Not All Instances Contribute Equally: Instance-Adaptive Class Representation Learning for Few-Shot Visual Recognition

Mengya Han, Yibing Zhan, Yong Luo, Bo Du, Han Hu, Yonggang Wen, Fellow, IEEE, and Dacheng Tao, Fellow, IEEE

Abstract—Few-shot visual recognition refers to recognize novel visual concepts from a few labeled instances. Many few-shot visual recognition methods adopt the metric-based meta-learning paradigm by comparing the query representation with class representations to predict the category of query instance. However, the current metric-based methods generally treat all instances equally and consequently often obtain biased class representation, considering not all instances are equally significant when summarizing the instance-level representations for the class-level representation. For example, some instances may contain unrepresentative information, such as too much background and information of unrelated concepts, which skew the results. To address the above issues, we propose a novel metric-based meta-learning framework termed instance-adaptive class representation learning network (ICRL-Net) for few-shot visual recognition. Specifically, we develop an adaptive instance revaluation network (AIRN) with the capability to address the biased representation issue when generating the class representation, by learning and assigning adaptive weights for different instances according to their relative significance in the support set of corresponding class. In addition, we design an improved bilinear instance representation and incorporate two novel structural losses, i.e., intraclass instance clustering loss and interclass representation distinguishing loss, to further regulate the instance revaluation process and refine the class representation. We conduct extensive experiments on four commonly adopted few-shot benchmarks: miniImageNet, tieredImageNet, CIFAR-FS, and FC100 datasets. The experimental results compared with the state-of-the-art approaches demonstrate the superiority of our ICRL-Net.

Index Terms—Few-shot, instance-adaptive, meta-learning, relative significance, visual recognition.

I. INTRODUCTION

DEEP learning models have achieved the state-of-the-art performance on various visual tasks, including image classification [1], [2], [3], object detection [4], [5], and segmentation [6]. However, most deep models have numerous parameters and require large amounts of labeled data for training. In most cases, obtaining abundant labeled data is time-consuming and laborious. The human annotation cost and data scarcity in some classes (e.g., rare species) significantly limit the applicability of current vision systems to learn new visual concepts efficiently. In contrast, learning from extremely few labeled instances is an important ability for humans. It is thus of great interest to develop machine learning algorithms that recognize new visual categories from only a limited amount of labeled instances for each novel category. The problem of learning to recognize unseen classes from limited instances, known as few-shot learning (FSL) [7], [8], has attracted increasing attention recently.

Few-shot visual recognition [9], [10], [11], [12], [13], [14], [15], [16], [17], as a specific FSL problem, attempts to learn a classifier with good generalization ability for novel visual concepts, each of which contains only a few labeled instances (support instances). To address this problem, a variety of solutions have been proposed. One mainstream solution is meta-learning [18], [19], [20], [21], where a series of independent few-shot tasks are used to learn a general model during training, and then the general model is applied to unseen target tasks in the testing phase. In this way, the learned meta-models solve a new task using the knowledge acquired from many similar tasks. Generally, the meta-learning methods for few-shot visual recognition can be classified into two categories: optimization-based and metric-based. The former [22], [23], [24] automatically learns a set of model parameters, such as general initialization conditions, learning rates, and parameter updating strategies. The latter [16], [25], [26] aims to learn a discriminative embedding space, in which the representations of different instances and classes can be...
The main contributions of this article are as follows. The optimization-based methods [19], [22], [23], [39], [40], [41] aim to find a single set of model parameters that can be adapted with a few steps of gradient descent to target tasks. For example, the well-known model-agnostic meta-learning (MAML) approach [22] meta-learns a good initial condition (a set of neural network weights), which enables the model to quickly adapt to new tasks. The few-shot optimization

2) We develop an adaptive instance revaluing network (AIRN) to alleviate the issue of bias in class representation. Specifically, the AIRN assigns different weights (or values) as the relative significance of instances by considering all the support instances in the same classes and obtains the class representation by averaging the revalued instance representations.

3) We design a joint loss function to improve the instance adaptive revaluation process and refine the discriminative representation space of metric-based meta-learning. The designed joint loss function contains three components: a commonly adopted classification loss for few-shot visual recognition and two newly designed structural losses for robust representation: intraclass instance clustering loss and interclass representation distinguishing loss.

We conducted extensive experiments on four popular few-shot benchmarks: miniImageNet [18], tieredImageNet [30], CIFAR-FS [31], and FC100 [32] datasets. The experimental results indicate that our ICRL-Net outperforms the state-of-the-art approaches. In specific, we achieved a significant 1.5% relative improvement compared with the most competitive counterpart. We further visualize the learned attention parts of ICRL-Net, and the results show that our ICRL-Net can appropriately identify the relative significance between instances in the same support set.

The rest of this article is organized as follows. Section II briefly introduces the related works. The details of ICRL-Net are depicted in Section III. Section IV provides the experimental analysis, and we conclude this article in Section V.

II. RELATED WORK

This section briefly introduces FSL and the attention mechanism.

A. Few-Shot Learning

FSL is an emerging research topic that aims to learn a model from a set of data (base classes) and adapt the model to a disjoint set (new classes) with limited training data [33], [34]. Up to now, few-shot visual recognition [35], [36], [37], [38], which aims to recognize novel visual categories from limited labeled instances, has received great attentions in FSL. Earlier work on FSL tended to involve generative models with complex iterative inference strategies [8]. Most of the recent FSL approaches follow the meta-learning paradigm, which is usually performed by training a meta-learner that learns the transferable knowledge from similar tasks and then generalizes to new tasks. Under this paradigm, various meta-learning methods for FSL are developed and can be roughly classified into two categories: optimization-based methods and metric-based methods.

The optimization-based methods [19], [22], [23], [39], [40], [41] aim to find a single set of model parameters that can be adapted with a few steps of gradient descent to target tasks. For example, the well-known model-agnostic meta-learning (MAML) approach [22] meta-learns a good initial condition (a set of neural network weights), which enables the model to quickly adapt to new tasks. The few-shot optimization
approach Meta-long short-term memory (LSTM) [19] goes further to learn not only a good initial condition but also an LSTM-based meta-learner that is used to learn appropriate parameter updating rules. Meta-SGD [39] further improves the meta-learning ability by learning the parameter initialization, gradient update direction, and learning rate within a single step. Although effective sometimes, using only a few instances to compute gradients in a high-dimensional parameter space could make generalization difficult. This issue is addressed by latent embedding optimization (LEO) [23], where a low-dimensional latent embedding of model parameters is learned and optimization-based meta-learning is performed in the embedding space. It turns out that the performance of meta-learning in the low-dimensional parameter space is much better than that of meta-learning in the high-dimensional space. Meta-transfer learning (MTL) [40] leverages the idea of transferring pretrained weights and learns to effectively transfer large-scale pretrained deep neural network weights for solving few-shot tasks. Overall, the approaches mentioned above still need to be fine-tuned on the target tasks. In contrast, the metric-based methods solve target tasks without any model updates, thus avoiding gradient computation during testing.

The metric-based methods [18], [27], [28], [32], [42], [43], [44], [45] aim to learn a discriminative representation space, in which the distances between samples should be small in the same class, and large otherwise. The classification is performed in the space by simply finding the nearest neighbor of the query. For example, Koch et al. [42] calculate correlations between input instances via supervised metric based on the Siamese neural networks, and then predict the most relevant classes for query instances. Vinyals et al. [18] design a matching network that introduces an episodic training strategy for FSL and trains a neural network to embed examples. In addition, an attention mechanism was used over the learned representations of the support set to predict the labels of the query set, which can be interpreted as a weighted nearest neighbor classifier. The popular prototypical network [27] is built upon [18], which takes a class’s prototype to be the mean of its support set in the representation space. Then, it calculates the distances between the class prototypes and the query representation to predict the category for a query instance. Inspired by semisupervised clustering, Ren et al. [30] propose an extension of the prototypical network, which uses massive unlabeled instances to generate refined prototypes. These approaches focus on learning representations for data such that they can be recognized with a fixed metric [18], [27] or linear [27], [42] classifier. In the relationship network [28], a deep distance metric is learned for comparing the relationship between the query instances and the support instances. The metric is learnable and equivalent to a nonlinear classifier. The cross attention network [46] is developed based on the relationship network, where an attention module is designed to highlight the correct region of interest in query instance to help classification. Meta-Baseline [29] further improves the ability of the metric-based methods by pretraining a classifier on all the base classes and meta-learning on a nearest-centroid-based few-shot classification algorithm. Few-shot embedding adaptation w/ transformer (FEAT) [47] takes advantage of the set-to-set function to generate task-adaptive feature representations. Meta-confidence transduction (MCT) [48] meta-learns the confidence for each query sample and then updates the class prototypes for each transduction step using all the query examples with meta-learned scores. Rectified metric propagation (ReMP) [49] proposes to refine the prototypes by considering the similarity information of the support set and query set to rectify the metric space, which aims to reduce the metric inconsistently between the training and testing phases. Instance credibility inference (ICI) [50], [51] assumes that not every unlabeled query instance is equally important and aims to exploit the unlabeled instances to augment the training set. It measures the credibility of pseudolabeled examples and selects the most trustworthy pseudolabeled samples according to their credibility as augmented labeled instances. However, most of the existing metric-based methods generally ignore the negative influence of support instances that contain too much interference information and often obtained biased class representations [27], [29]. To address this issue, we propose a metric-based meta-learning method, ICRL-Net, which refines the class representation by revaluing the significance of different support instances based on the attention mechanism.

B. Attention Mechanism

The attention mechanism can focus on the discriminative area adaptively and has been widely exploited for various tasks [3], [46], [52]. For example, Hu et al. [3] propose the squeeze-and-excitation network (SENet) to weight each channel of the feature map. Woo et al. [53] propose the convolutional block attention module (CBAM), which uses hybrid spatial and channel features for attention design. These attention blocks focus on either the channel encoding or the spatial context connection. In addition, there are other forms of attention mechanisms, such as graph attention [54] and self-attention [55], [56]. In FSL, there are many works [18], [46], [47], [57], [58] that also adopt the attention mechanism and achieved excellent performance. For example, in the matching network [18], the attention mechanism is used together with the softmax function to fully specify the prediction of the meta-learner classifier. Similar to the matching network, the cross attention network [46] models the semantic dependency between support instances and query instances. Hence, the relevant regions on the query instances are adaptively localized so that the discrimination ability of embedding features can be improved. In general, the matching network focuses on embeddings, while the cross attention network manipulates feature maps. FEAT [47] uses the self-attention [55] to generate task-adaptive feature representations. In contrast to these works, our developed AIRN mainly focuses on learning robust class prototypes by identifying important support instances and suppressing irrelevant information.

III. INSTANCE-ADAPTIVE CLASS REPRESENTATION LEARNING

This section depicts our ICRL-Net.

A. Problem Formulation

We first introduce the problem formulation of FSL. Let $S$ denote the support set, which contains $N$ classes and $K$ labeled
In few-shot learning (FSL), the episodic training strategy [18] is often used to obtain a well-trained model by sampling few instances for every episode (an individual \(N\)-way \(K\)-shot task) under the meta-learning paradigm. We follow the episodic training strategy, which efficiently learns transferable knowledge from a relatively large labeled dataset \(D_{\text{train}}\) that contains a set of classes \(C_{\text{train}}\). Then, the trained model is applied to a novel testing dataset \(D_{\text{test}}\) that contains a set of classes \(C_{\text{test}}\). There are only a few labeled instances provided for each category in \(C_{\text{test}}\), and \(C_{\text{train}} \cap C_{\text{test}} = \emptyset\). The process of training is conducted on a series of episodes. In each episode, a small subset of \(N\) classes are sampled from \(C_{\text{train}}\) to construct an \(N\)-way \(K\)-shot task: a support set \(S = \{(x_n^k, y_n^k)\mid n = 1, \ldots, N; k = 1, \ldots, K\}\) and a query set \(Q = \{(q_m^m, y_m^m)\mid n = 1, \ldots, N; m = 1, \ldots, M\}\), where \(N\) is the number of classes in each episode, \(K\) is the number of support instances in each class, and \(M\) is the number of query instances in each class. In each episode, the model is trained by minimizing the prediction loss of the query set \(Q\).

In testing, the generalization performances of the learned models are measured on the testing set episodes, each of which consists of a support set \(S\) and a query set \(Q\), where \(S\) and \(Q\) are sampled from \(D_{\text{test}}\) that contains classes distinct from those used in \(D_{\text{train}}\). The instance labels in the support set are known, while those in the query set are unknown and used only for evaluation. The predicted label of the query instance is given by taking the class that has the highest classification score.

Fig. 2 presents the overview of our ICRL-Net. As shown in Fig. 2, ICRL-Net consists of three modules: an attentional bilinear feature extractor to extract instance-level representations, an AIRN to learn the relative significance of support instances in the same class and accordingly obtains the class-level representations and a classifier to classify each query instance based on the learned class-level representations. In addition, ICRL-Net is trained based on a designed joint loss, which contains a commonly adopted classification loss and two newly designed structural losses: intraclass instance clustering loss and interclass representation distinguishing loss. More details of the different components and losses are explained as follows.

C. Attentional Bilinear Feature Extractor

The metric-based methods [18], [27], [29] perform classification by comparing the (support and query) instance representations, and thus, it is critical to learn discriminative instance representation. It is well-known that exploiting higher-order information [59], [60], [61] such as by utilizing bilinear pooling, can improve the discriminative capability of feature representations compared with the low-order information [62]. Several few-shot works recently exploited a variety of variants of bilinear pooling to learn the better feature representation. For example, the second-order similarity network [63] leverages second-order pooling (i.e., homogeneous bilinear pooling) to learn the second-order statistics for similarity pooling, can improve the discriminative capability of feature representations compared with the low-order information [62].

Several few-shot works recently exploited a variety of variants of bilinear pooling to learn the better feature representation. For example, the second-order similarity network [63] leverages second-order pooling (i.e., homogeneous bilinear pooling) to learn the second-order statistics for similarity learning. Following SoSN, Saliency-guided Hallucination Network [64], MsSoSN [65], and Few-shot Localizer [66] also use...
second-order statistics to improve the accuracy which demonstrates the usefulness of second-order pooling in FSL. The recent work [67] also gives a detailed theoretical analysis to demonstrate that the element-wise product of feature pairs can learn the correlations (higher order statistics). Therefore, the element-wise product of the outputs of two $1 \times 1$ convolutions can capture higher order statistics (correlations) for learning better feature representation. To learn a better instance-level feature representation and improve the reevaluation process of AIRN, we propose an efficient feature extractor, namely, the attentional bilinear feature extractor consists of a convolutional neural network followed by two parallel $1 \times 1$ convolutional layers, an attention pooling, and an L2 normalization.

Specifically, given an input instance $x_n$, its initial feature representation is the output of a convolutional neural network given by $f(x_n) \in \mathbb{R}^{d \times h \times w}$, where $d$, $h$, and $w$ are the channel, height, and width number of the instance feature, respectively. Then two parallel convolutional layers with kernel size 1 are applied to each initial representation to generate two feature representations $f_1 = f_{w_1}(f(x_n)) \in \mathbb{R}^{d \times h \times w}$ and $f_2 = f_{w_2}(f(x_n)) \in \mathbb{R}^{d \times h \times w}$, where $w_1$ and $w_2$ are the parameters of two $1 \times 1$ convolutional layers. Afterward, we compute the Hadamard product of two feature representations $f_1$ and $f_2$ to generate an intermediate feature representation $\hat{f}(x_n) = f_1 \odot f_2$, where $\odot$ denotes the elementwise multiplication. Attention pooling is applied on the intermediate representation $\hat{f}(x_n)$ for dimension reduction, where the spatially weight $A_n \in \mathbb{R}^{h \times w}$ for attention pooling is calculated as

$$A_n = \delta(w_{w}(f(x_n)))$$

where $w_{w}$ denotes the parameter of a $1 \times 1$ convolutional layer, and $\delta$ is the sigmoid function. Formally, the intermediate feature representation can be reshaped as $\hat{f}(x_n) \in \mathbb{R}^{d \times h \times w}$, and the spatially attention weights can be reshaped as $A_n \in \mathbb{R}^{h \times w \times 1}$. Therefore, the final instance-level representation $F_n \in \mathbb{R}^{d}$ of instance $x_n$ is given by

$$F_n = L_2(\hat{f}(x_n) \cdot A_n)$$

where $L_2$-normalization is adopted. Both the query and support instances share the same feature extractor.

D. Adaptive Instance Revaluing Network

After learning the instance-level representation, one crucial problem of the metric-based paradigm for few-shot visual recognition is learning effective class-level representations from few support instance-level representations. Previous works treat all support instances equally when generating class-level representation [27], [29]. However, they ignore the instance diversity, and hence the resulting representation may be biased. For example, some support instances may contain too much background or unrelated concepts information. Therefore, when learning the class-level representation, different instances should have different contributions. In light of the above analysis, we design an AIRN, which revalues the relative significance of all support instances in the same class. The class-level representation is obtained by adaptively and weightedly integrating instance-level representations according to their contributions. In this way, we can dynamically increase the importance value of more informative instances and decrease the weight of less informative instances.

The proposed architecture of AIRN is illustrated in Fig. 3. Specifically, given a support class $n$ with $K$-shot instances, when $K > 1$, the union set of instance-level representations is denoted as

$$U = [F_1, \ldots, F_K]$$

where each instance-level representation $F_n = [r_1, \ldots, r_d]$, $d$ is the feature dimension and $k \in \{1, \ldots, K\}$. Inspired by SENet [3], we first calculate the “summary statistics” of each instance-level representation and output a statistical vector $V = [v_1, \ldots, v_K]$, where each value $v_k$ is calculated as follows:

$$v_i = T_{\text{summary}}(F_n) = \frac{1}{K} \sum_{n=1}^{K} r_i$$

Then we use two fully connected layers to learn the weight of each instance based on the statistical vector $V$. The set of instance weights $A_n$ for class representation learning can be calculated as

$$A_n = T_{\text{weight}}(V) = \delta(w_3 \sigma(w_2 V))$$

where $A_n = [a_1, \ldots, a_K]$ is a vector, and each value represents the relative significance of a corresponding instance, $\delta$ denotes the sigmoid function, $\sigma$ denotes the rectified linear unit (ReLU) function, and $w_2$ and $w_3$ are the parameters of two fully connected layers.

Finally, the class-level representation of class $n$ is given by

$$c_n = T_{\text{combine}}(A_n, U) = \sum_{k=1}^{K} a_k \cdot F_n^k$$

where $a_k$ is the relative significance of the $k$th instance.

E. Classifier

Following [29], we use the cosine classifier to classify query instances in the final classification stage. Specifically, given a query instance $q_n^m$, a classifier is followed by a softmax activation to generate a probability distribution, which is defined as

$$p(y = n|c_n, q_n^m) = \frac{\exp(w_n^T \cdot F_n^m)}{\sum_{i=1}^{N} \exp(w_i^T \cdot F_n^m)}$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
and the predicted label of the query instance is

\[ \hat{y} = \arg \max_{n \in \{1, \ldots, N\}} (w_n^T, F_n^m) \]  

(8)

where \( F_n^m \) is the instance-level representation of \( q_n^m \), and \( w_n = (c_n/\|c_n\|_2) \) represents the weight parameter of class \( n \) in the cosine classifier.

**F. Optimization Loss**

After predicting the labels of the instances in the query set \( Q \), we have the following classification loss on the query set:

\[ L_{\text{cls}} = \sum_{(q_n, y_n) \in Q} \log(\hat{y} = y_n^m | q_n^m, \{c_n\}) \]  

(9)

where \( L_{\text{cls}} \) is the cross-entropy loss, and \( y_n^m \) and \( \hat{y} \) are the ground-truth label and predicted label for the query instance \( q_n^m \), respectively.

Meanwhile, to facilitate the learning of class-level representation, one solution is to decrease the similarities between the class-level representation of different classes and increase the similarities between the support instances and its own class-level representation. For example, Goldblum et al. [68] exploit the intraclass to interclass variance ratio to measure the feature clustering and minimize the feature clustering loss to reduce intraclass variation among features during training. This helps increase the interclass distance and decrease the intraclass distance in the representation space and enables the support instances to be more similar in the same class and dissimilar to other categories. Then the query instances can be easily classified. To achieve this goal, we design two novel structural losses, i.e., intraclass instance clustering loss and dissimilar to other categories. Then the query instances of class-level representation distinguishing loss denoted as \( L_{\text{inter}} \), to further refine the class-level representation, i.e.,

\[ L_{\text{inter}} = \sum_{i \neq j} \log(\hat{y} = y_{ij}^k | x_{ij}^k, \{c_n\}) \]  

(10)

\[ L_{\text{intra}} = \sum_{(c_i, c_j) \in s} \log(\hat{y} = y_{ij}^k | x_{ij}^k, \{c_n\}) \]  

(11)

where \( \hat{c}_i = (c_i/\|c_i\|_2) \) and \( \hat{c}_j = (c_j/\|c_j\|_2) \).

The overall objective function is a combination of \( L_{\text{cls}}, L_{\text{intra}}, \) and \( L_{\text{inter}} \), i.e.,

\[ L_{\text{joint}} = L_{\text{cls}} + \lambda_1 L_{\text{intra}} + \lambda_2 L_{\text{inter}} \]  

(12)

where \( \lambda_1 \) and \( \lambda_2 \) are the trade-off hyperparameters. The training procedure and the inference procedure of the ICRL-Net are summarized in Algorithms 1 and 2, respectively.

**IV. EXPERIMENTS**

To evaluate the effectiveness of ICRL-Net, we conducted extensive experiments on four publicly available and widely used few-shot visual recognition benchmarks, i.e., miniImageNet, tieredImageNet, CIFAR-FS, and FC100 datasets. In this section, we first introduce dataset details and experimental settings, and then comparisons with the state-of-the-art approaches. Finally, comprehensive ablation study is conducted to verify the effectiveness of different components.

**A. Dataset**

1) **MiniImageNet**: The miniImageNet dataset was initially proposed by Vinyals et al. [18], which is a standard benchmark for few-shot visual recognition. MiniImageNet is a subset randomly sampled from the ImageNet [72] dataset. MiniImageNet includes a total number of 100 classes and 600 images per class. We follow the split strategy proposed in [19] to split all the 100 classes into three subsets. One subset that contains 64 classes is used for training. The other two subsets used for validation and testing include 16 and 20 classes, respectively. All the images are resized to 84 × 84.

2) **tieredImageNet**: The tieredImageNet dataset was proposed by Ren et al. [30]. TieredImageNet is a relatively large subset sampled from the ImageNet [72] dataset and consists of 608 classes that can be grouped into 34 high-level categories. The tieredImageNet dataset is split into three subsets: a

---

**Algorithm 1** Training Process of ICRL-Net

**Input**: Training set \( D_{\text{train}} \)

1. while not done do
2. Sample a \( N \)-way \( K \)-shot task (\( S, Q \)) from \( D_{\text{train}} \)
3. for \( n = 1, \ldots, N \) do
4. for \( k = 1, \ldots, K \) do
5. Generate \( F_n^k \) for support instance \( x_n^k \) using (2)
6. end for
7. Generate \( c_n \) for class \( n \) using (3)–(6)
8. end for
9. for \( n = 1, \ldots, N \) do
10. for \( m = 1, \ldots, M \) do
11. Generate \( F_n^m \) for instance \( q_n^m \) using (2)
12. Predict \( \hat{y} \) for query instance \( q_n^m \) using Eq. (8)
13. end for
14. end for
15. Compute the loss function \( L_{\text{joint}} \) using (12)
16. Update the parameters of ICRL-Net with \( \nabla_{L_{\text{joint}}} \) using SGD
17. end while

**Algorithm 2** Inference Process of ICRL-Net

**Input**: Testing set \( D_{\text{test}} \)

**Require**: The trained ICRL-Net

1. Sample a \( N \)-way \( K \)-shot task (\( S, Q \)) from \( D_{\text{test}} \)
2. for \( n = 1, \ldots, N \) do
3. for \( k = 1, \ldots, K \) do
4. Generate \( F_n^k \) for support instance \( x_n^k \) using (2)
5. end for
6. Generate \( c_n \) for each class \( n \) using (3)–(6)
7. end for
8. for \( n = 1, \ldots, N \) do
9. for \( m = 1, \ldots, M \) do
10. Generate \( F_n^m \) for instance \( q_n^m \) using (2)
11. Predict \( \hat{y} \) for query instance \( q_n^m \) using (8)
12. end for
13. end for
14. Compute the predict accuracy for each episode (task)
training set, a validation set, and a testing set with 20, 6, and 8 high-level categories, respectively. The corresponding numbers of classes are 351, 97, and 160, respectively. All the images are of size $84 \times 84$.

3) CIFAR-FS: The CIFAR-FS dataset [31] is a recently proposed few-shot visual recognition benchmark, consisting of all the 100 classes from CIFAR-100 [73]. All the images on these datasets are $32 \times 32$, and the number of images per class is 600. Following previous works [20], [25], we divide the whole dataset into 64, 16, and 20 classes for training, validation, and testing, respectively.

4) FC100: The FC100 dataset [32] is another benchmark derived from CIFAR-100 [73]. There are 60 classes from 12 different superclasses for training, 20 classes from four different superclasses for validation, and 20 classes from four different superclasses for testing. Similar to the CIFAR-FS dataset, every class has 600 images of size $32 \times 32$.

B. Implementation Details

1) Pretraining Process: Following previous works [29], [47], we apply an additional pretraining phase to train the backbone network. Thus, the backbone network, appended with a linear layer, is trained to classify all the training classes (e.g., 64 classes in the miniImageNet) based on the cross-entropy loss. We follow the conventional deep learning pipeline and divide each training class into two parts: model training and validation. In this stage, the learning rate is set to 0.1. For all the datasets, random resized crop and horizontal flip data augmentations are used for model optimization. The classification performance over features of sampled one-shot tasks from the model validation split is evaluated to select the best pretrained model, whose weights are then used to initialize the backbone network in the meta-training phase.

2) Meta-Training Process: We follow the episodic training strategy [18] to train the ICRL-Net in the meta-training phase. Specifically, the pretrained backbone network is fine-tuned at a learning rate of 0.001. Meanwhile, other parts of ICRL-Net are initialized randomly and optimized with a learning rate of 0.01. We set the total number of the training epoch to be 200, and each epoch contains 100 episodes. All the experiments are implemented in PyTorch on an Ubuntu server with a single NVIDIA V100 GPU card. We adopt SGD with a Nesterov momentum of 0.9 and a weight decay of 0.0005 for model optimization in both pretraining and meta-training. Moreover, the learning rate decay is set to 0.1 in two in both pretraining and meta-training. We adopt the same data augmentation (i.e., random horizontal flip and random resized crop) when dealing with all the datasets during meta-training. More implementation details can be found in the Supplementary Material.

C. Performance Comparison

We compared the proposed method with the state-of-the-art methods, such as MAML [22], LEO [23], MTL [40], Matching Networks [18], Prototypical Networks [27], Relation Networks [28], Cross Attention Network (CAN) [46], Meta-Baseline [29], and Deep Subspace Network (DSN) [25].

1) Performance on miniImageNet and tieredImageNet: Table I shows the performance of all the comparison methods on miniImageNet and tieredImageNet. The best results are highlighted in boldface.

On the miniImageNet dataset, we observe that the proposed method ICRL-Net can achieve comparable performance compared with the state-of-the-art methods under both five-way one-shot and five-shot settings and can achieve the accuracies with 65.55% and 81.87% on one-shot and five-shot, respectively. Compared with several state-of-the-art approaches, such as Pareto self-supervised training (PSST) [36], ConstellationNet [58], and Meta-Baseline [29], we observe that the proposed ICRL-Net achieves the best performance. Note that our baseline is the same as Meta-Baseline [29], and they have a similar performance. Compared with Meta-Baseline, ICRL-Net brings significant improvements and improves one-shot accuracy by 2.38% and five-shot accuracy by 2.61%. Moreover, our approach has such an improvement attributed to the adaptive instance revaluing strategy that facilitates learning optimal class representations. The improvements demonstrate the necessity of considering the instance diversity for improving class representation learning.

On the tieredImageNet dataset, it is worth noting that the principle of dividing the training set and testing set is according to the disjoint sets, where the similarity of classes in each disjoint set is relatively high. Thus, it is more difficult to distinguish the categories in the training set and testing set. Nevertheless, on the tieredImageNet dataset, we observe that the proposed ICRL-Net still achieves comparable performance compared with other state-of-the-art algorithms. For example, on tieredImageNet, the five-way one-shot and five-shot accuracies of FEAT are 70.80% and 84.79%, respectively, and the accuracies of ICRL-Net are 70.56% and 85.62%, respectively. Note that our baseline is the same as Meta-Baseline [29], and they have a similar performance. Compared with Meta-Baseline [29], our ICRL-Net achieves a significant improvement of 1.94% and 2.33% accuracies for one-shot classification and five-shot classification, respectively. The improvements also prove the effectiveness of our approach.

2) Performance on CIFAR-FS and FC100: The experimental results on CIFAR-FS and FC100 are shown in Table II. The best results are highlighted in boldface. Our method also consistently outperforms the other state-of-the-arts methods under both one-shot and five-shot settings on the CIFAR-FS and FC100 datasets. For the CIFAR-FS dataset, our method outperforms the suboptimal method by 0.4% on one-shot and 0.5% on five-shot. Note that our ICRL-Net performs much better than the methods based on the average class representation: MetaOptNet [20] and Meta-Baseline [29]. For the FC100 dataset, our method also improves the classification performance by 0.7% and 0.1% under one-shot and five-shot settings, respectively. The results show the necessity and effectiveness of learning an optimal class representation.

D. Ablation Study

In this section, we conduct ablation studies to assess the effectiveness of each component of ICRL-Net, including AIRN, the attentional bilinear feature extractor, and the
designed loss function. We refer to the Supplementary Material for more studies.

1) Influence of the AIRN: In our baseline, class-level representation is generated by directly averaging a few support instance representations, and the performance is similar to Meta-Baseline [29]. Our ICRL-Net improves this baseline by designing the AIRN module, the attentional bilinear feature extractor, and novel loss function. To evaluate the effectiveness

### Table I

Comparisons with Other State-of-the-Art Methods on MiniImageNet and TieredImageNet. We use the officially provided results of all the other methods. The mean accuracy (%) of the proposed method ICRL-Net is obtained by over 600 testing episodes followed by the 95% confidence intervals (%). For each setting, the best result is highlighted. "-": Not reported.

| Model                        | Backbone   | MiniImageNet 5-way | TieredImageNet 5-way |
|------------------------------|------------|--------------------|----------------------|
|                              |            | 1-shot 5-shot      | 1-shot 5-shot        |
| Matching Networks [18]        | ConvNet-4  | 43.56 ± 0.84      | 55.31 ± 0.73        |
| Prototypical Networks [27]   | ConvNet-4  | 49.42 ± 0.78      | 68.20 ± 0.66        |
| MAML [22]                    | ConvNet-4  | 48.70 ± 1.84      | 63.10 ± 0.92        |
| Relation Networks [28]       | ConvNet-4  | 50.44 ± 0.82      | 65.32 ± 0.70        |
| Das et al. [59]              | ConvNet-4  | 52.68 ± 0.51      | 70.91 ± 0.85        |
| wDAE-GNN [33]                | WRN-28-10  | 62.96 ± 0.15      | 78.85 ± 0.10        |
| LEO [23]                     | WRN-28-10  | 61.76 ± 0.08      | 77.59 ± 0.12        |
| AWGIM [35]                   | WRN-28-10  | 63.12 ± 0.08      | 78.40 ± 0.11        |
| PSST [36]                    | WRN-28-10  | 64.16 ± 0.44      | 80.64 ± 0.32        |
| MTU-Net [37]                 | WRN-28-10  | 56.12 ± 0.43      | 71.93 ± 0.40        |
| BPM [9]                      | WRN-28-10  | 61.77 ± 0.72      | 73.73 ± 0.56        |
| AFIN [38]                    | ResNet-18  | 62.38 ± 0.72      | 78.16 ± 0.56        |
| VI-Net [16]                  | ResNet-18  | 61.05 ± 0.70      | 78.60 ± 0.30        |
| TADAM [32]                   | ResNet-12  | 58.50 ± 0.30      | 76.70 ± 0.30        |
| MTL [40]                     | ResNet-12  | 61.20 ± 1.80      | 75.50 ± 0.80        |
| MetaOptNet [20]              | ResNet-12  | 62.64 ± 0.61      | 76.63 ± 0.46        |
| CAN [46]                     | ResNet-12  | 63.85 ± 0.48      | 79.44 ± 0.34        |
| METANAS [24]                 | ResNet-12  | 61.70 ± 0.30      | 78.80 ± 0.20        |
| DSN [25]                     | ResNet-12  | 62.64 ± 0.66      | 78.83 ± 0.45        |
| RPS [26]                     | ResNet-12  | 62.02 ± 0.63      | 79.64 ± 0.44        |
| SLA-AG [14]                  | ResNet-12  | 62.93 ± 0.63      | 79.63 ± 0.47        |
| FIAT [47]                    | ResNet-12  | 66.78 ± 0.72      | 82.05 ± 0.70        |
| DeepEMD [70]                 | ResNet-12  | 65.9 ± 0.82       | 82.41 ± 0.56        |
| MAN [17]                     | ResNet-12  | 61.70 ± 0.47      | 78.42 ± 0.34        |
| ConstellationNet [58]        | ResNet-12  | 64.89 ± 0.23      | 79.95 ± 0.17        |
| Meta-UAPS [15]               | ResNet-12  | 64.22 ± 0.67      | 79.99 ± 0.49        |
| Meta-Baseline [29]           | ResNet-12  | 63.17 ± 0.23      | 79.26 ± 0.17        |
| Baseline                     | ResNet-12  | 62.71 ± 0.77      | 79.34 ± 0.56        |
| ICRL-Net (OURS)              | ResNet-12  | 65.55 ± 0.79      | 81.87 ± 0.51        |

### Table II

Comparison of Different Methods on CIFAR-FS and FC100. We use the officially provided results of all the other methods. The mean accuracy (%) over 600 testing episodes is reported followed by the 95% confidence intervals (%). For each setting, the best result is highlighted. "-": Not reported.

| Model                        | Backbone   | CIFAR-FS 5-way | FC100 5-way |
|------------------------------|------------|---------------|-------------|
|                              |            | 1-shot 5-shot | 1-shot 5-shot |
| MAML [18]                    | ConvNet-4  | 58.9 ± 1.9    | 71.5 ± 1.0  |
| Prototypical Networks [27]   | ConvNet-4  | 55.5 ± 0.7    | 72.0 ± 0.6  |
| Relation Networks [28]       | ConvNet-4  | 55.0 ± 1.0    | 69.3 ± 0.8  |
| TADAM [32]                   | ResNet-12  | 72.0 ± 0.7    | 84.3 ± 0.5  |
| MetaOptNet [20]              | ResNet-12  | 72.3 ± 0.8    | 85.1 ± 0.6  |
| DSN [24]                     | ResNet-12  | 73.5 ± 0.7    | 86.7 ± 0.5  |
| SLA-AG [14]                  | ResNet-12  | 75.4 ± 0.2    | 88.6 ± 0.2  |
| Baseline                     | ResNet-12  | 73.6 ± 0.6    | 85.4 ± 0.5  |
| ICRL-Net (OURS)              | ResNet-12  | 75.8 ± 0.7    | 87.3 ± 0.5  |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
of the proposed AIRN, we directly add this module on the baseline. Table III reports the comparison results under the one-shot and five-shot settings. From the results, we observe that using AIRN leads to significant improvements on five-shot tasks compared with the baseline. These improvements verify our intuition that the relative importance of few-shot support instances in the same class should be different when learning class-level representations. For one-shot tasks, since there is only one support instance per class. Therefore, class-level representation is just the instance-level representation, and AIRN does not take effect in one-shot tasks.

To further demonstrate the effectiveness of the proposed AIRN, we add some experiments that adopt other approaches to aggregate $K$-shot instances, such as similarity-type weights and attention-type weights. The experimental results are shown in Table IV. The AIRN module allows the model to pay more attention to the most informative support instance features while suppressing those unimportant support instance features and has the best performance when learning class-level representations.

2) Influence of the Attention Bilinear Feature Extractor:
To verify the effectiveness of the attentional bilinear feature extractor, we develop various variants of ICRL-Net to perform the process from the instance features to instance-level representations. In “Model-1,” global average pooling (instead of attention pooling) is adopted in the feature extractor of our ICRL-Net; in “Model-2,” attention pooling is removed from our extractor; in “Model-3,” two $1 \times 1$ convolutional layers are removed from our extractor; “Model-4” is to utilize only one $1 \times 1$ convolutional layer to transform the features in our extractor; finally, “Model-5” is to utilize naive bilinear pooling [59] in the feature extractor. Table V shows the comparison results under one-shot and five-shot settings. The results indicate that the attentional bilinear feature extractor yields better results than other variants. Therefore, it can conclude that the attention bilinear strategy is conducive to learning a discriminative and informative instance-level representation.

3) Influence of the Joint Loss Functions:
To demonstrate the effectiveness of the designed loss function, we compare the joint loss with three loss functions, including the original classification loss function $L_{\text{cls}}$, the $L_{\text{cls}} + L_{\text{inter}}$ loss function, and the $L_{\text{cls}} + L_{\text{intra}}$ loss function. The results are shown in Table VI. Compared with other loss functions, the designed loss function achieves the highest accuracy in the five-way one-shot and five-way five-shot tasks. Moreover, the accuracies of using $L_{\text{cls}} + L_{\text{inter}}$ and $L_{\text{cls}} + L_{\text{intra}}$ are higher than the original classification loss function $L_{\text{cls}}$, indicating that the proposed loss function is efficient for the representation learning of FSL.

4) Comparison on Different $K$-Shot Settings:
To further demonstrate the effectiveness of the adaptive instance revaluation strategy, we compare ICRL-Net with the baseline method under more $K$-shot ($K = 1, 3, 5, 7, 10, 20$) settings. The comparison results are shown in Fig. 4. From the results, we observe that ICRL-Net consistently outperforms the baseline under various $K$-shot settings on the miniImageNet and tieredImageNet datasets. The improvements again demonstrate the ability of ICRL-Net and validate the necessity and effectiveness of revaluing the importance of support instances when obtaining the class-level representation.

5) Hyperparameter Analysis:
We also conduct sensitivity analysis for the two trade-off hyperparameters in our joint loss. The experiments are performed by fixing the value of one hyperparameter and then changing the value of another hyperparameter. In specific, when analyzing the hyperparameter $\lambda_1$, the value of the hyperparameter $\lambda_2$ is set to 0.1.
TABLE V

| Method       | miniImageNet 5-way | tieredImageNet 5-way |
|--------------|-------------------|----------------------|
|              | 1-shot            | 5-shot               | 1-shot             | 5-shot             |
| Model-1      | 63.86 ± 0.80      | 81.23 ± 0.50         | 69.50 ± 0.82       | 84.94 ± 0.62       |
| Model-2      | 64.86 ± 0.81      | 81.67 ± 0.54         | 70.13 ± 0.83       | 85.33 ± 0.66       |
| Model-3      | 64.62 ± 0.84      | 81.56 ± 0.53         | 70.22 ± 0.84       | 85.10 ± 0.60       |
| Model-4      | 64.30 ± 0.77      | 81.25 ± 0.52         | 69.86 ± 0.80       | 85.14 ± 0.63       |
| Model-5      | 64.05 ± 0.76      | 80.78 ± 0.56         | 69.44 ± 0.86       | 84.76 ± 0.58       |
| ICRL-Net     | 65.55 ± 0.79      | 81.87 ± 0.51         | 70.56 ± 0.91       | 85.62 ± 0.64       |

TABLE VI

| Loss Function  | miniImageNet 5-way | tieredImageNet 5-way |
|----------------|-------------------|----------------------|
|                | 1-shot            | 5-shot               | 1-shot             | 5-shot             |
| $L_{cls}$      | 64.32 ± 0.80      | 81.12 ± 0.52         | 69.62 ± 0.84       | 85.07 ± 0.61       |
| $L_{cls} + L_{inter}$ | 65.26 ± 0.78      | 81.43 ± 0.56         | 70.32 ± 0.91       | 85.38 ± 0.61       |
| $L_{cls} + L_{intra}$ | 65.03 ± 0.80      | 81.50 ± 0.58         | 69.87 ± 0.84       | 85.24 ± 0.62       |
| $L_{joint}$    | 65.55 ± 0.79      | 81.87 ± 0.51         | 70.56 ± 0.91       | 85.62 ± 0.64       |

Similarly, when analyzing the hyperparameter $\lambda_2$, the value of the hyperparameter $\lambda_1$ is set to 0.1. The experimental results are shown in Fig. 5. We observe that the best performances are achieved when $\lambda_1 = \lambda_2 = 0.1$, and this setting is adopted in all the experiments.

E. Visualization of Relative Significance

We visualize the values of the relative significance of support instances to verify the effectiveness of the proposed method. We randomly selected several group examples from miniImageNet and tieredImageNet, and each group contains five support instances from the same class. The visualization results of the relative importance are shown in Fig. 6. Each column (group) of support instances is sorted by relative importance. It shows that more informative instances are generally assigned with larger importance weights according to ICRL-Net, except column (e). This might be due to multiple instances containing information related to other concepts, and our method does not accurately express the actual concepts of the instances. Therefore, in the future, we will focus on correctly expressing the target information related to the category when the instance contains multiple targets. We also visualize the learned attention maps in the Supplementary Material.

F. Cross-Domain FSL

Beyond the standard single-domain few-shot classification setting, we introduce a more challenging and realistic cross-domain setting. In this new setting, we aim to test the generalization of a trained model across previously unseen domains (datasets) with different data distributions. Cross-domain FSL can better evaluate the model’s generalization ability to novel
tasks. Following the instructions of [74], we conducted a cross-domain few-shot experiment by training the models on the miniImageNet training set and evaluating the model on the CUB200 testing set. This set of experiments is designed to evaluate the performance of different algorithms when the distribution divergence between the training and testing sets is large. To fairly compare with other approaches, we adopt the CUB testing split presented in their original work. The comparison results are reported in Table VII, where we can see that the proposed ICRL-Net can outperform other approaches. The cross-domain few-shot settings further demonstrate the effectiveness of the proposed method.

![Fig. 6. Visualization of the learned importance. (a)–(e) Support instances sampled from miniImageNet, and each class contains five support instances. (f) and (g) Support instances sampled from tieredImageNet. Each column of support instances is sorted by relative importance.](image)

| Table VII: Five-Shot Classification Accuracy Under Dataset Shift |
|---------------------------------------------------------------|
| **Model (%)** | miniImageNet → CUB |
| Linear Classifier [74] | 65.57 ± 0.70 |
| Cosine Classifier [74] | 62.04 ± 0.76 |
| MetaOptNet-SVM [20] | 54.67 ± 0.56 |
| **Baseline** | 64.86 ± 0.73 |
| **ICRL-Net** | 67.08 ± 0.55 |

**G. Transductive FSL**

We further extend the proposed method for transductive FSL by integrating our AIRN with the transductive inference algorithm developed by CAN [46]. Specifically, we first use the proposed AIRN to obtain the class features and use the initial class features to predict the labels of the unlabeled query samples. Then we select several pseudolabeled unlabeled query samples according to the label confidence criterion proposed by CAN [46]. Finally, the selected pseudolabeled unlabeled query samples are used together with the initial class features to generate more representative class features. From Table VIII, we can see that the performance of the proposed ICRL-Net under the transductive setting (ICRL-Net+T) is further improved.
In this article, we present a novel framework for metric-based FSL by considering the information diversity of instances. The designed ICRL-Net can automatically learn the relative significance of different support instances that belong to the same class. We conducted extensive experiments on four few-shot visual recognition benchmarks, and from the results, we mainly conclude that: 1) exploiting the significance of different instances is critical in obtaining the class-level representation, and the proposed ICRL-Net can adaptively and effectively evaluate the relative significance of different instances and 2) the proposed method outperforms the state-of-the-art FSL counterparts by a large margin, and all the different modules in our ICRL-Net are essential to achieving satisfactory performance. One disadvantage of our method may be that the importance is assessed at the instance level, while the relative importance of different regions in a visual instance is not considered. How to revaluing the significance of different regions inside one instance remains a potential future work.

ACKNOWLEDGMENT

The authors would like to thank the handling associate editor and all the anonymous reviewers for their constructive comments.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.
[3] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7152–7161.
[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2015, pp. 91–99.
[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.
[6] K. He, X. Glorot, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2961–2969.
[7] L. Fei-Fei, “A Bayesian approach to unsupervised one-shot learning of object categories,” in Proc. 9th IEEE Int. Conf. Comput. Vis., Oct. 2003, pp. 1134–1141.
[8] 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.
[9] N. Lai, M. Kan, C. Han, X. Song, and S. Shan, “Learning to learn adaptive classifier-predictor for few-shot learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 8, pp. 3458–3470, Aug. 2020.
[10] J. Lu, S. Jin, J. Liang, and C. Zhang, “Robust few-shot learning for user-provided data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 4, pp. 1433–1447, Apr. 2021.
[11] H.-G. Jung and S.-W. Lee, “Few-shot learning with geometric constraints,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 11, pp. 4660–4672, Nov. 2020.
[12] S. Gidaris and N. Komodakis, “Dynamic few-shot visual learning without forgetting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4367–4375.
[13] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, “Laplacian regularized few-shot learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 11660–11667.
[14] H. Lee, S. J. Hwang, and J. Shin, “Self-supervised label augmentation via input transformations,” in Proc. 37th Int. Conf. Mach. Learn. (ICML), 2020, pp. 5714–5724.
[41] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, “A simple neural attentive meta-learner,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018, pp. 1–17.

[42] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in Proc. ICML, 2015.

[43] D. Q. M. Brown and D. G. Lowe, “Low-shot learning with imprinted weights,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5822–5830.

[44] B. Hariharan and R. Girshick, “Low-shot visual recognition by shrinking and hallucinating features,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3018–3027.

[45] H. Qi, M. Brown, and D. G. Lowe, “Low-shot learning with imprinted weights,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5822–5830.

[46] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, “Few-shot network via embedding adaptation with set-to-set functions,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8808–8817.

[47] S. Min Kye, H. Beom Lee, H. Kim, and S. Ju Hwang, “Meta-learned confidence for few-shot learning,” 2020, arXiv:2002.12017.

[48] Y. Song, C. Li, F. Yu, and C. Chen, “ReMP: Rectified metric propagation for few-shot learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 2581–2590.

[49] 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. (CVPR), Jun. 2020, pp. 12836–12845.

[50] Y. Wang, L. Zhang, Y. Yao, and Y. Fu, “How to trust unlabeled data instance credibility inference for few-shot learning,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jun. 3, 2021, doi: 10.1109/TPAMI.2021.3086140.

[51] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2, 2018.

[52] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.

[53] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” 2017, arXiv:1710.10903.

[54] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[55] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, “Transformer in transformer,” 2021, arXiv:2103.00112.

[56] P. Wang, L. Liu, C. Shen, Z. Huang, A. Van Den Hengel, and H. T. Shen, “Multi-attention for one-shot learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2721–2729.

[57] W. Xu et al., “Attentional constellation nets for few-shot learning,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021, pp. 1–16.

[58] T.-Y. Lin, A. RoyChowdhury, and S. Maji, “Bilinear CNN models for fine-grained visual recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1449–1457.

[59] C. Jonescu, O. Vantzos, and C. Sminchisescu, “Matrix backpropagation for deep networks with structured layers,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2965–2973.

[60] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, “Hadamard product for low-rank bilinear pooling,” 2016, arXiv:1610.04325.

[61] M. Lin, Q. Chen, and S. Yan, “Network in network,” 2013, arXiv:1312.4400.

[62] H. Zhang and P. Koniusz, “Power normalizing second-order similarity network for few-shot learning,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WCAC), Jan. 2019, pp. 1185–1193.

[63] H. Zhang, J. Zhang, and P. Koniusz, “Few-shot learning via saliency-guided hallucination of samples,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2770–2779.

[64] H. Zhang, P. H. S. Torr, and P. Koniusz, “Few-shot learning with multi-scale self-supervision,” 2020, arXiv:2001.01600.

[65] D. Wertheimer and B. Hariharan, “Few-shot learning with localization in realistic settings,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6558–6568.

[66] P. Koniusz and H. Zhang, “Power normalizations in fine-grained image, few-shot image and graph classification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 2, pp. 591–609, Feb. 2022.
Bo Du received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010. He is currently a Professor with the School of Computer, Wuhan University. He has authored more than 100 scientific articles, such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, the Association for the Advancement of Artificial Intelligence (AAAI), and the International Joint Conference on Artificial Intelligence (IJCAI). His research interests include pattern recognition, hyperspectral image processing, and signal processing.

Han Hu received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2007 and 2012, respectively. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include multimedia networking, edge intelligence, and space–air–ground integrated network. Dr. Hu has served as a TPC Member of Infocom, the ACM International Conference on Multimedia (ACM MM), AAAI, and the International Joint Conference on Artificial Intelligence (IJCAI). He received several academic awards, including the Best Paper Award of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) 2019, the Best Paper Award of the IEEE Multimedia Magazine 2015, and the Best Paper Award of the IEEE Globecom 2013. He has served as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA (TMM) and Ad Hoc Networks.

Yonggang Wen (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science (minor in western literature) from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 2008. He served as the Acting Director of the Nanyang Technopreneurship Centre (NTC), Nanyang Technological University (NTU), Singapore, from 2017 to 2019, and the Assistant Chair (Innovation) of the School of Computer Science and Engineering (SCSE), NTU, from 2016 to 2018, where he has been serving as the Associate Dean (Research) of the College of Engineering since 2018 and is currently a Professor of computer science and engineering. His research interests include cloud computing, green data center, distributed machine learning, blockchain, big data analytics, multimedia network, and mobile computing. Dr. Wen was a co-recipient of multiple journal best papers awards, including IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2019 and IEEE MULTIMEDIA in 2015, and several best paper awards from international conferences, including 2020 IEEE Visual Communications and Image Processing (VCIP), 2016 IEEE Globecom, 2016 IEEE Infocom MuSiC Workshop, 2015 EAI/ICST Chinacom, 2014 IEEE International Conference on Wireless Communications and Signal Processing (WCSP), 2013 IEEE Globecom, and 2012 IEEE International Conference on Embedded and Ubiquitous Computing (EUC). He also received the 2016 IEEE ComSoc MMTTC Distinguished Leadership Award.

Dacheng Tao (Fellow, IEEE) is currently an Advisor and a Chief Scientist of the Faculty of Engineering, Digital Science Institute, The University of Sydney, Darlington, NSW, Australia. He is also the Director of the JD Explore Academy, Beijing, China, and a Vice President of JD.com, Beijing. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences. Dr. Tao is a fellow of the Australian Academy of Science, the American Association for the Advancement of Science (AAAS), and ACM. He received the 2015 Australian ScopusEureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award.