IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 12, 2022, accepted May 11, 2022, date of publication May 17, 2022, date of current version May 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175798

A Multiview Metric Learning Method for Few-Shot Fine-Grained Classification

ZHUANG MIAO, XUN ZHAO, JIABAO WANG, BO XU, YANG LI, AND HANG LI
College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Jiabao Wang (jiabao_1108@163.com)

This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20200581, in part by the National Natural Science Foundation of China under Grant 61806220, and in part by the China Postdoctoral Science Foundation under Grant 2020M683754 and Grant 2021T140799.

ABSTRACT Few-shot fine-grained image classification aims to solve the learning problem with few limited labeled examples. The existing methods use data augmentation to randomly transform the original examples to get new examples, and then use the new examples to train the model to improve the robustness and generalization ability of the learnt model. Due to each iteration of these methods uses a random transformation to get a new example, it will cause the unstable problem of the class center in the feature measurement stage. To solve this problem, a Multi-view Metric Learning (MML) method is proposed, which is based on a new concept (View Bag) and its effective similarity measurement method to achieve better few-shot fine-grained image classification. Firstly, a new example obtained by a kind of data augmentation is defined as a view, and a set of views generated by multiple data augmentation is defined as a view bag. Then, the view bag is sent into the model to extract the features, and a multi-view metric method with the view bag as the object is proposed to overcome the unstable problem of the class center. Finally, classification is performed by measuring the similarity between view bags. Experiments are conducted on three public datasets, CUB-2011-200, Stanford-Dogs and Stanford-Cars. The proposed method achieves 71.61±0.87%, 57.78±0.96% and 74.02±0.84% for the 5-way 1-shot classification task, and 88.72±0.51%, 76.30±0.68% and 92.94±0.37% for the 5-way 5-shot classification task, which have the state-of-the-art performances. Under the condition of the same backbone network, the proposed multi-view metric method can measure the similarity between examples more effectively, and improve the robustness and generalization ability of the model.

INDEX TERMS Few-shot fine-grained image classification, fine-grained image classification, metric learning, date augmentation, multi-view metric.

I. INTRODUCTION

Fine-grained visual classification (FGVC) aims to distinguish sub-categories of a general category, such as birds, cars and aircrafts. Due to the subtle inter-class differences and large intra-class variations, FGVC tasks are more challenging than the general image classification tasks. Over the past years, the methods of FGVC have made great achievements [1]–[6] with the development of datasets [7]–[10]. Compared with general classification datasets, it is more difficult to annotate fine-grained dataset due to the subtle difference among sub-categories. Furthermore, fine-grained annotations is labor intensive, which limits both scalability and practicality of real-world fine-grained applications. In order to tackle the problem of annotation, researchers [11]–[18] gradually focus on few-shot fine-grained classification task, which requires only few labeled examples.

Recently, few-shot fine-grained image classification is mainly divided into two categories: meta-learning method [19]–[23] and metric learning method [24]–[28]. The former focuses on learning a meta model for classification, while the latter focuses on measuring the similarity between examples for classification. Although the two methods are different, they both adopt data augmentation method to increase training data. In the data pre-processing, these methods use data augmentation function to randomly process the original image to get a new image before model training, such as random cropping or random flipping. In each iteration, the original example generates a different new example for training, so the data augmentation method increases the

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

number of training examples, can significantly improve the robustness and generalization ability of the model, and has been widely used in the field of image classification. However, using random augmentation method to generate new examples will make the feature of the model different in each iteration. When the number of intra-class examples is small, it will lead to a large difference for the class center in each iteration, which makes the calculation of the similarity measurement between the query example and the class center unstable. To solve this problem, researchers [29]–[31] propose transductive learning, which updates the class centers with unlabeled query examples. Although this kind of method can overcome the model instability caused by the uncertainty of examples, it has to update the class center for multiple times in each iteration, resulting in a large cost of computing time.

If the original image transformed by a kind of data augmentation method is regarded as an example under a certain perspective, which can be called a view of example. Thus, in the whole training process, each original image will generate multiple views, while the traditional method only randomly uses a single view of each image in each iteration, so each class contains a random variety of views, which leads to the instability of the class center. In practice, given two objects in two images, we tend to measure the similarities of the two images from different views and combine them as the standard to measure whether the two images belong to the same category. Inspired by this, in order to solve the problem of model instability caused by traditional data augmentation and make full use of new examples after augmentation, we proposed a Multi-view Metric Learning (MML) method, which mainly consists of view bag creation module and multi-view metric module. Firstly, a view bag is constructed from a variety of view examples obtained from the original example by multiple data augmentation method, which contains a number of different views. Then, the view bag is used as the input of the model to learn, and the multi-view metric module is used to measure the similarity between different views. Finally, the similarity between query examples in all views and original examples in all views is used for classification.

The main contributions of this paper include three aspects:

- A new concept, View Bag, is proposed to describe the set of the multiple transformations of the same image. The view bag creation method built multiple views by augmenting from the same original example in each iteration for training.
- A Multi-view Metric Learning method is proposed for few-shot fine-grained image classification, in which a new multi-view metric method is proposed to measure the similarity between two view bags. It improve the robustness of the model training in each iteration.
- Experimental results show that our method reaches a new state-of-the-art performance on three public fine-grained datasets. It also verified the effectiveness of several commonly used data augmentation methods and our multi-view metric method for FGVC.

II. RELATED WORK

A. FINE-GRAINED VISUAL CLASSIFICATION

In recent years, researchers have conducted many studies on fine-grained visual classification tasks and achieved abundant results [1]–[6], [32]–[35]. Fine-grained objects belong to the same super-category and their inter-class differences exist in local regions, so the key is discriminant feature extraction of local regions. In order to extract more discriminant features, on the one hand, researchers classify objects by locating local discriminant regions. For example, Shroff *et al.* [1] designed a circular attention structure to classify targets by extracting discriminant regional features at multiple moments. Ding *et al.* [33] proposed a Selective Sparse Sampling Networks (S3Ns) to learn sparse attention from class peak responses, which typically corresponds to informative object parts. Zhuang *et al.* [2] proposed an Attentive Pairwise Interaction Network (API-NET) based on the principle that a person classifies fine-grained objects by comparing them in pairs. API-Net can adaptively detect contrast cues from a pair of images and has strong discriminant characteristics through pair interaction learning. And more, He *et al.* [34] also proposed a multi-scale and multi-granularity deep reinforcement learning approach (M2DRL), which learns multi-granularity discriminative region attention and multi-scale region-based feature representation for fine-grained classification. To locate the discriminative regions fast, He *et al.* [35] proposed a weakly supervised discriminative localization approach for fast fine-grained classification.

On the other hand, researchers improve the feature extraction ability by designing effective loss functions to supervise the learning process of the model. For example, Chang *et al.* [3] proposed a channel loss to extract efficient channel information. Rao *et al.* [32] presented a counterfactual attention learning (CAL) method to supervise attention learning based on causal inference. Gao *et al.* [6] adopted a contrastive loss to push the features of different classes away while pulling the positive pairs close.

However, the above methods are always trained based on large-scale fine-grained datasets. The annotation of fine-grained datasets is more difficult than that of general classified data sets, so the generalization of these methods is limited.

B. FEW-SHOT FINE-GRAINED IMAGE CLASSIFICATION

Recently, few-shot fine-grained image classification has been developed rapidly, and the current methods are mainly divided into two categories: meta-learning method and metric learning method. The former focuses on learning a meta-model for classification, such as, multi-attention meta-learning [36] uses multiple attention mechanisms to extract regional discriminant features for classification, and uses gradient-based meta-learning method to update model parameters to focus on the different parts adaptively. The latter focuses on the classification by measuring the similarity between examples and the class center. In metric learning methods, most researchers focus on improving metric

algorithms to improve the accuracy of model recognition. Li *et al.* [24] proposed a covariance measurement method, which calculates the similarity between the covariance matrices of features to obtain an effective similarity measurement. Huang *et al.* [26] proposed a low-rank pairwise alignment bilinear network by extracting subtle differences between images. Li *et al.* [28] proposed a dual similarity measurement network that uses two measurement methods to measure the similarity between examples simultaneously because a single measurement method often has limitations when calculating the similarity between examples.

In addition, in order to avoid the instability of the class center caused by randomly angmented examples, researchers [29]–[31] improved the accuracy of measurement by calculating more accurate class center methods. For example, Meta-Confidence Transduction (MCT) method [29] obtained more accurate class center by updating the class center by weighting all intra-class examples. Such methods usually require multiple iterations to update the class center, which is too expensive to calculate. In order to avoid the instability of the class center caused by data augmentation and measure the similarity between examples more accurately, a multi-view metric method is proposed in this paper. This method does not update the class center repeatedly, but directly calculates the similarity between examples by measuring the similarity between various augmented examples. Since all the expanded examples are covered in the measure stage, the similarity between examples can be measured more accurately.

III. DEFINITION

A. PROBLEM DEFINITION

Few-shot fine-grained image classification aims to distinguish C categories, but each category only has N labeled examples, which specifically called C-way N-shot classification task. To measure the classification ability of the model with few labeled examples, a dataset is divided into target set $\mathcal{T}$ and auxiliary set $\mathcal{A}$ (as shown in Figure 1), where $\mathcal{A} \cap \mathcal{T} = \phi$, and the model trained on $\mathcal{A}$ are tested on $\mathcal{T}$. It is worth noting that the label space of set $\mathcal{T}$ is disjoint with the label space of set $\mathcal{A}$.

In the training process, the episodic training mechanism is adopted. Specifically, for each epoch in training, one episode randomly selects the examples in C categories from the set $\mathcal{A}$ to form the support set $\mathcal{A}_S$ and query set $\mathcal{A}_Q$. The definition is as follows:

$$\mathcal{A}_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{C\times N}, \mathcal{A}_Q = \{(\mathbf{x}_j, y_j)\}_{j=1}^{M}, \tag{1}$$

where $\mathcal{A}_S \cap \mathcal{A}_Q = \phi$, $\mathbf{x}_i$ denotes the i-th example of $\mathcal{A}_S$, $\mathbf{x}_j$ denotes the j-th example of $\mathcal{A}_Q$, $y_i, y_j \in \{1, \ldots, C\}$ denotes the label. N and M are the number of examples in each category of $\mathcal{A}_S$ and query set $\mathcal{A}_Q$ respectively. After t epochs, t episodes have been used to train the mapping function.

Once trained, we predict the labels of the set $\mathcal{T}$ via the mapping function conditioned on the set $\mathcal{A}$. In the test process, the set $\mathcal{T}$ are divided into support set $\mathcal{T}_S$ and query set $\mathcal{T}_Q$, and the difference is that $\mathcal{T}_S$ contains N labeled examples for each category and $\mathcal{T}_Q$ contains M unlabeled examples, and $M \gg N$. The definition is as follows:

$$\mathcal{T} = \{\mathcal{T}_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N\times C} \cup \mathcal{T}_Q = \{\mathbf{x}_m\}_{m=1}^{M}\}, \tag{2}$$

where $\mathcal{T}_S \cap \mathcal{T}_Q = \phi$, $\mathbf{x}_i$ denotes the i-th example of $\mathcal{T}_S$, $y_i \in \{1, \ldots, C\}$, and $\mathbf{x}_m$ denotes the m-th example of $\mathcal{T}_Q$.

B. VIEW BAG

In order to solve the problem of model robustness caused by a small mount of examples, data augmentation method has been widely used in few-shot fine-grained image classification task. Such methods usually use a data augmentation method, including random cropping, random flipping and color jittering, to transform an image into another new one, which is used for model training. We formulate the transformation as follows:

An example $(\mathbf{x}, y)$ is transformed into $(\mathbf{x}', y)$, where y is the label, and the transformation is represented as:

$$\mathbf{x}' = T(\mathbf{x}), \tag{3}$$

where $T(\cdot)$ is the data augmentation function, $\mathbf{x}$ and $\mathbf{x}'$ denotes the original image and the transformed image, respectively.

Although the above method can increase the amount of training data and improve the robustness of the model, it may have a negative impact on the few-shot fine-grained image classification. This is because the new example generated by the original example after a transformation method can be regarded as the example with modified local pixel values, and the features of the same example in different views are different, so a certain view feature cannot represent the whole features of the example. Therefore, the similarity between examples measured by the state of a view cannot fully measure the similarity between the original examples. In addition, after data augmentation, an example will often produce a variety of views, such as noisy view, partial view and flipped view, and different views reflect different characteristics.

Therefore, we proposed a new concept, View Bag, to represent the set of the multiple transformations of the same image. The view bag is formalized as follows:

Given an example $(\mathbf{x}, y)$, $\mathbf{x}$ is the original image, y is the label, and the view bag is obtained through P augmentation methods, denoted as $(\mathbf{B}(\mathbf{x}), y)$.

$$\mathbf{B}(\mathbf{x}) = \{\mathbf{x}^p | \mathbf{x}^p = T_p(\mathbf{x})\}_{p=1}^{P}, \tag{4}$$

where $T_p(\cdot)$ is the p-th augmentation function, $\mathbf{x}^p$ denotes the p-th view of $\mathbf{x}$.

The view bag is constructed by different views of the same original images. It is important to note that a raw examples are used to build a view bag by P augmentation functions. When $P = 1$, the view bag contains only one view, that is, the original example $\mathbf{x}$ is transformed into a view example $\mathbf{x}^1$, which is the same as in (3). In this simplification case, it becomes the traditional data augmented learning method. It will become more effectiveness to measure the similarity between different original examples by measuring the similarity between different view bags.

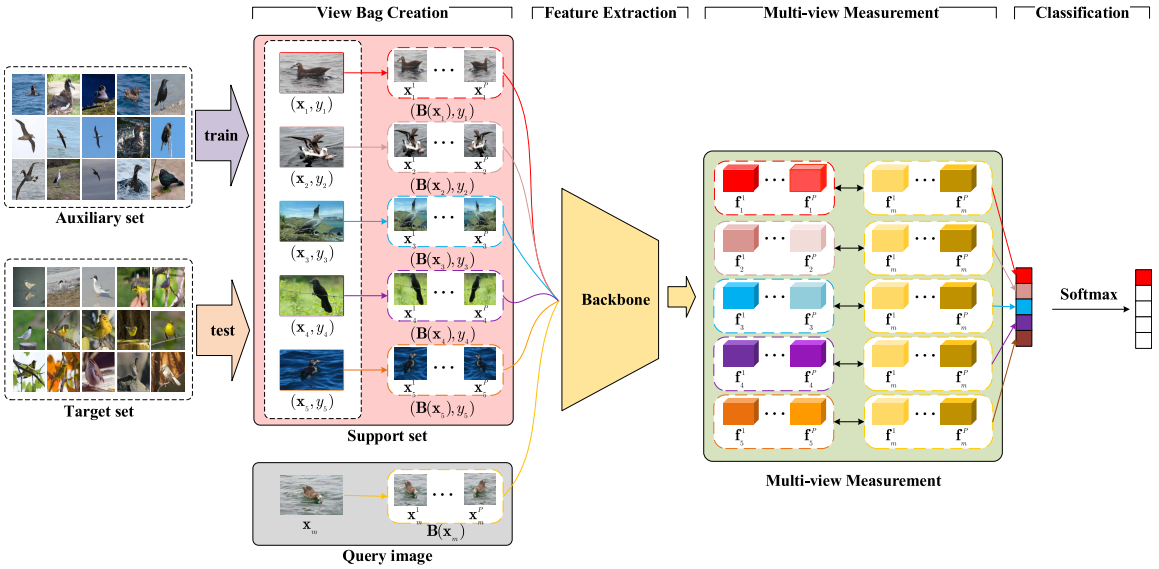


FIGURE 1. Architecture of our MML.

IV. METHOD

The architecture of our MML is shown in Figure 1. It consists of four stages: view bag creation, feature extraction, multi-view metric, and classification. Firstly, an image $\mathbf{x}$ from support set or query set is transformed into a view bag $\mathbf{B}(\mathbf{x})$, and then the feature of the view bag is extracted by the backbone. Finally, the similarity between a query view bag and a support view bag is computed for the next classification.

A. VIEW BAG CREATION

Different from the extensive few-shot fine-grained image classification method, the view bag is used as the input of the backbone network, as shown in Figure 1.

The support set $\mathcal{A}_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{C\times N}$ contains C category and each category contains N images. Each image $\mathbf{x}_i$ generates a labeled view bag $(\mathbf{B}(\mathbf{x}_i), y_i)$ through P data augmentation functions. The view bag set of the support set $\mathcal{A}'_S$ contains a total of C-class examples, and each category contains $N \times P$ examples. It can be expressed as:

$$\mathcal{A}'_S = \{(\mathbf{B}(\mathbf{x}_i), y_i)\}_{i=1}^{C\times N}. \tag{5}$$

Similarly, query set $\mathcal{A}_Q = \{\mathbf{x}_m\}_{m=1}^{M}$ becomes $\mathcal{A}'_Q$ by the same P data augmentation functions:

$$\mathcal{A}'_Q = \{\mathbf{B}(\mathbf{x}_m)\}_{m=1}^{M}, \tag{6}$$

where $\mathbf{x}_m$ denotes the m-th image in $\mathcal{A}_Q$, and $\mathbf{B}(\mathbf{x}_m) = \{\mathbf{x}_m^1, \ldots, \mathbf{x}_m^P\}$ denotes m-th view bag.

B. FEATURE EXTRACTION

Given an image $\mathbf{x}$, its feature is got by $\mathbf{f}_i^p = \varphi(\mathbf{x}_i^p)$, where $\varphi(\cdot)$ represents the mapping function of feature extraction. In practice, the mapping function is the backbone network.

Feature sets $\mathcal{F}_S$ and $\mathcal{F}_Q$ are obtained for $\mathcal{A}'_S$ and $\mathcal{A}'_Q$. They can be obtained by extracting the features of the examples in one view bag through the backbone network.

$$\mathcal{F}_S = \{\mathbf{F}_i\}_{i=1}^{C\times N} = \{\mathbf{f}_i^1, \ldots, \mathbf{f}_i^p, \ldots, \mathbf{f}_i^P\}_{i=1}^{C\times N}, \tag{7}$$

where $\mathbf{F}_i = \{\mathbf{f}_i^1, \ldots, \mathbf{f}_i^p, \ldots, \mathbf{f}_i^P\}$ denotes the feature set of i-th view bag in $\mathcal{A}_S$, $\mathbf{f}_i^p \in \mathbb{R}^{d\times W\times H}$ represents the feature of the i-th image by the p-th augmentation function, and d, W, H represent the channel, width and height of the feature.

$$\mathcal{F}_Q = \{\mathbf{F}_m\}_{m=1}^{M} = \{\mathbf{f}_m^1, \ldots, \mathbf{f}_m^q, \ldots, \mathbf{f}_m^P\}_{m=1}^{M}, \tag{8}$$

where $\mathbf{F}_m = \{\mathbf{f}_m^1, \ldots, \mathbf{f}_m^q, \ldots, \mathbf{f}_m^P\}$ denotes the feature set of m-th view bag in $\mathcal{A}_Q$, $\mathbf{f}_m^q \in \mathbb{R}^{d\times W\times H}$ represents the feature of the m-th image by the q-th augmentation function.

C. MULTI-VIEW METRIC

To more accurately calculate the similarity between query view bag $\mathbf{B}(\mathbf{x})$ and each category, we propose a new multi-view metric method to calculate the similarity between different view bags as the final similarity. To describe our multi-view metric method clearly, the c-th category feature set $\{\mathbf{F}_{(c-1)\times N}, \mathbf{F}_{(c-1)\times N+1}, \ldots, \mathbf{F}_{(c-1)\times N+N}\}$ is rewritten as $\{\mathbf{F}_c^1, \ldots, \mathbf{F}_c^p, \ldots, \mathbf{F}_c^P\}$, where $\mathbf{F}_c^p = \{\mathbf{f}_n^p\}_{n=(c-1)\times N+1}^{(c-1)\times N+N}$ represents the feature set of the p-th augmentation in all N examples of class c. We define a convert function $g(\cdot)$ to transform the feature tensor $\mathbf{f} \in \mathbb{R}^{d\times W\times H}$ into a feature matrix $\bar{\mathbf{f}} = g(\mathbf{f}) \in \mathbb{R}^{d\times L}$, which can be treated as a set of $L(L = W \times H)$ d-dimensional local descriptors.

The similarity between $\mathbf{B}(\mathbf{x}_m)$ and c-th category is calculated based on covariance matrix measurement method [24] and image-to-class measurement method [25], and we noted them as MML(C) and MML(D) respectively.

1) MML(C)

Firstly, a covariance matrix $\mathbf{S}_c^p \in \mathbb{R}^{d\times d}$ is calculated for the feature $\mathbf{F}_c^p$ of the p-th augmentation in the c-th category, which can be treated as the feature center of the p-th augmentation in the c-th category,

$$\mathbf{S}_c^p = \frac{1}{LN-1}\sum_{n=(c-1)\times N+1}^{(c-1)\times N+N}(\bar{\mathbf{f}}_n^p - \tau)(\bar{\mathbf{f}}_n^p - \tau)^T, \tag{9}$$

where $\bar{\mathbf{f}}_n^p = g(\mathbf{f}_n^p) \in \mathbb{R}^{d\times L}$, $\tau \in \mathbb{R}^{d\times L}$ is a matrix of mean vectors, with each of its column the same mean vector of all the $L \times N$ descriptors.

Then, we calculate the similarity between $\mathbf{f}_m^q$ and $\mathbf{S}_c^p$ as

$$sim(\mathbf{f}_m^q, \mathbf{S}_c^p) = w^T diag(\bar{\mathbf{f}}_m^{q\,T} \mathbf{S}_c^p \bar{\mathbf{f}}_m^q), \tag{10}$$

where $\bar{\mathbf{f}}_m^q = g(\mathbf{f}_m^q) \in \mathbb{R}^{d\times L}$, $diag(\bar{\mathbf{f}}_m^{q\,T} \mathbf{S}_c^p \bar{\mathbf{f}}_m^q)$ represents the local similarity between $\mathbf{f}_m^q$ and $\mathbf{S}_c^p$, and $diag(\cdot)$ returns a column vector of the main diagonal elements of a matrix. The similarity $sim(\mathbf{f}_m^q, \mathbf{S}_c^p)$ between the feature $\mathbf{f}_m^q$ and the class center $\mathbf{S}_c^p$ is obtained by a fully-connection layer, and w represents the weight of the layer.

Finally, the similarity between $\mathbf{B}(\mathbf{x}_m)$ and c-th category is calculated as

$$sim(\mathbf{B}(\mathbf{x}_m), c) = \frac{1}{P^2} \sum_{p=1}^{P} \sum_{q=1}^{P} sim(\mathbf{f}_m^q, \mathbf{S}_c^p), \tag{11}$$

where $sim(\mathbf{B}(\mathbf{x}_m), c)$ averages the P^2 similarities between $\mathbf{f}_m^q$ and $\mathbf{S}_c^p$.

2) MML(D)

For the p-th augmentation of the m-th query image $\mathbf{x}_m$, it can be transformed into a feature $\mathbf{f}_m^q = [\mathbf{f}_m^q(1), \ldots, \mathbf{f}_m^q(L)] \in \mathcal{R}^{d\times L}$, where each $\mathbf{f}_m^q(i)$ is a deep local descriptor. For each descriptor $\mathbf{f}_m^q(i)$, we can find its k-nearest neighbors $\mathbf{f}_c^p(i,j)|_{j=1}^k$ in the p-th augmentation of the c-th category. Then we compute the similarity between $\mathbf{f}_m^q(i)$ and each $\mathbf{f}_c^p(i,j)$, and sum the $L \times k$ similarities as the similarity between query $\mathbf{f}_m^q$ and the p-th augmentation of category c.

Mathematically, the measure can be expressed as

$$sim(\mathbf{f}_m^q, \mathbf{f}_c^p) = \sum_{i=1}^{L} \sum_{j=1}^{k} cos(\mathbf{f}_m^q(i), \mathbf{f}_c^p(i,j)), \tag{12}$$

where $cos(\cdot)$ represents the cosine similarity measurement.

Finally, we obtain the final similarity by summing the similarities between query $\mathbf{f}_m^q$ and category c for all possible P transformations.

$$sim(\mathbf{B}(\mathbf{x}_m), c) = \frac{1}{P^2} \sum_{p=1}^{P} \sum_{q=1}^{P} sim(\mathbf{f}_m^q, \mathbf{f}_c^p), \tag{13}$$

D. CLASSIFICATION

The class prediction probability p_m^c of the $\mathbf{B}(\mathbf{x}_m)$ is generated by softmax operation

$$p_m^c = \frac{\exp(sim(\mathbf{B}(\mathbf{x}_m), \mathcal{S}_c))}{\sum_{j=1}^{C} \exp(sim(\mathbf{B}(\mathbf{x}_m), \mathcal{S}_j))}, \tag{14}$$

where $\mathcal{S}_j$ is the j-th category, C is the number of classes.

The cross-entropy loss is adopted to supervise the training, and the loss L is calculated as follows

$$L = -\frac{1}{M} \sum_{m=1}^{M} y_m \log \hat{y}_m, \tag{15}$$

TABLE 1. The details of datasets.

Datasets	all classes	train classes in $\mathcal{A}$	test classes in $\mathcal{A}$	classes in $\mathcal{T}$
CUB-200-2011	200	130	20	50
Stanford-Dogs	120	70	20	30
Stanford-Cars	196	130	17	49

where $\hat{y}_m = \arg\max_{c=1,\ldots,C} p_m^c$ is the predicted label of m-th examples, y_m is the ground-truth label of m-th examples, M is the number of training examples. As $\mathbf{B}(\mathbf{x}_m)$ has the same label as $\mathbf{x}_m$, the classification result $\hat{y}_m^c$ is the result of $\mathbf{x}_m$.

V. EXPERIMENTS

Experiments are conducted on three widely-used fine-grained datasets, including CUB-200-2011 [7], Stanford-Dogs [8], and Stanford-Cars [9]. Each dataset is divided into auxiliary set and target set in the same way as methods [24]–[27], among which the auxiliary set is divided into training set and validation set. The statistical information of datasets is shown in Table 1.

Our method is implemented in PyTorch on one NVIDIA 2080Ti GPU. During the training process on each dataset, the view bag was constructed by random cropping, horizontal flipping and color enhancement, and the processed image size was 84 × 84. Meanwhile, the model adopted Conv4-64 as backbone and was trained by episodic training mechanism, and 250,000 episodic were trained in total. In the 5-way 1-shot classification task, each episodic contains 5 classes, and each class contains 1 labeled support image and 15 query images. Similarly, in the 5-way 5-Shot classification task, each episodic contains 5 classes, and each class contains 5 labeled support images and 15 query set images. Besides, we adopt Adam algorithm with an initial learning rate of 0.005 to optimize our model, where the learning rate is reduced by half for every 100,000 episodes. In the test process on all three datasets, 600 episodes were randomly formed from the test set for testing to calculate the top-1 mean accuracy as well as the corresponding confidence interval.

A. COMPARISON WITH STATE-OF-THE-ART METHODS

Our method was compared with state-of-the-art methods on three datasets. The results are shown in Table 2, where the optimal value and the sub-optimal value are highlighted in **bold** and underline respectively. Our MML(C) and MML(D) calculate the similarity based on covariance matrix measurement method [24] and image-to-class measurement method [25] respectively.

From Table 2, we can find that our method achieves 71.61±0.87%, 57.78±0.96% and 74.02±0.84% on CUB-200-2011, Stanford-Dogs and Stanford-Cars in 5-way 1-shot classification task respectively, and 88.72±0.51%, 76.30±0.68% and 92.94±0.37% in 5-way 5-shot classification task, which are the best results. As can be seen from Table 2, on the three public datasets of CUB-2011-200, Stanford-Dogs and Stanford-Cars, in 5-way 1-shot setting, our MML (C) increased by 17.63%, 8.68% and

TABLE 2. Comparison with state-of-the-art methods.

method	augmentation	5-way 1-shot			5-way 5-shot		
		CUB-2011-200	Stanford-Dogs	Stanford-Cars	CUB-2011-200	Stanford-Dogs	Stanford-Cars
MatchingNets [12]	✓	45.30±1.03	35.80±0.99	34.80±0.98	59.50±1.01	47.50±1.03	44.70±1.03
Prototypical [13]	✓	37.36±1.00	37.59±1.00	40.90±1.01	45.28±1.03	48.19±1.03	52.93±1.03
GNN [37]	✓	51.83±0.98	46.98±0.98	55.85±0.97	63.69±0.94	62.27±0.95	71.33±0.62
LRPABNcpt [26]	✓	63.63±0.77	45.72±0.75	60.28±0.76	76.06±0.58	60.94±0.66	73.29±0.58
PABN [27]	✓	66.71±0.43	55.47±0.46	56.80±0.45	76.81±0.21	66.65±0.23	68.78±0.22
DN4-DA [25]	✓	53.15±0.84	45.73±0.76	61.51±0.85	81.90±0.60	66.33±0.66	89.60±0.44
BSnet(D&C) [28]	✓	67.58±0.95	-	57.16±0.97	83.43±0.64	–	86.17±0.57
CovaMNet [24]	–	52.42±0.76	49.10±0.76	56.65±0.86	63.76±0.64	63.04±0.65	71.33±0.62
MattML [36]	✓	66.29±0.56	54.84±0.53	66.11±0.54	80.34±0.30	71.34±0.38	82.80±0.28
VFD [38]	✓	68.42±0.92	57.03±0.86	–	82.42±0.61	73.00±0.66	–
our MML(C)	✓	70.05±0.92	**57.78±0.96**	71.48±0.91	83.63±0.63	72.64±0.73	85.26±0.53
our MML(D)	✓	**71.61±0.87**	57.39±0.92	**74.02±0.84**	**88.72±0.51**	**74.33±0.67**	**92.94±0.37**

14.83% respectively compared with CovaMNet [24], and MML (D) increased by 18.64%, 11.66% and 12.51% respectively compared with DN4-DA [25]. In 5-way 5-shot setting, MML (C) increased by 19.87%, 9.6% and 13.93% respectively compared with CovaMNet [24], and MML (D) increased by 6.82%, 8% and 3.44% respectively compared with DN4-DA [25]. From the experimental results, it can be seen that compared with the methods in [24] or [25] that only measures the single-view distance as the final similarity, we adopt the measurement methods of [24] or [25] as the atomic distance measurement, which can effectively improve the classification accuracy. The proposed multi-view metric method is to computes the similarities between two examples from multiple views, and synthesize these similarities as the final similarity, which has more robust distance measurement ability.

It is worth noting that DN4-DA, BSnet(D&C), MattML and PABN all use data augmentation, including random cropping, color jittering and horizontal flipping. From Table 2, it can be seen that our method achieves the best results under the same data augmentation. What's more, DN4-DA, BSnet(D&C) and MML(D) all use image-to-class measurement method [25] to calculate the similarity between views. DN4-DA uses the traditional data augmentation for training, while BSnet(D&C) uses both cosine measurement method and image-to-class measurement method to calculate the similarity between examples, and our MML(D) adopts the proposed multi-view metric method. The result shows that our method is the best among all classification tasks, which verifies that it can obtain more efficient similarity measurement.

B. ABLATION STUDIES

1) MULTI-VIEW METRIC

In order to verify the validity of the proposed multiple view metric method, on the basis of CovaMNet, we use different measurement methods to conduct comparative experiments. The experimental method adopts the same data augmentation to train model, and the results are shown in Figure 2, where ''Benchmark'' means the result without data augmentation, ''Traditional'' means that multiple data augmentation methods are used to process the original example, and the single-view measurement is used to compute the similarity, ''Multi-view Metric'' means that our multi-view measurement is used to calculate the similarity.

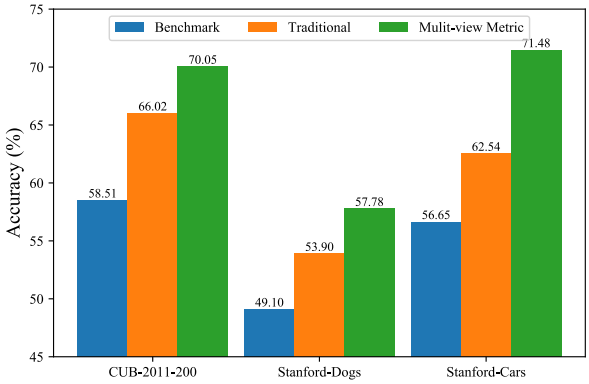


FIGURE 2. Comparison of different measurements.

As can be seen from Figure 2, the ''Traditional'' method can significantly improve the accuracy by 7.51%, 4.80% and 5.89% on three datasets, respectively. This is because the original data can generate various views after data augmentation, which expand the training example and improves the robustness and generalization ability of the learnt model. However, under the same data augmentation, the accuracy of the ''Multi-view Metric'' method is higher than that of the ''Traditional'' method. On the three datasets, the accuracy of the ''Multi-view Metric'' method is 11.54%, 8.68% and 8.94% higher than that of the ''Benchmark'' method, and 4.03%, 3.88 and 8.94% higher than that of the ''Traditional'' method. This is because the ''Traditional'' method only uses a single view of each example in each iteration, although various data augmentation are used to augment the original example to obtain different views, such as random cropping. On the contrary, although ''Multi-view Metric'' also augment the original example, it takes all views as a view bag instead of a single example as input. When measuring the similarity between view bags, all the similarities between views are integrated as the final similarity. Therefore, compared with the ''Traditional'' method, ''Multi-view Metric'' method can measure the similarity between examples more accurately.

2) DATA AUGMENTATION

To verify data augmentation for few-shot fine-grained image classification task, we adopt four kinds of data augmentations to augment examples, including horizontal flipping, vertical flipping, color jittering and random cropping, and use the ''Traditional'' method and our ''Multi-view Metric'' method to calculate the similarity between examples. It is important to note that horizontal flipping and vertical flipping are conducted in accordance with a certain probability, so there is two kinds of training examples, the original image and its augmented example, in the whole training process. Therefore, during the construction of view bags, we adopt the original image and the augmented example to construct the view bag for training, that is, the view bag contains the original image and the augmented example. The experimental results are shown in Figure 3. It can be found that regardless of the ''Traditional'' method or our ''Multi-view Metric'' method, the improvement of random cropping for data augmentation is the largest.

In order to better analyze the effects of different data augmentation, Figure 4 shows the results of original examples processed by different data augmentation. It can also be obtained from the analysis of the original image that the proportion of the object in the whole image is small in the fine-grained image, and the random cropping method can remove part of the background and improve the proportion of the object, so it can better extract the features of the object. The second best is the horizontal flipping. The examples obtained by horizontal flipping are more consistent with the original example and improve the robustness of the model. As can be seen from Figure 3, the result of our ''Multi-view Metric'' is similar to that of ''Traditional'' method when color jittering is used. This is because color jittering changes the color of the original example by adjusting the image's saturation, contrast and brightness, and the feature variation in the regional area after processing are not obvious. In addition, on Stanford-Cars, when the vertical flipping method was used, the accuracy of the ''Traditional'' method decreases while our ''Multi-view Metric'' method and ''Benchmark'' method have the similar accuracy. This is because our ''Multi-view Metric'' method compute both the similarity between the vertical flipped images and the similarity between the original images, to reduce the negative impact of vertical flipping. Therefore, compared with the ''Traditional'' method, our ''Multi-view Metric'' method can effectively reduce the impact of negative examples in data augmentation, make full use of the augmented examples, and greatly improve the robustness of the learnt model.

3) DATA AUGMENTATION COMBINATION

In order to verify the impact of the combination of different types of view bags on support set or query set, different combinations are constructed for comparison. According to the analysis results in sub-section V-B2, different combinations of three data augmentation methods, including random cropping, horizontal flipping and color jittering, were selected to construct view bags of support set or query set respectively, and 5-way 1-shot classification task was carried out on the CUB-2011-200 dataset. The results are shown in Table 3.

TABLE 3. Comparison of different data augmentation combination.

Support Set (P=2)	Query Set (Q=2)	Accuries
HP&CJ	HP&CJ	66.66±0.92
RC&CJ	RC&CJ	66.87±0.92
HP&RC	HP&RC	69.73±0.97
Support Set (P=3)	Query Set (Q=1)	Accuries
RC&HP&CJ	RC	68.86±0.96
RC&HP&CJ	HP	64.97±0.85
RC&HP&CJ	CJ	63.17±0.91
Support Set (P=3)	Query Set (Q=2)	Accuries
RC&HP&CJ	RC&HP	69.83±0.94
RC&HP&CJ	HP&CJ	67.15±0.91
RC&HP&CJ	RC&CJ	68.89±0.88
Support Set (P=3)	Query Set (Q=3)	Accuries
RC&HP&CJ	RC&HP&CJ	70.05±0.92

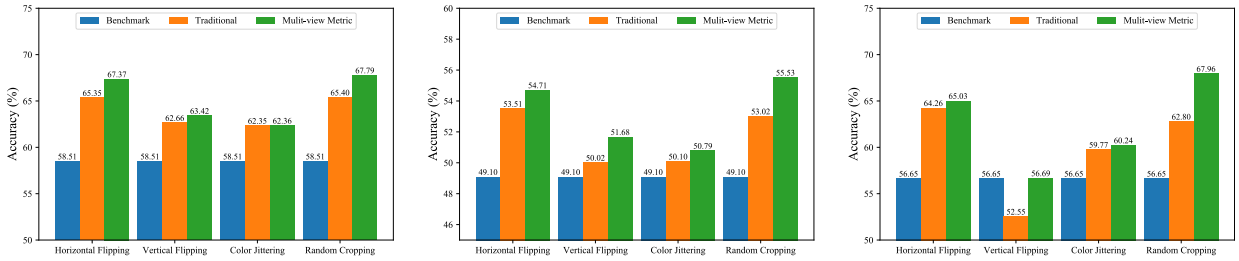
TABLE 4. Results with ResNet-12 backbone.

method	5-way 1-shot	
	CUB-2011-200	Stanford-Dogs
Baseline [39]	63.90 ± 0.88	63.53 ± 0.89
Baseline++ [39]	68.46 ± 0.85	58.30 ± 0.35
MatchingNet [12]	72.62 ± 0.90	65.87 ± 0.81
VFD [38]	79.12 ± 0.83	76.24 ± 0.87
our MML(C)	68.85 ± 0.74	55.94 ± 0.78
method	5-way 5-shot	
	CUB-2011-200	Stanford-Dogs
Baseline [39]	82.54 ± 0.54	79.95 ± 0.59
Baseline++ [39]	81.02 ± 0.46	73.77 ± 0.68
MatchingNet [12]	84.14 ± 0.50	80.70 ± 0.42
VFD [38]	91.48 ± 0.39	88.00 ± 0.47
our MML(C)	78.94 ± 0.53	69.74 ± 0.69

In Table 3, P and Q represent the number of data augmentation contained in the view bag of the support set or query set respectively. ''RC'' represents the transformation obtained by random cropping, ''HP'' represents the transformation obtained by horizontal flipping, and ''CJ'' represents the transformation obtained by color jittering. It can be seen from Table 3 that the results are positively correlated with the number of data augmentation types, that is, with the increase of data augmentation types, the accuracy is constantly improved. When a view bag contained three types of transformation (P = 3, Q = 3), the best accuracy (70.43%) is achieved. It is verified that the combination of multiple data augmentation types can calculate the similarity between examples more robustly.

C. DISCUSSION

Our backbone network is Conv4-64, which consists of 4 convolution layers with 64/64/64/64 filters. To comparing with ResNet-12 backbone as in [38], we replace Conv4-64 with ResNet-12 and reported the results in Table 4. From Table 4, we can find that our MML(C) has a big performance gap with [38] and [12] when we use the ResNet-12 backbone. The main reason is that our multi-view metric is inefficient in deal with the output feature of the ResNet-12 backbone.



(a) (b) (c)

FIGURE 3. Comparison of different data augmentation.

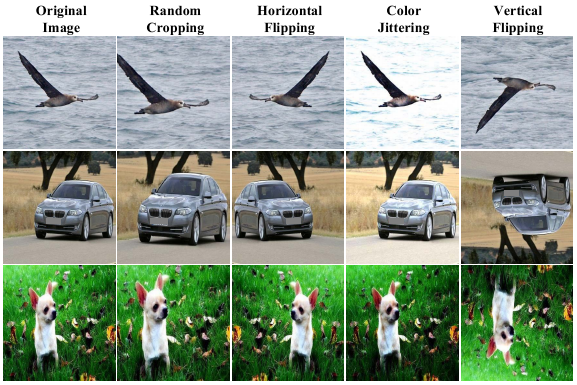


FIGURE 4. Results of different data augmentation.

Since the ResNet-12 backbone contains 4 down-sampling convolution operations, it make the size of the output feature maps only 5×5. The computation of the covariance matrix (Eq. 9) or the k-nearest neighbor similarity (Eq. 12) does not have good representation ability, because the few local deep descriptors of each category is hard to calculate a effective covariance representation or find the effective k-nearest neighbors for this category. Comparing with the ResNet-12 backbone, the Conv4-64 backbone has only 2 max-pooling operations and outputs the feature maps of 21×21 size, which make the covariance matrix or the k-nearest neighbor similarity have strong representation ability.

VI. CONCLUSION

In this paper, a multi-view metric learning method is proposed to comprehensively measure the similarity from a new viewpoint by making full use of the original example and its data augmentations. For 5-way 1-shot and 5-way 5-shot classification tasks, on three public fine-grained datasets, experimental results show that our MML achieves the state-of-the-art accuracies. It improves the robustness of the learnt model by measuring the similarity between our proposed view bags.

However, our method focuses on the view bag and its similarity measurement, neglecting to mining the regional discriminative parts. In the next step, attention mechanism or part detection modules will be further studied to extract the discriminative feature representation to improve the accuracies.

REFERENCES

[1] P. Shroff, T. Chen, Y. Wei, and Z. Wang, "Focus longer to see better: Recursively refined attention for fine-grained image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2020, pp. 3791–3798.

[2] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proc. AAAI*, New York, NY, USA, Feb. 2020, pp. 13130–13137.

[3] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4683–4695, 2020, doi: 10.1109/TIP.2020.2973812.

[4] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *Proc. ECCV*, Glasgow, U.K., Aug. 2020, pp. 153–168.

[5] H. Hanselmann and H. Ney, "Fine-grained visual classification with efficient end-to-end localization," 2020, *arXiv:2005.05123*.

[6] Y. Gao, X. Han, X. Wang, W. Huang, and M. R. Scott, "Channel interaction networks for fine-grained image categorization," in *Proc. AAAI*, New York, NY, USA, Feb. 2020, pp. 10818–10825.

[7] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "TheCaltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[8] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *Proc. CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 1–2.

[9] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Sydney, NSW, Australia, Dec. 2013, pp. 1–8.

[10] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.

[11] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006, doi: 10.1109/TPAMI.2006.79.

[12] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016, pp. 3630–3638.

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

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

[15] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *Proc. ICLR*, New Orleans, LA, USA, May 2019, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=SyVuRiC5K7

[16] 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)*, Snowmass Village, CO, USA, Mar. 2020, pp. 2207–2216.

[17] Z. Wu, Y. Li, L. Guo, and K. Jia, "PARN: Position-aware relation networks for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6658–6666.

[18] H. Xu, K. Zhang, and W. Wang, "A feature selection method for small samples," *J. Comput. Res. Develop.*, vol. 55, no. 10, pp. 2321–2330, Oct. 2018, doi: 10.7544/issn1000-1239.2018.20170748.
[19] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Proc. NeurIPS*, Vancouver, BC, Canada, Dec. 2019, pp. 4005–4016.
[20] J. Zhao, X. Lin, J. Zhou, J. Yang, L. He, and Z. Yang, "Knowledge-based fine-grained classification for few-shot learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, London, U.K., Jul. 2020, pp. 1–6.
[21] K. Yan, Z. Bouraoui, P. Wang, S. Jameel, and S. Schockaert, "Few-shot image classification with multi-facet prototypes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 1740–1744.
[22] X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, and Q. Li, "Few-shot learning for domain-specific fine-grained image classification," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3588–3598, Apr. 2021, doi: 10.1109/tie.2020.2977553.
[23] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *Proc. ICLR*, New Orleans, LA, USA, May 2019, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=BJgklhAcK7
[24] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proc. AAAI*, Honolulu, HI, USA, Jan. 2019, pp. 8642–8649.
[25] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7260–7268.
[26] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Trans. Multimedia*, vol. 23, pp. 1666–1680, Jul. 2021, doi: 10.1109/TMM.2020.3001510.
[27] H. Huang, J. Zhang, J. Zhang, Q. Wu, and J. Xu, "Compare more nuanced: Pairwise alignment bilinear network for few-shot fine-grained learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Shanghai, China, Jul. 2019, pp. 91–96.
[28] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, and J.-H. Xue, "BSNet: Bi-similarity network for few-shot fine-grained image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 1318–1331, 2021, doi: 10.1109/TIP.2020.3043128.
[29] S. M. Kye, H. B. Lee, H. Kim, and S. J. Hwang, "Meta-learned confidence for few-shot learning," 2020, *arXiv:2002.12017*.
[30] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11–20.
[31] 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, doi: 10.1109/TIP.2019.2924811.
[32] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 1005–1014.
[33] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6598–6607.
[34] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1235–1255, Sep. 2019, doi: 10.1007/s11263-019-01176-2.
[35] X. He, Y. Peng, and J. Zhao, "Fast fine-grained image classification via weakly supervised discriminative localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1394–1407, May 2019, doi: 10.1109/TCSVT.2018.2834480.
[36] Y. Zhu, C. Liu, and S. Jiang, "Multi-attention meta learning for few-shot fine-grained image recognition," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1090–1096.
[37] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *Proc. ICLR*, Vancouver, BC, Canada, Apr. 2018, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=BJj6qGbRW
[38] J. Xu, H. Le, M. Huang, S. Athar, and D. Samaras, "Variational feature disentangling for fine-grained few-shot classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 8792–8801.
[39] 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. ICLR*, New Orleans, LA, USA, May 2019, pp. 1–11. [Online]. Available: https://openreview.net/pdf?id=HkxLXnAcFQ

ZHUANG MIAO received the Ph.D. degree from the PLA University of Science and Technology, Nanjing, China, in 2007. He is currently a Professor with the Army Engineering University of PLA, Nanjing. His current research interests include artificial intelligence, pattern recognition, and computer vision.

XUN ZHAO received the B.S. degree from the Army Engineering University of PLA, Nanjing, China, in 2019, and the M.S. degree from the Command and Control Engineering College, Army Engineering University of PLA, in 2021. His current research interests include image classification and deep learning.

JIABAO WANG received the Ph.D. degree in computational intelligence from the PLA University of Science and Technology, Nanjing, China, in 2013. He is currently an Associate Professor with the Army Engineering University of PLA, Nanjing. His current research interests include computer vision and machine learning.

BO XU received the Ph.D. degree from the PLA University of Science and Technology, Nanjing, China, in 2011. He is currently an Associate Professor with the Army Engineering University of PLA, Nanjing. His current research interests include cyberspace security and computer networks.

YANG LI received the M.S. degree from the PLA University of Science and Technology, Nanjing, China, in 2010, and the Ph.D. degree from the Army Engineering University of PLA, Nanjing, in 2018. He is currently an Associate Professor with the Army Engineering University of PLA, Nanjing. His current research interests include computer vision, deep learning, and image processing.

HANG LI received the Ph.D. degree in computer science and technology, in 2018. He is currently a Postdoctoral Researcher with the Army Engineering University of PLA. His research interests include computer vision and information fusion.