Different from loss-attention block, relying on the corresponding loss function to calculate the combination weights, this block introduces a self-attention mechanism [36] to obtain the weights through finding the views' relations. Notably, most of the methods use self-attention to capture the relations among different samples, but this paper employs it to reflect the relations among different views.

Specifically, the aligned novel feature is $\mathbf{P} = [\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \cdots, \mathbf{P}^{(V)}] \in \mathbb{R}^{dim2\times(N\times V)}$. We reshape it to $\mathbf{P}_{tmp} \in \mathbb{R}^{(dim2\times N)\times V}$. The complete self-attention block includes 2 steps:

(i) We compute the view self-attention weight matrix $\mathbf{T} \in \mathbb{R}^{V\times V}$ by:

$$\mathbf{T} = softmax\left(\mathbf{P}_{tmp}{}^T\mathbf{P}_{tmp}\right) \tag{10}$$

where $softmax$ is the operation to compute the probability.

(ii) Following, we can get the self-attention feature matrix $\mathbf{P}_{sa} \in \mathbb{R}^{(dim2\times N)\times V}$ by:

$$\mathbf{P}_{sa} = \mathbf{P}_{tmp}\mathbf{T} \tag{11}$$

where $\mathbf{P}_{sa} = [\mathbf{P}_{sa}^1, \mathbf{P}_{sa}^2, \cdots, \mathbf{P}_{sa}^V]$.

After that, we concatenate the self-attention feature matrix and obtain the final collaborative feature $\mathbf{Z} = [\mathbf{z}_{(1)}, \mathbf{z}_{(2)}, \cdots, \mathbf{z}_{(N)}] \in \mathbb{R}^{dim3\times N}$ by:

$$\mathbf{Z} \leftarrow \mathbf{z}_{(n)} = Concat\left(\mathbf{p}_{sa}{}_{(n)}^{(1)}, \mathbf{p}_{sa}{}_{(n)}^{(2)}, \cdots, \mathbf{p}_{sa}{}_{(n)}^{(V)}\right) \tag{12}$$

where $\mathbf{p}_{sa}{}_{(n)}^{(v)}, \mathbf{z}_{(n)}(n = 1, 2, \cdots, N)$ denote the n_{th} vector of $\mathbf{P}^{(v)}$ and $\mathbf{Z}$.

3) Graph-Attention Based Multi-View Feature Collaboration: Furthermore, inspired by [37], we introduce a novel Graph-Attention based Multi-View Feature Collaboration (GA-MVFC) method. The classical graph-attention aims to compute the weights among different connected nodes. In our task, the node corresponds to the view, and we purpose to focus on the relations among different views. According to reshape the aligned novel feature, we obtain the $\mathbf{P}_{tmp} = [\mathbf{p}_{tmp}^{(1)}, \mathbf{p}_{tmp}^{(2)}, \cdots, \mathbf{p}_{tmp}^{(v)}, \cdots, \mathbf{p}_{tmp}^{(V)}] \in \mathbb{R}^{(dim2\times N)\times V}$,

where $\mathbf{p}_{tmp}^{(v)} \in \mathbb{R}^{(dim2\times N)\times 1}$. Then we formulate the graph-attention block, which includes 3 steps:

(i) We take the v_{th} view as the central node, and calculate its weight with other views as:

$$\alpha^{(v)(j)} = \frac{exp\left(e^{(v)(j)}\right)}{\sum_{k\in V} exp\left(e^{(v)(k)}\right)} \tag{13}$$

where $\alpha^{(v)(j)}$ denotes the weight between v_{th} view and j_{th} view, $j = 1, 2\cdots, v, \cdots, V$; $e^{(v)(j)} = Sim(\mathbf{p}_{tmp}^{(v)}, \mathbf{p}_{tmp}^{(j)})$ denotes the similarity between v_{th} view and j_{th} view, here we use the cosine similarity.

(ii) Notably, taking any view as the central node, we can get the corresponding weights. In this paper, we average them to get an unified weight for each view, which can be formulated as:

$$\alpha^{(v)} = \frac{1}{V}\sum_{j=1}^{V} \alpha^{(v)(j)} \tag{14}$$

(iii) After that, we assign the weights to the corresponding features and fuse them by:

$$\mathbf{Z} \leftarrow \mathbf{z}_{(n)} = Concat\left(\alpha^{(1)}\mathbf{p}_{(n)}^{(1)}, \alpha^{(2)}\mathbf{p}_{(n)}^{(2)}, \cdots, \alpha^{(V)}\mathbf{p}_{(n)}^{(V)}\right) \tag{15}$$

where $\mathbf{p}_{(n)}^{(v)}, \mathbf{z}_{(n)} (n = 1, 2, \cdots, N)$ denote the n_{th} vector of $\mathbf{P}^{(v)}$ and $\mathbf{Z}$.

4) Discussion About Different Attention Mechanisms: All the three attention mechanisms have their cons and pros. To be more specific: (1) The loss-attention is based on the loss obtained by the classifier, and its consideration is more about the coordination between the feature and the classifier. It is more purposeful compared with other attention strategies and can obtain more accurate classification accuracy, which is suitable for the case of known features and classifiers.

(2) The self-attention is more concerned with the internal connection among features. It is more universal and only depends on features. Although it loses a certain accuracy compared to the loss-attention, it improves flexibility and reduces computational consumption.

(3) The graph-attention has the same insight as the self-attention, i.e., also pays attention on the feature's relations. The difference between them is that: graph-attention will first send diverse features into the graph space for encoding, so it can represent a higher-order relationship among samples. This good property makes the graph-attention strategy suitable for cases where the features have high similarity.

D. Classification

After obtaining the collaborative feature $\mathbf{Z}$ through Equation (9) or (12), we utilize it to replace $\mathbf{X}$ and obtain the final collaborative classifier $\mathbf{W}_z \in \mathbb{R}^{C\times dim3}$:

$$\mathbf{W}_z = \mathbf{Y}\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T + \mu\mathbf{I}\right)^{-1} \tag{16}$$

Finally, given a testing sample feature $\mathbf{x}_{ts}^{(v)}, (v = 1, 2, \cdots, V)$. Through the above descriptions, we can obtain the collaborative feature $\mathbf{z}_{ts} \in \mathbb{R}^{dim3}$, and then predict its label by:

$$\mathcal{A}(\mathbf{z}_{ts}) = max\left\{\mathbf{W}_z\mathbf{z}_{ts}\right\} \tag{17}$$

E. Multi-View Feature Collaboration for Few-Shot Learning

The feature of $\left(\mathbf{X}_{novel}^{(v)}\right)$ after denoising can be formulated as $\mathbf{P}_{novel}^{(v)} = [\mathbf{P}_s^{(v)}, \mathbf{P}_u^{(v)}, \mathbf{P}_q^{(v)}]$. Compared with traditional classification task, the *support data* equals to the *labeled training data*, the *unlabeled data* equals to the *unlabeled training data*, and the *query data* equals to the *testing data*. Depending on whether to use unlabeled samples when constructing the classifier, the FSL can be split into supervised setting (not use) and unsupervised setting (use). Notably, in our paper, we assume that the *query data*'s feature is given in advance, *i.e.*, our supervised setting is transductive.

In order to design more robust classifier, we introduce a standard self-training strategy [52]. In semi-supervised case, (1) we first design classifier by Equation (16); (2) then predict the category of *unlabeled data* by Equation (17); (3) next select one most confidence sample from the *unlabeled data* according to the prediction and add it to the *support data* without putting back; (4) finally repeat the process until the classifier is stable and use it classify the *query data*. In supervised case, we just need to replace the *unlabeled data* with *query data* in (1)(2)(3) steps, and finally classify the *query data*.

F. Multi-View Feature Extraction Model

All adopted multi-view features come from different, existed FEMs. As examples: (1) Standard-feature (Std-Fea) [45]. The FEM is based on a classical CNN-based classification structure. (2) Meta-feature (Meta-Fea), which is similar to [35], integrates the meta-learning strategy to the algorithm. (3) Self-supervised-feature (SS-Fea) [30]. The FEM introduces the auxiliary loss to cooperate with traditional CNN to enhance the network's robustness. The experimental results of all stacking ways are discussed in Section V-E.5.

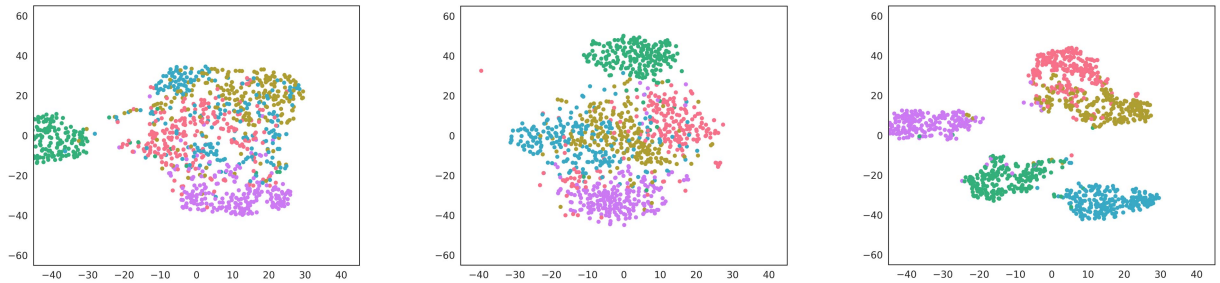
For convenience, we fuse two categories of SS-Feas for most experiments. For the first category, the FEM is designed by introducing standard classification loss $\mathcal{L}_c$ and auxiliary rotation loss $\mathcal{L}_r$. $\mathcal{L}_c$ can be formulated as:

$$\mathcal{L}_c = -\sum_c y_{(c,x)} log(p_{(c,x)}) \tag{18}$$

The probabilities of the the truth label is indicated as $y_{(c,x)}$, and predicted label is denoted as $p_{(c,x)}$, which represents the x_{th} sample belongs to the c_{th} class. Then, each sample is rotated to r degree and $r \in \mathcal{C}_{\mathcal{R}} = \{0°, 90°, 180°, 270°\}$. We define rotation loss as:

$$\mathcal{L}_r = -\sum_r y_{(r,x)} log(p_{(r,x)}) \tag{19}$$

where $y_{(r,x)}$ denotes probabilities of the truth label, and $p_{(r,x)}$ denotes the probabilities of predicted label, which represents the x_{th} sample belongs to the r_{th} class. Therefore, the first loss function is formulated as $\mathcal{L}_c + \mathcal{L}_r$, and the feature based on this



(a) SS-R-Fea (b) SS-M-Fea (c) Fuse-Fea

Fig. 3. t-SNE visualization of features on mini-ImageNet. SS-R-Fea and SS-M-Fea represent two categories of different single-view features (see Section IV-F). Fuse-Fea denotes the fusion feature with loss-attention.

kind of FEM is dubbed SS-R-Fea. We extract another category from self-supervised FEM as the second feature, named as SS-M-Fea. Specifically, in order to predict image mirrors, this FEM adds the loss $\mathcal{L}_c$ and auxiliary mirror loss $\mathcal{L}_m$ to the neural network. Assume that there are m ways and $m \in \mathcal{C}_{\mathcal{M}} = \{vertically, horizontally, diagonally\}$, we define the mirror loss as:

$$\mathcal{L}_m = -\sum_{m} y_{(m,x)} log(p_{(m,x)}) \tag{20}$$

where $y_{(m,x)}$ denotes the probabilities that the truth label, and $p_{(m,x)}$ illustrates the probabilities of the predicted label of x_{th} sample belongs to the m_{th} class. Next, the loss function can be summarized as $\mathcal{L}_c + \mathcal{L}_m$.

V. Experiments

A. Datasets

Our experiments are carried out on four benchmark datasets, including mini-ImageNet [70], tiered-ImageNet [68], CIFAR-FS [59], FC100 [43], and CUB [71]. mini-ImageNet and tiered-ImageNet are selected from the ImageNet dataset [72] and as the subsets. mini-ImageNet consists of 100 classes with 600 images per class, and tiered-ImageNet has 608 classes and each class contains $1,281$ images on average. Both of them resize the image to 84×84. Following the standard split way as [45], for mini-ImageNet, the base set contains 64 selected classes, the validation is composed of 16 classes, and the novel set includes 20 classes. Similarly, for tiered-ImageNet, the base set includes 351 classes, the validation set contains 97 classes, and 160 classes are prepared for the novel set. The CIFAR-FS and FC100 are the subsets of the CIFAR-100 dataset [73], which includes 100 classes. According to the split introduced in [59], CIFAR-FS is divided into 64 classes, and it can be seen as the base set, the validation set consists of 16 classes, and the novel set includes 20 classes. And for FC100, we divided it into 60 classes as the base set, the validation set contains 20 classes, and the novel set includes 20 classes. The image size of CIFAR-FS and FC100 datasets are set to 32×32.

B. Implementation Details

All the FEMs on various views in this paper utilize ResNet12 as the backbone network. ResNet12 contains four residual blocks, four 2×2 max-pooling layers, and four dropout layers. The optimizer is stochastic gradient descent with Nesterov momentum (0.9). For the parameter η in Equation (6), we define it to 1.4 for convenience and discuss more choices in Figure 5. The training epoch is 120, and we evaluate 600 episodes with 15 query samples per class for all the models. Additionally, all the selected subspace learning approaches refer to the scikit-learn [74] default implementation. This paper excludes the fine-tuning process for the novel data classification task. Other experimental settings follow the ICI [45].

C. Experimental Results

We compare our proposed LA-MVFC, SA-MVFC, and GA-MVFC (only fuse SS-R-Fea and SS-M-Fea) with several SOTAs. The supervised comparison results are listed in Table II, III. And the semi-supervised comparison results are reported in Table IV. Here, we list several observations.

(1) First, we discuss the supervised results from Table II, III. Obviously, our LA-MVFC and SA-MVFC have far surpassed other approaches. To be more specific, in mini-ImageNet, our methods can exceed others at least 2.6% on 5-way 1-shot case, 2.2% on 5-way 5-shot case; in tiered-ImageNet, our methods can exceed others at least 3.6% on 5-way 1-shot case, 0.9% on 5-way 5-shot case; in CIFAR-FS, our methods can exceed others at least 5.6% on 5-way 1-shot case, 3.1% on 5-way 5-shot case; in FC100, our methods can exceed others at least 3.9% on 5-way 1-shot case, 5.0% on 5-way 5-shot case.

(2) Then, we observe the semi-supervised results from Table IV. Our reported results are based on 100 unlabeled samples. About the influence of the number of unlabeled samples, please refer to Figure 6. Specifically, in mini-ImageNet, our methods can exceed others at least 4.4% on 5-way 1-shot case, 3.6% on 5-way 5-shot case; in tiered-ImageNet, our methods can exceed others at least 2.1% on 5-way 1-shot case, 2.8% on 5-way 5-shot case.

(3) Next, we compare our LA-MVFC, SA-MVFC, and GA-MVFC with other advanced multi-view based methods, containing DenseCls [47], DWC [29], DivCoop [28], URT [48]. In mini-ImageNet, our methods can outperform others at least 2.6% on 5-way 1-shot case, 2.2% on 5-way 5-shot case; in tiered-ImageNet, our methods can outperform others at least 4.1% on 5-way 1-shot case, 2.2% on 5-way 5-shot case.

(4) Finally, just looking at the comparison results with ResNet12 backbone. In mini-ImageNet, our methods can outperform others at least 2.6% on 5-way 1-shot case, 2.2%

TABLE II

THE 5-WAY SUPERVISED FEW-SHOT CLASSIFICATION ACCURACIES ON MINI-IMAGENET AND TIERED-IMAGENET WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES

Method	Backbone	mini-ImageNet		tiered-ImageNet	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet [42]	4CONV	49.42 ± 0.78	68.20 ± 0.66	-	-
MAML [34]	4CONV	48.70 ± 1.84	63.11 ± 0.92	-	-
RelationNet [53]	ResNet18	52.48 ± 0.86	69.83 ± 0.68	-	-
CLLO [54]	ResNet18	51.75 ± 0.80	74.27 ± 0.63	-	-
CLLO++ [54]	ResNet18	51.87 ± 0.77	75.68 ± 0.63	-	-
LEO [19]	WRN	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09
TPN [55]	4CONV	52.78 ± 0.27	66.42 ± 0.21	55.74 ± 0.29	71.01 ± 0.23
AM3 [56]	ResNet12	65.30 ± 0.49	78.10 ± 0.36	69.08 ± 0.47	82.58 ± 0.31
TapNet [57]	ResNet12	61.65 ± 0.15	76.36 ± 0.10	63.08 ± 0.15	80.26 ± 0.12
CTM [58]	ResNet18	64.12 ± 0.82	80.51 ± 0.13	-	-
DenseCls [47]	ResNet12	62.53 ± 0.19	79.77 ± 0.19	-	-
MetaOpt [59]	ResNet12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
TEAM [60]	ResNet12	60.07 ± 0.63	75.90 ± 0.52	-	-
DWC [29]	ResNet12	63.73 ± 0.62	81.19 ± 0.43	70.44 ± 0.32	85.43 ± 0.21
S2M2 [30]	WRN	64.93 ± 0.18	83.18 ± 0.11	73.71 ± 0.22	88.59 ± 0.14
Fine-tuning [61]	WRN	65.73 ± 0.68	78.40 ± 0.52	73.34 ± 0.71	85.50 ± 0.50
DSN-MR [62]	ResNet12	64.60 ± 0.72	79.51 ± 0.50	67.39 ± 0.82	82.85 ± 0.56
ICI [45]	ResNet12	66.80	79.26	80.79	87.92
MABAS [63]	ResNet12	64.21 ± 0.82	81.01 ± 0.57	-	-
DivCoop [28]	ResNet12	64.14 ± 0.62	81.23 ± 0.42	-	-
HGNN [64]	4CONV	60.03 ± 0.51	79.64 ± 0.36	64.32 ± 0.49	83.34 ± 0.45
URT [48]	ResNet12	72.23	83.35	80.30	88.63
DC [65]	WRN	68.57 ± 0.55	82.88 ± 0.42	78.19 ± 0.25	89.90 ± 0.41
MELR [66]	ResNet12	67.40 ± 0.43	83.40 ± 0.28	72.14 ± 0.51	87.01 ± 0.35
ODE [67]	ResNet12	67.76 ± 0.46	82.71 ± 0.31	71.89 ± 0.52	85.96 ± 0.35
LA-MVFC	ResNet12	**74.81** ± 1.12	**85.58** ± 0.61	83.95 ± 1.13	**90.75** ± 0.58
SA-MVFC	ResNet12	74.17 ± 0.96	84.66 ± 0.51	**84.43** ± 0.77	90.68 ± 0.51
GA-MVFC	ResNet12	74.05 ± 0.54	85.16 ± 0.62	84.27 ± 0.51	90.12 ± 0.27

TABLE III

THE 5-WAY SUPERVISED FEW-SHOT CLASSIFICATION ACCURACIES ON CIFAR-FS AND FC100 WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES

Method	Backbone	CIFAR-FS		FC100	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ProtoNet [42]	4CONV	55.50 ± 0.70	72.00 ± 0.60	35.30 ± 0.60	48.60 ± 0.60
MAML [34]	4CONV	58.90 ± 1.90	71.50 ± 1.00	-	-
RelationNet [53]	4CONV	55.00 ± 1.00	69.30 ± 0.80	-	-
TADAM [43]	ResNet12	-	-	40.10 ± 0.40	56.10 ± 0.40
DenseCls [47]	ResNet12	-	-	42.04 ± 0.17	57.63 ± 0.23
MetaOpt [59]	ResNet12	72.00 ± 0.70	84.20 ± 0.50	41.10 ± 0.60	55.50 ± 0.60
TEAM [60]	ResNet12	70.43 ± 1.03	81.25 ± 0.92	-	-
MABAS [63]	ResNet12	73.24 ± 0.95	85.65 ± 0.65	41.74 ± 0.73	57.11 ± 0.75
Fine-tuning [61]	WRN	76.58 ± 0.68	85.79 ± 0.50	43.16 ± 0.59	57.57 ± 0.55
DSN-MR [62]	ResNet12	75.60 ± 0.90	86.20 ± 0.60	-	-
LA-MVFC	ResNet12	81.83 ± 1.16	**89.27** ± 0.63	**47.02** ± 1.05	61.88 ± 0.80
SA-MVFC	ResNet12	**82.17** ± 0.93	89.06 ± 0.54	46.15 ± 0.76	**62.58** ± 0.49
GA-MVFC	ResNet12	81.00 ± 0.75	89.14 ± 0.69	45.32 ± 0.33	61.95 ± 0.55

TABLE IV

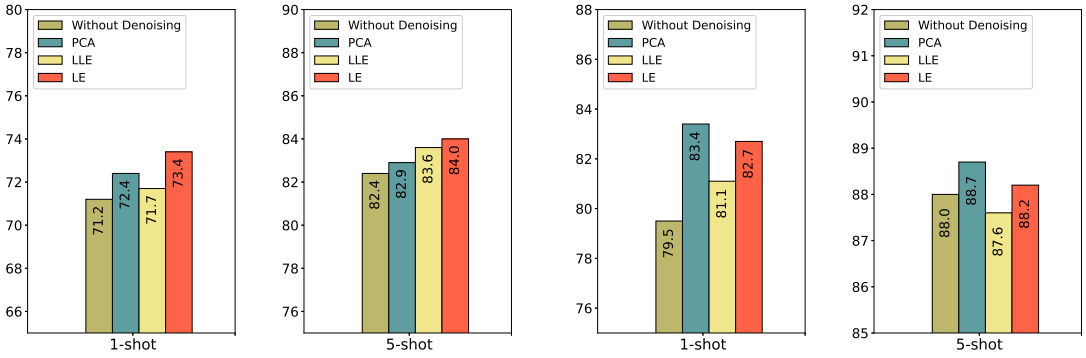
THE 5-WAY SEMI-SUPERVISED FEW-SHOT CLASSIFICATION ACCURACIES ON MINI-IMAGENET AND TIERED-IMAGENET WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES. WE USE 100 UNLABELED SAMPLES

Method	Backbone	mini-ImageNet		tiered-ImageNet	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MSkM [68]	4CONV	50.41 ± 0.31	64.39 ± 0.24	49.04 ± 0.31	62.96 ± 0.14
TPN [55]	4CONV	55.51 ± 0.86	69.86 ± 0.65	59.91 ± 0.94	73.30 ± 0.75
LST [44]	ResNet12	70.10 ± 1.90	78.70 ± 0.80	77.70 ± 1.60	85.20 ± 0.80
EPNet [31]	ResNet12	75.36 ± 1.01	84.07 ± 0.60	81.79 ± 0.97	88.45 ± 0.61
TransMatch [69]	WRN	63.02 ± 1.07	81.19 ± 0.59	-	-
ICI [45]	ResNet12	71.41	81.12	85.44	89.12
LA-MVFC	ResNet12	**79.76** ± 1.16	**87.64** ± 0.53	**87.56** ± 1.04	**91.90** ± 0.56
SA-MVFC	ResNet12	77.13 ± 0.78	87.39 ± 0.32	86.21 ± 0.82	90.41 ± 0.51
GA-MVFC	ResNet12	78.69 ± 0.55	87.42 ± 0.67	86.92 ± 0.76	89.87 ± 0.51

TABLE V

ABLATION STUDIES OF OUR METHOD IN 5-WAY SUPERVISED FEW-SHOT CASE. SS-R AND SS-M DENOTE THAT THE FEATURE EXTRACTION MODEL (FEM) ADOPTS ROTATION-BASED SEMI-SUPERVISION AND MIRROR-BASED SEMI-SUPERVISION. (SEE SECTION IV-F). LA, SA, GA DENOTE LOSS-ATTENTION, SELF-ATTENTION AND GRAPH ATTENTION STRATEGIES

	Baseline	Self-Train	SS-R	SS-M	Denoising	LA	SA	GA	mini-ImageNet	
									5-way 1-shot	5-way 5-shot
①	✓								56.06	75.70
②	✓	✓							64.21	77.48
③	✓	✓	✓						71.33	80.36
④	✓	✓		✓					70.95	81.82
⑤	✓	✓	✓	✓					71.24	82.38
⑥	✓	✓	✓	✓	✓				73.42	83.96
⑦	✓	✓	✓	✓	✓	✓			74.81	85.58
⑧	✓	✓	✓	✓	✓		✓		74.17	84.66
⑨	✓	✓	✓	✓	✓			✓	74.05	85.16



(a) mini-ImageNet (b) tiered-ImageNet

Fig. 4. Comparison results of supervised case with different feature denoising strategies.

on 5-way 5-shot case; in tiered-ImageNet, our methods can outperform others at least 3.6% on 5-way 1-shot case, 2.1% on 5-way 5-shot case; in CIFAR-FS, our methods can outperform others at least 6.6% on 5-way 1-shot case, 3.1% on 5-way 5-shot case; in FC100, our methods can outperform others at least 5.0% on 5-way 1-shot case, 5.0% on 5-way 5-shot case.

D. t-SNE Visualization

Here, we look at the changes brought about by our method intuitively from t-SNE visualization [75]. We show the t-SNE results in Figure 3 to observe the feature distribution. SS-R-Fea and SS-M-Fea represent two categories of different single-view features (see Section IV-F).Fuse-Fea denotes the fusion feature with loss-attention. The reported data is randomly selected from the mini-ImageNet. We observe that our fusion feature distribution is more discriminative than the single-view feature, which is more helpful for classification tasks.

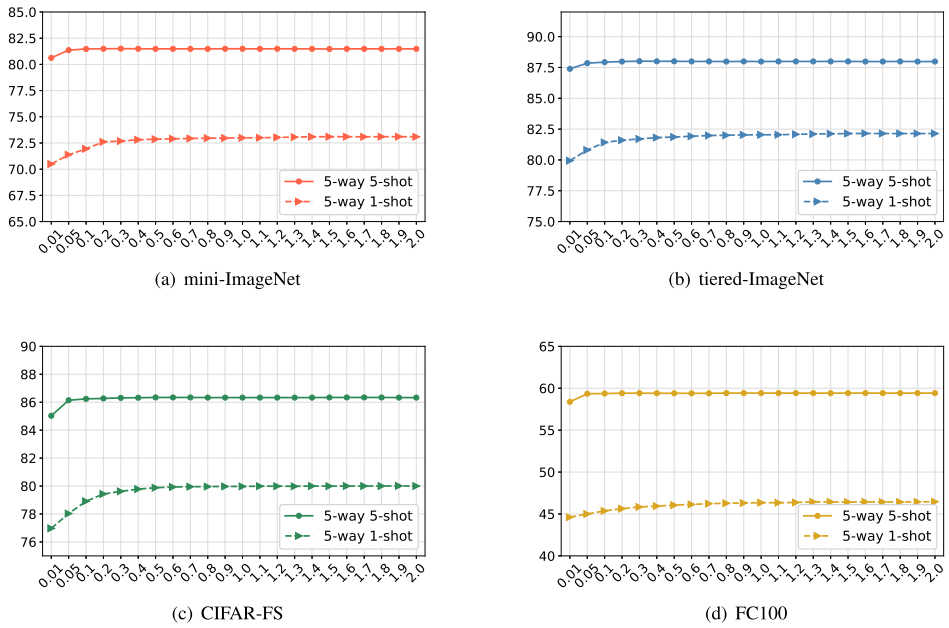
E. Ablation Studies

We design ablation studies in mini-ImageNet and tiered-ImageNet to analyze the efficiency of our methods block by block, including self-supervision, multi-view feature fusion, multi-view feature denoising, attention strategy, view number, and self-training. The detailed results are listed in Table V. The baseline denotes using ICI [45] based FEM and logistic regression classifier. Besides, we only adopt two kinds of features when completing the feature fusion process for convenience, and please see Table VI for more kinds of fusion results.

1) Influence of Self-Supervision: Seeing Table V, SS-R and SS-M denote the rotation based semi-supervision and mirror based semi-supervision (see Section IV-F). Comparison ② with ③, and ② with ④, we observe that different kinds of self-supervision strategies help a lot for the method, can improve the results 0.1%-5.9% on different cases.

2) Influence of Multi-View Feature Fusion: Looking at the comparison results of ③, ④, ⑤ in Table V. We observe that merely using the fusion strategy seems to have little improvements for the final performance or even negative effects. Not because this method is inadvisable, it just need some extra tricks, which are discussed in the following sections.

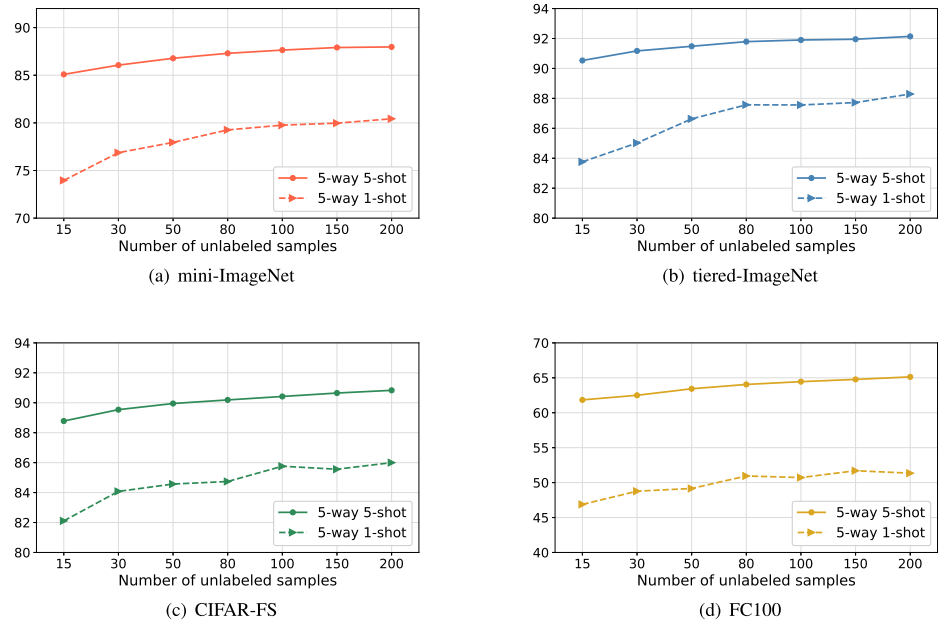
3) Influence of Multi-View Feature Denoising: The first trick is our designed feature denoising, which transforms the different features into a unified space through subspace learning. See the Table V, comparing ⑤ with ⑥, we find that the performances have significant improvements of 0.7%-3.9%. It has demonstrated the efficiency of this strategy. While there exist some subspace learning approaches, such as LE [50]



(a) mini-ImageNet (b) tiered-ImageNet

(c) CIFAR-FS (d) FC100

Fig. 5. Comparison results of LA-MVFC with different η.



(a) mini-ImageNet (b) tiered-ImageNet

(c) CIFAR-FS (d) FC100

Fig. 6. Comparison results of semi-supervised LA-MVFC with varied unlabeled samples.

PCA [51], LLE [49], and it is interesting to know the influences of different ways. Here, we test three kinds of subspace learning methods, the results are shown in Figure 4. Based on the above analysis, we find that all the strategies are helpful to our method, and LE is the most.

4) Influence of Attention Strategy: The next trick is the attention mechanism. We designed two kinds of strategies, including loss-attention (LA) and self-attention (SA). See Table V. Comparing ⑥ with ⑦, ⑥ with ⑧, and ⑥ with ⑨ we observe that the attention blocks improve the original performances of 0.6%-2.0%. To further evaluate the attention blocks, we compare the attention-based results with fixed weights, which is demonstrated in Table VII. The experimental results illustrate that the updated weight is more reasonable. Moreover, from Equation (6), it is evident that η is a parameter that influences the to-be-learned weights in LA-MVFC. For fairness and convenience, we froze the η to 1.4 for all the experiments. Here, we show the experimental results with other values in Figure 5 and find that our LA-MVFC is not sensitive to this parameter.

5) Influence of View Numbers: As the description above, Table V only fuse two views of features. Besides SS-R-Fea and SS-M-Fea, we introduce Std-Fea and Meat-Fea (described in Section IV-F) to evaluate the proposed method further. We list the performance on mini-ImageNet in Table VI. All the results are based on the supervised setting with a 5-way 1-shot case. We find that whether in LA-MVFC, SA-MVFC or GA-MVFC, the more features are fused, the better the results are obtained. The reason is that: the designed attention mechanisms for fusing features can automatically adjust the impact of distinct

TABLE VI
COMPARISON RESULTS OF FUSING MULTIPLE VIEWS OF FEATURES WITH THE SUPERVISED SETTING ON MINI-IMAGENET

Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Std-Fea	✓				✓	✓	✓				✓	✓	✓		✓
Meta-Fea		✓			✓			✓	✓		✓	✓		✓	✓
SS-R-Fea			✓			✓		✓		✓	✓		✓	✓	✓
SS-M-Fea				✓			✓		✓	✓		✓	✓	✓	✓
LA-MVFC	68.7	67.0	71.3	71.0	69.0	74.3	74.5	72.4	72.1	74.8	75.1	75.0	75.6	75.4	76.1
SA-MVFC	68.7	67.0	71.3	71.0	69.3	73.9	74.6	72.7	71.8	74.1	75.0	74.7	75.2	74.7	75.9
GA-MVFC	68.7	67.0	71.3	71.0	68.9	73.6	74.4	72.9	72.2	74.3	75.4	74.9	75.1	75.2	75.5

TABLE VII
COMPARISON RESULTS WITH FIXED WEIGHTS ON 5-WAY SUPERVISED FEW-SHOT CASE. (A, B) INDICATES THAT THE SS-R-FEA'S WEIGHT IS "A", AND SS-M-FEA'S WEIGHT IS "B". LA-MVFC, SA-MVFC AND GA-MVFC EMPLOY THE LOSS-ATTENTION, SELF-ATTENTION AND GRAPH-ATTENTION BLOCKS TO UPDATE THE WEIGHTS AUTOMATICALLY FOR EACH EPISODE

Weight	**mini-ImageNet**		**tiered-ImageNet**	
	1-**shot**	5-**shot**	1-**shot**	5-**shot**
(0.1, 0.9)	71.64	80.09	80.26	87.12
(0.3, 0.7)	72.11	84.27	81.47	87.94
(0.5, 0.5)	73.42	83.96	83.38	88.71
(0.7, 0.3)	72.86	81.98	80.96	87.38
(0.9, 0.1)	71.97	80.64	81.93	87.69
LA-MVFC	**74.81**	**85.58**	83.95	**90.75**
SA-MVFC	74.17	84.66	**84.43**	90.68
GA-MVFC	74.05	85.16	84.27	90.12

features on the results and ensures that the fusion result is not worse than that of a single one. Besides, we conclude that if the to-be-fused features have similar performances, the final stacking result may significantly improve.

6) Influence of Self-Training: From Table V, we observe that it can improve the performance of 0.9%-9.4%. Besides, self-training can extend the standard supervised FSL to the semi-supervised case. Here, we take LA-MVFC as an example to observe the impact of the number of unlabeled samples on the results, which is listed in Figure 6. The performance of the proposed method increases with the unlabeled instances. And the results become saturation after 100 unlabeled samples.

VI. CONCLUSION

There is a fundamental problem in decoupled FSL: the pre-trained feature extraction model (FEM) is challenging to adapt to the novel class in the cross-category setting. To address this challenge, we propose Loss-Attention Feature Collaboration (LA-MVFC), Self-Attention Feature Collaboration (SA-MVFC) and Graph-Attention Feature Collaboration (GA-MVFC), which fuses multi-view features to achieve collaboratively represent samples. It benefits from enhancing the efficiency and robustness of the FSL-based model. LA-MVFC, SA-MVFC, GA-MVFC are simple non-parametric methods that exploit the existing FEMs in a direct way. Experimental results have evaluated their effectiveness. In future work, on the one hand, we will focus more on attention-based few-shot learning methods, and it is interesting to complete the feature collaboration in the pre-training phase; on the other hand, we will discuss in depth whether the current FSL paradigm is suitable for real application scenarios, and strive to define a new FSL paradigm to make it closer to real applications.

REFERENCES

[1] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.

[2] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[3] B. Fan et al., "Contextual multi-scale feature learning for person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 655–663.

[4] Y. Zhang, J. Wu, Z. Cai, and S. Y. Philip, "Multi-view multi-label learning with sparse feature selection for image annotation," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2844–2857, Nov. 2020.

[5] S. Shao, L. Xing, R. Xu, W. Liu, Y.-J. Wang, and B.-D. Liu, "MDFM: Multi-decision fusing model for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5151–5162, Aug. 2022.

[6] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Jun. 2016.

[7] Y. Zhang, Y. Wang, X. Chen, X. Jiang, and Y. Zhou, "Spectral–spatial feature extraction with dual graph autoencoder for hyperspectral image clustering," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 5, 2022, doi: 10.1109/TCSVT.2022.3196679.

[8] S. Shao, R. Xu, W. Liu, B.-D. Liu, and Y.-J. Wang, "Label embedded dictionary learning for image classification," *Neurocomputing*, vol. 385, pp. 122–131, Apr. 2020.

[9] W. Liu, G. Lin, T. Zhang, and Z. Liu, "Guided co-segmentation network for fast video object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1607–1617, Apr. 2021.

[10] S. Wan, Y. Xia, L. Qi, Y. H. Yang, and M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1756–1768, Jul. 2020.

[11] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2019.

[12] Y. Wang, Y. Chen, W. Wang, and H. Zhu, "MSGAN: Multi-stage generative adversarial networks for cross-modality domain adaptation," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 520–524.

[13] Y. Wang, Y. Zhao, S. Ying, S. Du, and Y. Gao, "Rotation-invariant point cloud representation for 3-D model recognition," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10948–10956, Oct. 2022.

[14] E. Perez, D. Kiela, and K. Cho, "True few-shot learning with language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 11054–11070.

[15] L. Xing, S. Shao, Y. Ma, Y. Wang, W. Liu, and B. Liu, "Learning to cooperate: Decision fusion method for few-shot remote-sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] F. Zhao, J. Zhao, S. Yan, and J. Feng, "Dynamic conditional networks for few-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–35.

[17] L. Xing, Y. Ma, W. Cao, S. Shao, W. Liu, and B. Liu, "Rethinking few-shot remote sensing scene classification: A good embedding is all you need?" *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[18] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.

[19] A. A. Rusu et al., "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[20] X. Sun, B. Wang, Z. Wang, H. Li, H. Li, and K. Fu, "Research progress on few-shot learning for remote sensing image interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2387–2402, 2021.

[21] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "F-VAEGAN-D2: A feature generating framework for any-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10275–10284.

[22] S. Chen et al., "FREE: Feature refinement for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 122–131.

[23] S. Chen et al., "TransZero: Attribute-guided transformer for zero-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 2, 2022, p. 3.

[24] S. Chen et al., "MSDN: Mutually semantic distillation network for zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Jun. 2022, pp. 7612–7621.

[25] S. Chen et al., "HSVA: Hierarchical semantic-visual adaptation for zero-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 16622–16634.

[26] W. Wang, Y. Shi, S. Chen, Q. Peng, F. Zheng, and X. You, "Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1106–1112.

[27] S. Chen et al., "GNDAN: Graph navigated dual attention network for zero-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 4, 2022, doi: 10.1109/TNNLS.2022.3155602.

[28] N. Dvornik, C. Schmid, and J. Mairal, "Selecting relevant features from a multi-domain representation for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 769–786.

[29] N. Dvornik, J. Mairal, and C. Schmid, "Diversity with cooperation: Ensemble methods for few-shot classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3723–3731.

[30] P. Mangla, M. Singh, A. Sinha, N. Kumari, V. N. Balasubramanian, and B. Krishnamurthy, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2218–2227.

[31] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–138.

[32] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 266–282.

[33] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10836–10846.

[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[35] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[37] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.

[38] S. Shao et al., "MHFC: Multi-head feature collaboration for few-shot learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4193–4201.

[39] J. Zhang, J. Song, L. Gao, Y. Liu, and H. T. Shen, "Progressive meta-learning with curriculum," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5916–5930, Sep. 2022.

[40] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6116–6125, Dec. 2019.

[41] J. Zhang, J. Song, Y. Yao, and L. Gao, "Curriculum-based meta-learning," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1838–1846.

[42] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[43] B. Oreshkin, P. R. López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 721–731.

[44] X. Li et al., "Learning to self-train for semi-supervised few-shot classification," in *Proc. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 10276–10286.

[45] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu, "Instance credibility inference for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12836–12845.

[46] R. Xu et al., "GCT: Graph co-training for semi-supervised few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 4, 2022, doi: 10.1109/TCSVT.2022.3196550.

[47] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, "Dense classification and implanting for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9258–9267.

[48] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[49] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[50] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 585–591.

[51] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Roy. Statist. Soc. B*, vol. 61, no. 3, pp. 611–622, 1999.

[52] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 759–766.

[53] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[54] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[55] Y. Liu et al., "Learning to propagate labels: Transductive propagation network for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.

[56] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4847–4857.

[57] S. W. Yoon, J. Seo, and J. Moon, "TapNet: Neural network augmented with task-adaptive projection for few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7115–7123.

[58] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–10.

[59] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10657–10665.

[60] L. Qiao, Y. Shi, J. Li, Y. Tian, T. Huang, and Y. Wang, "Transductive episodic-wise adaptive metric for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3603–3612.

[61] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–20.

[62] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4136–4145.

[63] J. Kim, H. Kim, and G. Kim, "Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 599–617.

[64] C. Chen, K. Li, W. Wei, J. T. Zhou, and Z. Zeng, "Hierarchical graph neural networks for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 240–252, Jan. 2022.

[65] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–13.

[66] N. Fei, Z. Lu, T. Xiang, and S. Huang, "MELR: Meta-learning via modeling episode-level relationships for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–20.

[67] C. Xu et al., "Learning dynamic alignment via meta-filter for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5182–5191.

[68] M. Ren et al., "Meta-learning for semi-supervised few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[69] Z. Yu, L. Chen, Z. Cheng, and J. Luo, "TransMatch: A transfer-learning scheme for semi-supervised few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12856–12864.

[70] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.
[71] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Los Angeles, CA, USA, Tech. Rep., 2011.
[72] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[73] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Comput. Sci. Dept., Univ. of Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
[74] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[75] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

Shuai Shao received the M.S. and Ph.D. degrees from the College of Control Science and Engineering, China University of Petroleum (East China), where he is currently pursuing the Ph.D. degree. Currently, he is a Postdoctoral Researcher with the Zhejiang Laboratory. He was a Visiting Student at Tsinghua University, July from 2019 to July 2020. During his Ph.D., he published five papers as the first author in ACM Multimedia (ACMMM) and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT). His research interests include image processing, computer vision, and few-shot learning.

Lei Xing received the B.S. degree from the College of Oceanography and Space Informatics, China University of Petroleum (East China), where he is currently pursuing the M.S. degree. His main research interests include machine learning and computer vision.

Yanjiang Wang received the M.S. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1989, and the Ph.D. degree from Beijing Jiaotong University, Beijing, in 2001. Currently, he is a Professor with the College of Control Science and Engineering, China University of Petroleum, Qingdao, China. He is the Head of the Institute of Signal and Information Processing, China University of Petroleum. He has presided over and completed two projects of the National Natural Science Foundation of China. His research interests include pattern recognition, computer vision, and cognitive computation.

Baodi Liu (Member, IEEE) received the B.S. degree in signal and information processing from the China University of Petroleum (East China) in 2007 and the Ph.D. degree in the Department of Electronic Engineering, Tsinghua University in 2013. He was a Visiting Scholar at the University of California, Merced, from 2019 to 2020. He is currently an Associate Professor with the College of Control Science and Engineering, China University of Petroleum (East China). His research interests include image processing, computer vision, and machine learning.

Weifeng Liu (Senior Member, IEEE) received the double B.S. degree in automation and business administration and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2002 and 2007, respectively. He was a Visiting Scholar at the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia, from 2011 to 2012. He is currently a Professor with the College of Control Science and Engineering, China University of Petroleum (East China), China. He has authored or coauthored a dozen papers in top journals and prestigious conferences, including ten ESI Highly Cited Papers and three ESI Hot Papers. His current research interests include pattern recognition and machine learning. He is the Co-Chair of IEEE SMC Technical Committee on cognitive computing. He is an Associate Editor of *Neural Processing Letter* and the Guest Editor for the Special Issue on Signal Processing, *IET Computer Vision*, *Neurocomputing*, and *Remote Sensing*. He also serves dozens of journals and conferences.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees from Tufts University, MA, USA, all in electrical engineering.

He is currently a Professor and the Director of the Vision and Image Processing Laboratory with the Department of Computer and Information Science, University of Macau. His research interests include image processing, computer vision, machine learning, and multimedia security. He is a fellow of the International Society for Optical Engineering (SPIE), and a Senior Member of the China Computer Federation (CCF). He was a recipient of the Third Price of Macao Natural Science Award in 2014 and 2020. He is a Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He was listed as "World's Top 2% Scientists" on the Stanford University Releases List in 2020 and 2021 and the "Highly Cited Researcher" in the Web of Science in 2020 and 2021. He is an Associate Editor of IEEE TRANSACTIONS ON NEUTRAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), and four other journals.