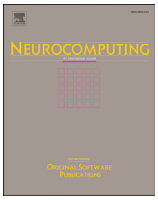
Neurocomputing 655 (2025) 131377

Contents lists available at ScienceDirect

Neurocomputing

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



Interactive triplet attention for few-shot fine-grained image classification

Xiaoxu Li [a,b], Shaoying Xue [a], Jiyang Xie [b], Xiaochen Yang [c,*], Zhanyu Ma [b], Jing-Hao Xue [d]

[a] School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China
[b] Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[c] School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8QQ, UK
[d] Department of Statistical Science, University College London, London, WC1E 6BT, UK

ARTICLE INFO

Communicated by G. Ciocca

Keywords:
Few-shot learning
Fine-grained image classification
Attention mechanisms

ABSTRACT

Few-shot fine-grained classification aims to identify novel fine-grained classes from extremely few examples with ultra-high semantic similarity between classes, hence a notoriously hard task. To extract discriminative features from *few samples* for recognizing subtle differences between *fine-grained classes*, it is pivotal to exploit comprehensive interactions across all dimensions in space and channel, which, however, is unexplored yet by state-of-the-art methods in this challenging area. To address this issue, in this paper we show that a simple adjustment to the existing triplet attention module (TAM) can be highly effective for few-shot fine-grained image classification. More specifically, building on TAM which comprises three parallel branches for pairwise interactions between height, width, and channel dimensions, we introduce an additional interaction between the outputs of these three branches, capable of modeling the dependency across all three dimensions; the revised method is dubbed interactive triplet attention module (ITAM). ITAM is a plug-and-play module, which can be inserted into any metric-based few-shot fine-grained image classifiers for performance enhancement. Extensive experiments, on CUB-200-2011, Flowers, Stanford-Cars, and Stanford-Dogs, showcase the superiority of ITAM against state-of-the-art few-shot fine-grained image classifiers.

1. Introduction

Fine-grained image classification is a challenging task in computer vision. Unlike other tasks, fine-grained classification aims to identify sub-categories of objects with subtle differences, such as different breeds of birds, dogs, or flowers, hence challenging. For example, as shown in Fig. 1, the focus should be on subtle differences in beak, wing, or belly feathers as clues for distinguishing between bird breeds. In general, a quality fine-grained classifier requires a large number of labeled samples for training. However, often only very few labeled samples are available in many fields, e.g., new species in biological research [1] and rare cases in medical research [2]. To tackle the problem of few training samples, in recent years many few-shot learning methods have been proposed [3–5]. However, in the case of few-shot fine-grained image classification [6], due to the dual issues of few shots and high semantic similarity of fine-grained classes, it is still an open question of how to extract discriminative features from *few samples* for recognizing subtle differences between *fine-grained classes*. To partly answer this question, we believe it is pivotal to exploit comprehensive interactions across all dimensions in space and channel, due to the following two reasons. Firstly, fine-grained objects often differ in only tiny regions and certain properties, such as distinct color patterns in belly feathers, and therefore both spatial and channel attention are crucial. Secondly, when objects have multiple discriminative regions, the importance of these properties may vary across different regions, necessitating the interaction between spatial and channel dimensions. However, as far as we know, such interaction is unexplored yet by state-of-the-art methods in this challenging area.

To model the interactions among spatial and channel dimensions, this paper proposes adapting the triplet attention module (TAM) [7] for fine-grained few-shot image classification. The original TAM is used in standard image classification and consists of three branches to establish the pairwise interactions between the dimensions of height, width and channel. By aligning and fusing spatial attention and channel attention, TAM can effectively extract discriminative features to highlight desired targets while reducing unwarranted interference from the background, hence we believe it can be exploited to better recognize subtle differences between fine-grained classes and improve their few-shot

* Corresponding author.
Email address: xiaochen.yang@glasgow.ac.uk (X. Yang).

https://doi.org/10.1016/j.neucom.2025.131377
Received 12 March 2025; Received in revised form 14 August 2025; Accepted 23 August 2025
Available online 26 August 2025

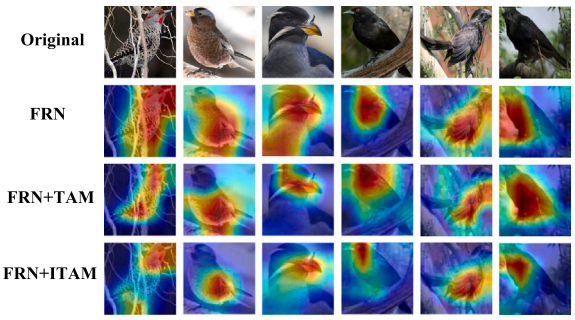


Fig. 1. Heatmaps of features for samples from the CUB dataset. Compared with FRN and FRN + TAM, the proposed FRN + ITAM focuses more on discriminative regions of birds.

classification. However, TAM is only capable of modeling first-order interactions between two dimensions. Therefore, to further enhance TAM, we introduce an additional interaction between every pair of the three branches to capture the dependency across all three dimensions, which enables explicit modeling of up to third-order interactions, thereby capturing more complex relationships between features. The resulting method, termed interactive triplet attention module (ITAM), can further improve the feature discriminativeness, as demonstrated in Fig. 1. It is also worth mentioning that ITAM does not introduce additional parameters, which is highly desirable when dealing with scarce support data, and we shall show that ITAM is a plug-and-play module, which can be inserted into any metric-based few-shot fine-grained image classifiers for performance enhancement.

In sum, the novelties and contributions of this paper are:

- We investigate TAM for a new task of few-shot fine-grained image classification and show that cross-dimensional attention enables the classifier to focus on the discriminative features, aiding in the recognition of subtle differences between fine-grained classes from only few labeled samples.
- We improve TAM, which models two-dimensional interactions, to ITAM, which models three-dimensional interactions, leading to more discriminative features.
- We show that ITAM can be readily inserted into various metric-based few-shot image classifiers with different network structures.
- Extensive experiments, on the benchmark datasets of CUB-200-2011, Flowers, Stanford-Cars and Stanford-Dogs, showcase the superiority of ITAM against state-of-the-art few-shot fine-grained image classifiers.

2. Related work

2.1. Metric-based few-shot learning

Metric-based approach, aiming to accurately measure the similarity between support samples and query samples, is de-facto popular and state-of-the-art approach to few-shot learning [4], with various metrics adopted or developed to classify samples, e.g., MatchingNet [8] uses cosine similarity, ProtoNet [9] adopts Euclidean distance, and DeepBDC [10] employs the Brownian distance covariance as the metric, which can capture non-linear relationships between features and is computationally efficient. Instead of using fixed metrics, RelationNet [11] designs a relation module to learn the distance between a query sample and a support class. As an improvement, SLTRN [12] uses a transformer architecture to learn the relationship between a query and all support samples, fully exploiting the information from the support set. NTK-FSCIL [13] introduces neural tangent kernel theory, combining meta-learning and regularization to improve generalization.

2.2. Few-shot fine-grained image classification

The discriminativeness and generalizability of feature representations are crucial to few-shot fine-grained image classification. In particular, low-level feature representations are often adopted as they contain local detailed information. DN4 [14] is a pioneering work in this line, which shifts the classification regime from image-to-image comparison to image-to-class comparison built on local features. It computes the distances between a local feature of a query image and the k most similar local features of support images in a class, and then aggregates the distances over all local features of the query. FRN [15] hypothesizes that a query image can be approximated by support images of its own class and thus uses ridge regression to reconstruct local features of a query image from the pool of support features and predicts the class of the query sample based on its distance to the reconstructed query. HelixFormer [16] refines support local features by considering their relations to the query local features, modeled through cross-attention, and similarly refines query local features. AGPF [17] extracts multi-scale features and reweights the features via the multi-level attention pyramid. LCCRN [18] designs a new cross-reconfiguration module, which can fully integrate the base feature representation and the local content-rich feature representation to enhance the semantic understanding of the network. KLSANet [19] crops random regions of an image to extract diverse local features and filters out irrelevant query parts before computing image-to-class similarity. FSC [20] addresses cross-domain few-shot learning by improving domain invariance through frequency-spatial fusion and style-based attacks. CoDF [21] employs self-supervised learning techniques to endow representations with richer information, thereby facilitating the acquisition of key information required for few-shot class-incremental learning tasks.

Orthogonal to the aforementioned methods, few-shot fine-grained image classification can be improved by advancing the distance or similarity measures. BSNet [22] combines cosine similarity and a relation module to form a double similarity module. DeepEMD [23] uses the earth mover's distance as a metric to calculate the structural distance between local features.

In this paper, a new approach is proposed to improve feature discriminativeness. Compared with other methods, ITAM uses only operations such as rotation and matrix multiplication, which do not introduce any trainable parameters and are particularly well-suited for the few-shot learning task.

2.3. Attention mechanisms

Squeeze and excitation network (SENet) [24] computes channel attention and provides incremental performance gains at a fairly low cost. Convolutional block attention module (CBAM) [25] provides robust representative attention by combining spatial attention with channel attention. Global-context network (GCNet) [26] proposes a novel non-local block that integrates with SE blocks to combine context representation with channel weighting. Efficient channel attention (ECA) [27] proposes a local cross-channel interaction strategy. Coordinate attention (CA) [28] captures cross-channel information, as well as directional perception and positional perception information. Normalize attention module (NAM) [29] reduces the weight of less significant features by applying sparse weight penalties to the attention module. The triplet attention module (TAM) [7] uses rotation operations and residuals to establish inter-dimensional dependencies with negligible computational overhead. Efficient multi-scale attention (EMA) [30] designs a multi-scale parallel sub-network to establish short and long dependencies. Some channel dimensions are reshaped into batch dimensions to avoid certain forms of dimensionality reduction through general convolution. In addition to constructing local cross-channel interaction in each sub-network, the output feature graphs of two parallel sub-networks are fused by cross-space learning. Efficient local attention (ELA) [31] improves the efficiency of computing spatial attention and avoids compressing channel attention. It encodes spatial positional information by

Table 1
Comparison of attention mechanisms in terms of their type, purpose, and structure. 'GAP' and 'GMP' refer to global average pooling and global maximum pooling, respectively; 'MLP' refers to multi-layer perceptron; 'Conv' refers to convolution.

Method	Attention type	Purpose	Structure
SENet	Channel attention	Explicitly model interdependencies between channels	2D GAP followed by a self-gating mechanism based on MLP
ECA	Channel attention	Learn channel attention more efficiently	2D GAP followed by 1D Conv along the channel dimension
CBAM	Spatial + channel attention	Extract meaningful information along both spatial and channel dimensions	1D channel attention (GAP, GMP + MLP), followed by 2D spatial attention (GAP, GMP + 2D Conv)
NAM	Spatial + channel attention	Utilize the variance of the trained model weights to adjust channel and spatial attention	Multiply features by a scaling factor based on batch normalization
CA	Spatial + channel attention	Capture both channel relationships and long-range spatial dependencies	Two 1D GAP to aggregate features along H and W dimensions, followed by separate 2D Conv and attention multiplication
EMA	Spatial + channel attention	Capture both channel relationships and short-range and long-range spatial dependencies	Divide the channel dimension into groups, followed by CA and 2D spatial dimension in parallel
ELA	Spatial + channel attention	Improve CA to learn spatial attention more efficiently without compressing channel dimension	Two 1D GAP to aggregate features along H and W dimensions, followed by separate 1D Conv, GroupNorm, and attention multiplication
DCAFE	Spatial + channel attention	Improve CA to preserve the most significant features	Parallel CA where one branch is CA and the other applies GMP for feature aggregation
GCNet	Spatial + channel attention	Integrate global spatial attention with channel attention	Global attention pooling via 2D Conv, feature transform via 2D Conv, and feature aggregation via addition
AA	New attention paradigm	Reduce redundancy of self-attention	Introduce a small set of agent token to replace the value tokens in attention
TAM	Spatial + channel attention	Model first-order interaction between spatial and channel attention	Rotation followed by residual transformations
ITAM	Spatial + channel attention	Model high-order interaction between spatial and channel attention to extract discriminative features	Rotation followed by residual transformations and pairwise multiplication between refined features

combining 1D convolution and group normalization without dimensionality reduction. Agent attention (AA) [32] introduces a small number of agent tokens to collect information from key and value tokens and then deliver it to query, which significantly reduces computational cost while preserving global context modeling capability. Dual coordinate attention feature extraction (DCAFE) [33] uses coordinate attention [28] that better captures complex features such as petal patterns and structural variations.

The proposed ITAM module is built on TAM [7]. However, unlike the above attention mechanisms, ITAM captures the full interaction between dimensions of the input tensors through three parallel branches, offering full cross-dimensional attention. Our goal is to stress the importance of full cross-dimensional interactions to efficiently and effectively capture from few samples the discriminative feature representations for recognizing subtle differences between classes, hence particularly fit for few-shot fine-grained image classification. A comparison between ITAM and other attention mechanisms is listed in Table 1.

3. Interactive triplet attention network

3.1. Problem formulation

Few-shot image classification aims to use only very few labeled samples from each category to learn knowledge on data representations and/or classification models, and apply the learned knowledge to classify new categories. To achieve this, the datasets are usually divided into training set D_{train}, validation set D_{val}, and test set D_{test}, with the class labels being disjoint between these three datasets. To avoid overfitting on the training set and achieve good generalization on the test set, we adopt the widely used episodic training mechanism [8].

Specifically, in each set, a series of episodes (a.k.a. tasks) is generated as follows. Firstly, randomly sample N classes from $D_{train}/D_{val}/D_{test}$. Then, from each class, randomly sample K images to form the support set S and M images to form the query set Q. This setting is known as an "N-way K-shot" classification problem. Samples in the query set Q are unlabeled and will be classified based on the knowledge learned from the labeled support samples.

3.2. Network overview

As shown in Fig. 2, the interactive triplet attention network consists of three main components.

The first component is the feature embedding network. Its purpose is to extract features from images.

The second component is the interactive triplet attention module (ITAM). This module first establishes three pairwise interactions between the dimensions of height, width, and channel by using three parallel branches. Secondly, to comprehensively fuse spatial and channel attention, the module performs the interaction between every pair of the three branches. In this way, the interaction among all three dimensions is considered, thereby generating discriminative features that can facilitate subsequent classification.

The third component is the metric module, which calculates the similarity between support features and query features. Any similarity measure or similarity learning module can be used. In this paper, we adopt the state-of-the-art feature reconstruction network (FRN) [15] to reconstruct query features from support features and use the Euclidean distance between the original query features and the reconstructed query features for classification.

3.3. Interactive triplet attention module

Fig. 3 shows the ITAM module. Given an image x, the feature embedding network is used to generate the corresponding embedded features $\hat{x} = f(x \mid \theta) \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of channels, and H and W denote the height and width of the feature map, respectively. Following TAM [7], to establish the pairwise interactions between the dimensions of height H, width W, and channel C, the feature tensor $\hat{x}$ is passed to three parallel branches, where the first branch is responsible for capturing the interaction between the height dimension H and the channel dimension C; the second branch is for the interaction between the channel dimension C and the width dimension W; and the third branch is for the spatial attention between H and W. In the first and second branches, the feature tensor $\hat{x}$ is rotated 90° counterclockwise along the H and W axes, respectively, and these two rotated features are represented as $\check{x}_1$ and $\check{x}_2$:

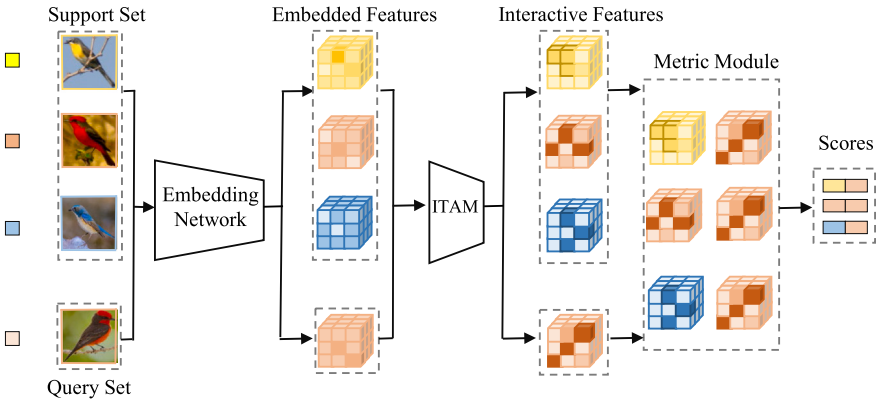


Fig. 2. Architecture of the interactive triplet attention network under the 3-way 1-shot setting, which includes an embedding network to extract image features, the interactive triplet attention module (ITAM) to generate more discriminative features, and a metric module to calculate the similarity between support features and query features. Darker blocks indicate that the corresponding features are more discriminative.

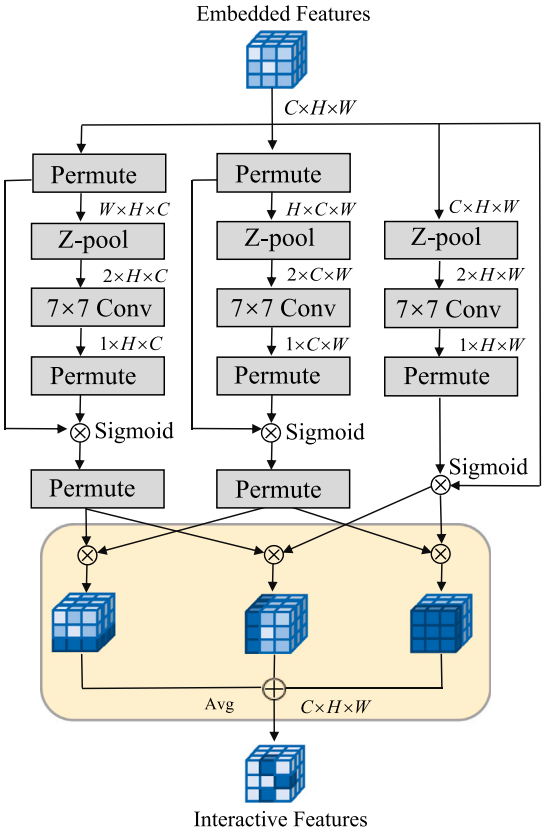


Fig. 3. Architecture of the interactive triplet attention module (ITAM). ITAM consists of three branches, where the left, middle, and right branches compute the interactions between the (H, C), (C, W), and (H, W) dimensions of the feature tensor, respectively. The novelty of ITAM compared with TAM lies in the yellow region. Since each branch only considers two-dimensional interactions, ITAM further performs interactions between every pair of the refined features, enabling information fusion across all three dimensions.

$$\check{x}_1 = pm_1\left(\hat{x}\right) \in \mathbb{R}^{W\times H\times C}, \tag{1}$$

$$\check{x}_2 = pm_2\left(\hat{x}\right) \in \mathbb{R}^{H\times C\times W}, \tag{2}$$

where pm_1 represents a counterclockwise rotation of 90° along the H axis and pm_2 represents a counterclockwise rotation of 90° along the W axis. The rotation operation refers to axis reordering and is implemented using PyTorch's `.permute()` function, which is functionally equivalent to a transpose extended to multiple dimensions. In the third branch, no rotation operation is applied in order to preserve the original feature tensor. For notation consistency, $\check{x}_3 = \hat{x}$ is introduced. Then a Z-pool layer is applied to aggregate the cross-dimensional features. Given a rotated feature tensor $\check{x}$, the Z-pool layer is calculated as:

$$Z\text{-pool}\left(\check{x}_n\right) = \left[GMP\left(\check{x}_n\right), GAP\left(\check{x}_n\right)\right], \; n \in [1,2,3], \tag{3}$$

where GMP and GAP denote the global max pooling operation and the global average pooling operation, respectively. Both operations are applied to the first dimension of the tensor. For example, for $\check{x}_1 \in \mathbb{R}^{W\times H\times C}$, the outputs of GMP and GAP are both of dimensional $2\times H\times C$. By utilizing the two pooling operations, the Z-pool can effectively reduce the first dimension to two while preserving the most prominent features and the overall pattern, thereby retaining the rich feature representation of the original image. Next, the attention weight s_n is generated through a standard convolution layer $Conv(\cdot)$ with a kernel size of 7×7 and an activation layer σ:

$$s_n = \sigma\left(Conv_{7\times 7}\left(Z\text{-pool}\left(\check{x}_n\right)\right)\right), \tag{4}$$

The resulting attention weights s_n are applied to $\check{x}_n$.

The final step is to generate the enhanced interactive features. To ensure all features have the same shape, the feature tensor in the first two branches is rotated 90° clockwise along the H and W axes, respectively, returning it to the original shape. Different from TAM [7] which averages over the three feature tensors, we perform a multiplication between every pair of the refined feature tensors and then aggregate the interactive features by using a simple average:

$$\tilde{x} = \frac{1}{3}\Big(pm_1^{-1}\left(s_1\check{x}_1\right)\times pm_2^{-1}\left(s_2\check{x}_2\right) + pm_2^{-1}\left(s_2\check{x}_2\right)\times s_3\check{x}_3 + pm_1^{-1}\left(s_1\check{x}_1\right)\times s_3\check{x}_3\Big), \tag{5}$$

where pm_1^{-1} represents a clockwise rotation of 90° along the H axis and pm_2^{-1} represents a clockwise rotation of 90° along the W axis. The obtained feature $\tilde{x}$ has the same shape as the input feature, i.e., with dimensions $C\times H\times W$.

In this process, ITAM aligns attention in both spatial dimensions and channel dimensions with only rotation and residual convolution, which does not introduce any additional parameters. Moreover, with each branch already capturing first-order interactions, the multiplication between pairs of the three branches brings in a third-order interaction among all three dimensions. This enables the modeling of more complex relationships between features, potentially improving their discriminability. As demonstrated in Section 4.5, feature embeddings resulting from ITAM are more discriminative than those from TAM.

3.4. Metric module

The enhanced features generated by ITAM can be used in any metric-based few-shot learning methods. In this section, we exemplify this by combining the ITAM with the state-of-the-art feature reconstruction network (FRN) [15].

The feature reconstruction module first converts the feature tensor into $r\,(r = H \times W)$ C-dimensional local descriptors $\tilde{x} = \left[\tilde{x}_{(1)}, \ldots, \tilde{x}_{(r)}\right]$, and then gathers the local descriptors of the same class in a support feature pool to represent a class. In other words, in the N-way K-shot classification task, for each class n, all the features in K support images are gathered into a support feature pool $S_n \in \mathbb{R}^{Kr \times C}$. Similarly, a query feature tensor can be transformed into $Q \in R^{r \times C}$. Next, ridge regression is applied to reconstruct the query feature by using the support feature pool of class n, leading to the reconstructed query feature $\bar{Q}_n$:

$$\bar{Q}_n = \rho Q S_n^{\top}\left(S_n S_n^{\top} + \lambda I\right)^{-1} S_n, \tag{6}$$

$$\lambda = \frac{KHW}{d} e^{\alpha}, \tag{7}$$

$$\rho = e^{\beta}, \tag{8}$$

where I is the identity matrix. To ensure that λ and ρ are always positive, they are defined as $\lambda = \frac{KHW}{d} e^{\alpha}$ and $\rho = e^{\beta}$, where α and β are learnable hyperparameters.

Finally, we calculate the squared Euclidean distance between the original query feature and the reconstructed query feature:

$$d\left(Q, \bar{Q}_n\right) = \frac{1}{r} \parallel Q - \bar{Q}_n \parallel^2, \tag{9}$$

where the squared Euclidean distance $d\left(Q, \bar{Q}_n\right)$ can be viewed as the mean squared error of reconstructing the query features based on the n^{th} class support features. For a query sample, it is more likely that its features can be reconstructed by using the features from the support samples of the same class than those from different classes. Therefore, the query sample is assigned to the class with the smallest Euclidean distance.

3.5. Loss function

The final classification is obtained by applying the softmax function to the computed Euclidean distances:

$$P\left(y_q = n \mid x_q\right) = \frac{e^{\left(-\tau d\left(Q, \bar{Q}_n\right)\right)}}{\sum_{n' \in \mathcal{N}} e^{\left(-\tau d\left(Q, \bar{Q}_{n'}\right)\right)}}, \tag{10}$$

where x_q denotes the query sample, y_q denotes the predicted label of the query sample, $\mathcal{N}$ denotes the set of class labels in an episode, and τ is a learnable hyperparameter.

The entire network is trained by minimizing the cross-entropy (CE) loss:

$$L_{CE} = -\frac{1}{M} \sum_{i=1}^{M} \left(\boldsymbol{y}_i^{\top} \log\left(\boldsymbol{p}_i\right)\right), \tag{11}$$

where $\boldsymbol{y}_i$ represents a one-hot vector, $\boldsymbol{p}_i$ represents the classification probability vector, and M represents the number of query samples.

4. Experimental results and analysis

4.1. Datasets

In order to test the effectiveness of the proposed method, we evaluated it on four fine-grained datasets: CUB-200-2011 (CUB), Stanford-Dogs (Dogs), Oxford-102-Flower (Flowers), and Stanford-Cars (Cars). A brief description of the datasets is as follows:

CUB-200-2011 [34]: The dataset contains 11,788 bird images of 200 classes, randomly divided into a training set of 100 classes, a validation set of 50 classes, and a test set of 50 classes. In addition,each image is cropped into a bounding box according to the preprocessing methods [23,35].

Stanford-Cars [36]: The dataset contains 16,185 images of 196 types of cars, randomly divided into a training set of 98 classes, a validation set of 49 classes, and a test set of 49 classes.

Stanford-Dogs [37]: The dataset contains 20,580 dog images of 120 categories, randomly divided into a training set of 60 classes, a validation set of 30 classes, and a test set of 30 classes.

Flowers [38]: The dataset contains 102 categories of flowers, randomly divided into a training set of 51 classes, a validation set of 26 classes, and a test set of 25 classes.

4.2. Implementation details

Two widely used feature extractor networks were adopted in the experiments: ResNet-12 and ResNet-18. For ResNet-12, we used the same implementation as in [23,35,39,40]. The input image size is 84×84, and the output feature map shape is 640×5×5. For ResNet-18, which is an improvement on ResNet-12, there are four residual blocks like ResNet-12, but the first two residual blocks are divided into two sub-residual blocks, each of which contains 3 convolutional layers with 3×3 convolution kernel. For both ResNet-12 and ResNet-18, the model was trained under a 10-way 5-shot setting, and a total of 1,200 epochs of training were conducted. The SGD optimizer was used on all datasets, with the initial learning rate set to 0.1 and the weight decay set to 0.0005. α and β are learnable hyperparameters, initialized to 0 and optimized jointly with the rest of the model. Following the settings in [15,41], the temperature coefficient τ in the loss function is also treated as a learnable hyperparameter, initialized to 1. In the validation stage, the strategy of validating once every 20 epochs was adopted, and 1,000 episodes were carried out each time to obtain the average accuracy. In the test phase, evaluation was conducted under the 5-way 1-shot and 5-way 5-shot settings. The test accuracy was obtained as the average accuracy across 10,000 episodes, where each episode contains 16 query samples.

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

Tables 2 and 3 list the classification performance of different few-shot learning methods based on the ResNet-12 backbone and the ResNet-18 backbone, respectively. Unless explicitly marked, all comparison experiments were implemented using the original authors' source code under the same settings as this paper.

As shown in Table 2, the proposed ITAM performs the best in seven out of eight settings. On the CUB dataset, ITAM achieves the best performance in the 1-shot setting at 83.72 % and in the 5-shot setting at 93.34 %, outperforming the second-best method by 0.35 % and 0.23 %. On the Flowers dataset, ITAM achieves the best performance in the 1-shot setting at 83.52 % and in the 5-shot setting at 94.36 %, outperforming the second-best method TDM [43] by 0.67 % and 0.76 %. On the Dogs dataset, ITAM achieves the best performance in the 1-shot setting at 76.73 % and in the 5-shot setting at 88.68 %, outperforming the second-best method by 0.24 % and 0.23 % in the 1-shot and 5-shot settings, respectively. Compared with the baseline method FRN [15], our method, including the additional component ITAM, consistently performs better.

A similar pattern can also be observed in Table 3, where ITAM achieves the best accuracy in five cases and the second best in the remaining cases. A further remark is that ITAM consistently and substantially outperforms AGPF [17], a method dedicated to fine granularity. All of these showcase the effectiveness of ITAM for few-shot fine-grained classification.

In addition to the standard evaluation of few-shot fine-grained classification, we investigate the effectiveness of ITAM in cross-domain scenarios, where there exists a domain gap between the base and novel data. Table 4 lists the cross-domain performance of ITAM and some comparison methods, where all models were trained on the mini-ImageNet dataset and tested on the CUB dataset. It can be seen that ITAM performs the best among these methods, which further verifies its superiority.

Table 2
Comparison of 5-way few-shot classification accuracy with ResNet-12 as the backbone. † represents the results reported in the original paper (missing performance is indicated by –). Average accuracy and 95 % confidence interval are reported. The best is in bold, and the second best is underlined.

Methods	CUB	Flowers	Cars	Dogs
	5-way 1-shot accuracy (%)			
RENet (ICCV-21) [42]	79.49$_{\pm 0.44}$	74.96$_{\pm 0.50}$	84.29$_{\pm 0.39}$	69.48$_{\pm 0.48}$
FRN (CVPR-21) [15]	83.16$_{\pm 0.19}$	81.07$_{\pm 0.20}$	86.48$_{\pm 0.18}$	76.49$_{\pm 0.21}$
HelixFormer (MM-22) [16]	81.66$_{\pm 0.30}$	63.30$_{\pm 0.26}$	79.40$_{\pm 0.43}$	65.92$_{\pm 0.49}$
TDM (CVPR-22) [43]	82.41$_{\pm 0.19}$	82.85$_{\pm 0.19}$	87.04$_{\pm 0.17}$	76.11$_{\pm 0.20}$
AGPF (PR-22) [17]	78.54$_{\pm 0.83}$	77.92$_{\pm 0.94}$	83.94$_{\pm 0.76}$	72.06$_{\pm 0.91}$
MCL (CVPR-22) [44]	83.25$_{\pm 0.25}$	76.55$_{\pm 0.26}$	85.04$_{\pm 0.36}$	71.49$_{\pm 0.28}$
BiFRN (AAAI-23) [45]	82.90$_{\pm 0.19}$	80.30$_{\pm 0.20}$	87.80$_{\pm 0.16}$	74.73$_{\pm 0.21}$
BSFA (TCSVT-23) [46]	83.11$_{\pm 0.41}$	75.33$_{\pm 0.54}$	**88.78**$_{\pm 0.38}$	73.54$_{\pm 0.50}$
QSFormer (TCSVT-23) [47]	75.26$_{\pm 0.17}$	75.24$_{\pm 0.25}$	80.02$_{\pm 0.28}$	68.87$_{\pm 0.72}$
C2-Net (AAAI-24) [48]	83.37$_{\pm 0.42}$	80.86$_{\pm 0.46}$	84.42$_{\pm 0.43}$	75.50$_{\pm 0.49}$
KLSANet (NN-24) [19]†	74.97$_{\pm 0.43}$	–	74.43$_{\pm 0.76}$	64.43$_{\pm 0.81}$
FicNet (TMM-24) [49]†	80.97$_{\pm 0.57}$	–	86.81$_{\pm 0.47}$	72.41$_{\pm 0.64}$
SAA (TCSVT-24) [50]†	75.57$_{\pm 0.48}$	74.22$_{\pm 0.49}$	–	70.32$_{\pm 0.50}$
SRML (PR-25) [51]†	83.05$_{\pm 0.43}$	79.82$_{\pm 0.47}$	87.49$_{\pm 0.36}$	72.97$_{\pm 0.47}$
ITAM	**83.72**$_{\pm 0.19}$	**83.52**$_{\pm 0.19}$	87.06$_{\pm 0.17}$	**76.73**$_{\pm 0.20}$
	5-way 5-shot accuracy (%)			
RENet (ICCV-21) [42]	91.11$_{\pm 0.24}$	81.95$_{\pm 0.38}$	91.94$_{\pm 0.23}$	81.75$_{\pm 0.35}$
FRN (CVPR-21) [15]	92.59$_{\pm 0.11}$	92.52$_{\pm 0.11}$	94.78$_{\pm 0.08}$	88.22$_{\pm 0.12}$
HelixFormer (MM-22) [16]	91.83$_{\pm 0.17}$	66.96$_{\pm 0.22}$	92.26$_{\pm 0.15}$	80.65$_{\pm 0.36}$
TDM (CVPR-22) [43]	92.37$_{\pm 0.10}$	93.60$_{\pm 0.10}$	96.11$_{\pm 0.07}$	88.45$_{\pm 0.11}$
AGPF (PR-22) [17]	89.85$_{\pm 0.44}$	91.96$_{\pm 0.45}$	94.11$_{\pm 0.36}$	84.83$_{\pm 0.50}$
MCL (CVPR-22) [44]	93.01$_{\pm 0.16}$	90.31$_{\pm 0.19}$	93.92$_{\pm 0.21}$	85.24$_{\pm 0.23}$
BiFRN (AAAI-23) [45]	93.11$_{\pm 0.10}$	92.30$_{\pm 0.11}$	96.49$_{\pm 0.06}$	87.76$_{\pm 0.12}$
BSFA (TCSVT-23) [46]	93.08$_{\pm 0.23}$	86.90$_{\pm 0.36}$	95.31$_{\pm 0.20}$	85.70$_{\pm 0.33}$
QSFormer (TCSVT-23) [47]	86.42$_{\pm 0.19}$	87.81$_{\pm 0.60}$	91.13$_{\pm 0.13}$	83.56$_{\pm 0.45}$
C2-Net (AAAI-24) [48]	92.20$_{\pm 0.23}$	91.54$_{\pm 0.27}$	92.72$_{\pm 0.23}$	87.65$_{\pm 0.28}$
KLSANet (NN-24) [19]†	88.92$_{\pm 0.41}$	–	87.84$_{\pm 0.45}$	81.07$_{\pm 0.31}$
FicNet (TMM-24) [49]†	93.17$_{\pm 0.32}$	–	95.36$_{\pm 0.22}$	85.11$_{\pm 0.37}$
SAA (TCSVT-24) [50]†	88.03$_{\pm 0.29}$	90.19$_{\pm 0.28}$	–	84.61$_{\pm 0.32}$
SRML (PR-25) [51]†	92.74$_{\pm 0.23}$	91.97$_{\pm 0.26}$	95.34$_{\pm 0.16}$	86.01$_{\pm 0.30}$
ITAM	**93.34**$_{\pm 0.10}$	**94.36**$_{\pm 0.09}$	**96.97**$_{\pm 0.06}$	**88.68**$_{\pm 0.11}$

Table 3
Comparison of 5-way few-shot classification accuracy with ResNet-18 as the backbone. † represents the results reported in the original paper (missing performance is indicated by -). Other captions are as in Table 2.

Methods	CUB	Flowers	Cars	Dogs
	5-way 1-shot accuracy (%)			
MatchingNet (NeurIPS-16) [8]	72.88$_{\pm 0.89}$	76.07$_{\pm 0.82}$	75.03$_{\pm 0.95}$	65.59$_{\pm 0.95}$
RelationNet (CVPR-18) [11]	68.82$_{\pm 1.04}$	69.04$_{\pm 0.97}$	64.08$_{\pm 1.05}$	54.21$_{\pm 1.00}$
Baseline++ (CVPR-19) [52]	65.67$_{\pm 0.95}$	67.90$_{\pm 0.96}$	67.41$_{\pm 0.99}$	62.54$_{\pm 0.87}$
Neg-Margin (ECCV-20) [53]	72.51$_{\pm 0.82}$	76.34$_{\pm 0.89}$	76.04$_{\pm 0.81}$	68.86$_{\pm 0.83}$
FRN (CVPR-21) [15]	83.40$_{\pm 0.19}$	81.22$_{\pm 0.21}$	87.63$_{\pm 0.17}$	77.53$_{\pm 0.21}$
TDM (CVPR-22) [43]	83.25$_{\pm 0.19}$	82.31$_{\pm 0.20}$	87.69$_{\pm 0.17}$	76.59$_{\pm 0.21}$
AGPF (PR-22) [17]	79.02$_{\pm 0.83}$	78.69$_{\pm 0.84}$	84.68$_{\pm 0.78}$	73.61$_{\pm 0.91}$
DeepBDC (CVPR-22) [10]	81.85$_{\pm 0.42}$	81.07$_{\pm 0.50}$	85.48$_{\pm 0.40}$	**78.81**$_{\pm 0.43}$
LCCRN (TCSVT-23) [18]	82.80$_{\pm 0.19}$	82.86$_{\pm 0.19}$	86.24$_{\pm 0.18}$	77.29$_{\pm 0.20}$
QGN (PR-23) [54]†	83.82	–	–	–
ITAM	**84.09**$_{\pm 0.18}$	**82.88**$_{\pm 0.19}$	**87.97**$_{\pm 0.17}$	78.36$_{\pm 0.20}$
	5-way 5-shot accuracy (%)			
MatchingNet (NeurIPS-16) [8]	85.25$_{\pm 0.57}$	87.46$_{\pm 0.51}$	87.02$_{\pm 0.56}$	80.94$_{\pm 0.60}$
RelationNet (CVPR-18) [11]	82.68$_{\pm 0.58}$	85.46$_{\pm 0.58}$	91.45$_{\pm 0.44}$	80.42$_{\pm 0.62}$
Baseline++ (CVPR-19) [52]	81.53$_{\pm 0.58}$	84.34$_{\pm 0.62}$	85.50$_{\pm 0.58}$	79.04$_{\pm 0.61}$
Neg-Margin (ECCV-20) [53]	89.25$_{\pm 0.43}$	90.83$_{\pm 0.47}$	93.06$_{\pm 0.38}$	85.75$_{\pm 0.52}$
FRN (CVPR-21) [15]	92.69$_{\pm 0.10}$	92.33$_{\pm 0.11}$	95.35$_{\pm 0.08}$	89.05$_{\pm 0.11}$
TDM (CVPR-22) [43]	92.98$_{\pm 0.10}$	93.46$_{\pm 0.11}$	96.06$_{\pm 0.17}$	88.87$_{\pm 0.11}$
AGPF (PR-22) [17]	89.92$_{\pm 0.42}$	92.78$_{\pm 0.40}$	94.87$_{\pm 0.33}$	85.68$_{\pm 0.52}$
DeepBDC (CVPR-22) [10]	93.00$_{\pm 0.24}$	93.19$_{\pm 0.24}$	95.84$_{\pm 0.16}$	91.33$_{\pm 0.22}$
LCCRN (TCSVT-23) [18]	**93.60**$_{\pm 0.10}$	93.87$_{\pm 0.10}$	96.34$_{\pm 0.07}$	**89.54**$_{\pm 0.10}$
QGN (PR-23) [54]†	91.22	89.9	91.3	–
ITAM	93.41$_{\pm 0.10}$	**94.12**$_{\pm 0.09}$	**97.18**$_{\pm 0.06}$	89.15$_{\pm 0.11}$

Table 4
The 5-way few-shot classification accuracy in the cross-domain setting: mini-ImageNet → CUB. The accuracy of the methods labeled ◇ is quoted from Ref. [55], and those labeled △ are quoted from Ref. [15]. ResNet-12 was used as the backbone for all methods.

Methods	1-shot	5-shot
ProtoNet (NeurIPS-17) [9]◇	47.51 ± 0.38	67.96 ± 0.70
MetaOptNet (CVPR-19) [39]△	44.79 ± 0.75	64.98 ± 0.68
FEAT (CVPR-20) [35]◇	50.67 ± 0.78	71.08 ± 0.73
SCL (ECCV-20) [40]◇	49.58 ± 0.70	68.81 ± 0.60
FRN (CVPR-21) [15]△	54.11 ± 0.19	77.09 ± 0.15
BSFA (TCSVT-23) [46]	55.72 ± 0.56	67.49 ± 0.48
SRCPT (ICMLC-24) [55]◇	54.73 ± 0.22	75.97 ± 0.18
ITAM	**62.81** ± 0.20	**79.17** ± 0.14

4.4. Ablation studies

In this section, the effectiveness of ITAM was demonstrated by first comparing it with other attention modules while keeping the feature extraction network and metric module fixed. Then, each branch of ITAM was removed to assess the impact of individual dimension interactions on classification performance, and the design of the interaction strategy was also investigated. Next, ITAM was inserted into different metric-based few-shot learning methods to showcase its flexibility as a plug-and-play module. Finally, the influence of different shot numbers was evaluated to understand the universality of the method.

Comparison with other attentions. First, in order to verify the rationality of the attention mechanism selected in this paper, we replaced ITAM in the network with other attention modules and conducted comparative experiments on CUB and Flowers. As shown in Table 5, TAM outperforms the existing attention methods in most cases, which indicates the effectiveness of considering the interaction of different dimensions. On top of TAM, the proposed ITAM achieves even better performance, demonstrating that the full interaction across all dimensions can further enhance the extraction of highly discriminative features.

The importance of each dimension interaction. To demonstrate the validity of each dimension interaction, ablation studies were conducted on the branches of ITAM: removing all branches (which is equivalent to FRN); keeping a single branch for the pairwise interaction between (H, C), (C, W), (H, W) dimensions; keeping two branches; or keeping all three branches (i.e., the proposed ITAM). Note that when two branches are kept, the interaction between the two branches can still be computed, although for only one pair. Table 6 lists the experimental results of 5-way 1-shot and 5-way 5-shot on CUB and Flowers, under the ResNet-12 backbone.

The table presents the following observations. Firstly, using a single pairwise interaction generally improves performance over the baseline (a), although the effect varies across datasets and tasks. The pairwise interaction between (H, W) yields the best performance in most cases, highlighting the importance of spatial information. Secondly, distinct performances can be noticed when considering two pairwise interactions. Some combinations, such as (e) and (g), lead to lower accuracy on CUB, particularly in the 1-shot setting. In contrast, the interaction between the second and third branches (i.e., (f)) consistently improves performance. Constructing a third-order interaction using only two branches is likely to overemphasize one dimension, such as the channel dimension in the case of (H, C) and (C, W), which we hypothesize contributes to the poorer results. The best performance is achieved with the proposed strategy, which jointly integrates all three interaction branches. This indicates that the third-order attention mechanism is more than just a sum of its individual components; it benefits from coordinated learning across spatial and channel dimensions, enabling richer feature representations. Moreover, Table 4 shows that ITAM

Table 5
Comparison of 5-way few-shot classification performance under different attention mechanisms. Both ResNet-12 and ResNet-18 were considered as the backbone.

Methods	CUB		Flowers	
	1-shot	5-shot	1-shot	5-shot
ResNet-12				
FRN	$83.16_{\pm 0.19}$	$92.59_{\pm 0.11}$	$81.07_{\pm 0.20}$	$92.52_{\pm 0.11}$
+SeNet (CVPR-18) [56]	$83.21_{\pm 0.19}$	$92.61_{\pm 0.10}$	$80.22_{\pm 0.21}$	$91.46_{\pm 0.12}$
+CBAM (ECCV-18) [25]	$83.31_{\pm 0.19}$	$92.52_{\pm 0.11}$	$81.27_{\pm 0.21}$	$92.16_{\pm 0.12}$
+GCNet (ICCV-19) [26]	$82.77_{\pm 0.19}$	$92.25_{\pm 0.10}$	$80.93_{\pm 0.21}$	$91.94_{\pm 0.12}$
+ECA (CVPR-20) [27]	$83.07_{\pm 0.19}$	$92.37_{\pm 0.11}$	$80.71_{\pm 0.21}$	$91.92_{\pm 0.12}$
+CA (ICCV-21) [28]	$80.78_{\pm 0.20}$	$91.23_{\pm 0.11}$	$81.65_{\pm 0.21}$	$92.40_{\pm 0.11}$
+NAM (CVPR-21) [29]	$80.83_{\pm 0.20}$	$91.28_{\pm 0.11}$	$78.28_{\pm 0.22}$	$90.83_{\pm 0.13}$
+TAM (WACV-21) [7]	$83.22_{\pm 0.19}$	$92.94_{\pm 0.10}$	$82.18_{\pm 0.20}$	$93.03_{\pm 0.11}$
+EMA (ICASSP-23) [30]	$83.27_{\pm 0.19}$	$92.64_{\pm 0.10}$	$81.04_{\pm 0.21}$	$92.07_{\pm 0.12}$
+SRU (CVPR-23) [57]	$83.15_{\pm 0.19}$	$92.02_{\pm 0.11}$	$76.59_{\pm 0.22}$	$90.02_{\pm 0.13}$
+ELA (arXiv-24) [31]	$83.70_{\pm 0.19}$	$92.53_{\pm 0.10}$	$81.30_{\pm 0.20}$	$92.50_{\pm 0.11}$
+AA (ECCV-24) [32]	$82.90_{\pm 0.19}$	$93.02_{\pm 0.19}$	$82.92_{\pm 0.20}$	$92.96_{\pm 0.20}$
+CA (CEA-25) [33]	$83.06_{\pm 0.19}$	$93.08_{\pm 0.19}$	$83.12_{\pm 0.19}$	$93.24_{\pm 0.19}$
+ITAM	$\mathbf{83.72}_{\pm 0.19}$	$\mathbf{93.34}_{\pm 0.10}$	$\mathbf{83.52}_{\pm 0.19}$	$\mathbf{94.36}_{\pm 0.09}$
ResNet-18				
FRN	$83.40_{\pm 0.19}$	$92.69_{\pm 0.10}$	$81.22_{\pm 0.21}$	$92.33_{\pm 0.11}$
+SeNet (CVPR-18) [56]	$83.61_{\pm 0.19}$	$92.75_{\pm 0.10}$	$80.90_{\pm 0.21}$	$92.30_{\pm 0.11}$
+CBAM (ECCV-18) [25]	$82.98_{\pm 0.19}$	$92.29_{\pm 0.10}$	$81.02_{\pm 0.21}$	$91.93_{\pm 0.12}$
+GCNet (ICCV-19) [26]	$82.85_{\pm 0.19}$	$92.26_{\pm 0.11}$	$80.89_{\pm 0.21}$	$92.09_{\pm 0.11}$
+ECA (CVPR-20) [27]	$83.07_{\pm 0.19}$	$92.37_{\pm 0.11}$	$80.71_{\pm 0.21}$	$91.92_{\pm 0.12}$
+CA (ICCV-21) [28]	$83.55_{\pm 0.19}$	$92.67_{\pm 0.10}$	$82.00_{\pm 0.20}$	$92.91_{\pm 0.11}$
+NAM (CVPR-21) [29]	$81.15_{\pm 0.20}$	$91.10_{\pm 0.11}$	$79.19_{\pm 0.21}$	$91.18_{\pm 0.13}$
+TAM (WACV-21) [7]	$84.00_{\pm 0.18}$	$93.01_{\pm 0.10}$	$82.19_{\pm 0.20}$	$93.18_{\pm 0.11}$
+EMA (ICASSP-23) [30]	$83.30_{\pm 0.19}$	$92.49_{\pm 0.10}$	$81.04_{\pm 0.21}$	$92.13_{\pm 0.12}$
+SRU (CVPR-23) [57]	$80.66_{\pm 0.20}$	$90.80_{\pm 0.12}$	$75.21_{\pm 0.21}$	$89.40_{\pm 0.19}$
+ELA (arXiv-24) [31]	$82.69_{\pm 0.19}$	$92.36_{\pm 0.10}$	$81.52_{\pm 0.19}$	$92.64_{\pm 0.11}$
+AA (ECCV-24) [32]	$83.22_{\pm 0.19}$	$92.61_{\pm 0.10}$	$82.71_{\pm 0.19}$	$93.69_{\pm 0.11}$
+CA (CEA-25) [33]	$83.46_{\pm 0.19}$	$93.06_{\pm 0.10}$	$82.55_{\pm 0.19}$	$94.09_{\pm 0.11}$
+ITAM	$\mathbf{84.09}_{\pm 0.18}$	$\mathbf{93.41}_{\pm 0.10}$	$\mathbf{82.88}_{\pm 0.19}$	$\mathbf{94.12}_{\pm 0.09}$

Table 6
Ablation studies on each dimension interaction. ResNet-12 is the backbone.

	(H,C)	(C,W)	(H,W)	CUB		Flowers	
				1-shot	5-shot	1-shot	5-shot
(a)	×	×	×	83.16	92.59	81.07	92.52
(b)	✓	×	×	82.92	92.62	81.53	92.43
(c)	×	✓	×	83.13	92.53	81.97	92.94
(d)	×	×	✓	83.09	92.86	82.31	92.98
(e)	✓	✓	×	72.45	86.39	82.72	93.73
(f)	×	✓	✓	83.30	93.08	82.79	93.91
(g)	✓	×	✓	77.25	88.94	78.43	91.16
ITAM	✓	✓	✓	**83.72**	**93.34**	**83.52**	**94.36**

consistently outperforms TAM, supporting the view that the full third-order interaction is more effective than a combination of first-order interactions.

To model third-order interactions, we adopt a simple parameter-free strategy based on multiplication and averaging operations. Table 7 compares this approach with more advanced interaction and fusion techniques – cascade operations [58], adaptive weighting [59], and non-linear fusion [60]. The experimental results show that the proposed ITAM achieves superior performance across multiple configurations, with particularly notable results on the Flowers dataset – reaching 83.52 % and 94.36 % in the 1-shot and 5-shot settings, respectively, outperforming other methods by approximately 2 %. In addition, it also achieves the highest accuracy of 93.34 % on the 5-shot task of the CUB dataset. These results suggest that the multiplication and averaging design, though simple, is highly effective.

Flexibility as a plug-and-play component. To verify the flexibility of the proposed method as a plug-and-play component, TAM and ITAM

Table 7
The comparison of 5-way 1-shot and 5-way 5-shot accuracies under different interaction fusion methods on the CUB and Flowers datasets.

Methods	CUB		Flowers	
	1-shot	5-shot	1-shot	5-shot
Cascade Operations	$80.91_{\pm 0.20}$	$91.37_{\pm 0.11}$	$81.30_{\pm 0.20}$	$92.45_{\pm 0.11}$
Adaptive Weighting	$83.69_{\pm 0.19}$	$93.01_{\pm 0.10}$	$81.38_{\pm 0.21}$	$92.28_{\pm 0.11}$
Non-linear Fusion	$\mathbf{83.88}_{\pm 0.18}$	$93.33_{\pm 0.10}$	$80.96_{\pm 0.21}$	$92.28_{\pm 0.11}$
ITAM	$83.72_{\pm 0.19}$	$\mathbf{93.34}_{\pm 0.10}$	$\mathbf{83.52}_{\pm 0.19}$	$\mathbf{94.36}_{\pm 0.09}$

Table 8
Flexibility as a plug-and-play component for metric-based classifiers. ResNet-12 is the backbone.

Methods	CUB		Flowers		Cars	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [9]	79.64	91.15	75.41	89.46	82.29	93.11
+TAM	80.97	90.41	76.59	88.47	82.33	91.32
+ITAM	**81.52**	**91.26**	**77.13**	**89.52**	**87.89**	**94.98**
TDM [43]	82.41	92.37	82.85	93.60	87.04	96.11
+TAM	82.98	93.01	**83.27**	**94.00**	88.62	96.83
+ITAM	**83.58**	**93.11**	83.13	93.99	**89.19**	**97.35**
FRN [15]	83.16	92.59	81.07	92.52	86.48	94.78
+TAM	83.22	92.94	82.18	93.03	86.56	95.51
+ITAM	**83.72**	**93.34**	**83.52**	**94.36**	**87.06**	**96.97**

Table 9
Classification accuracy under different numbers of shots on the CUB and Flowers datasets. ResNet-12 is the backbone.

Methods	CUB				
	1-shot	3-shot	5-shot	7-shot	9-shot
FRN [15]	83.15	90.86	92.45	93.12	93.54
FRN+ITAM	**83.31**	**91.29**	**92.93**	**93.69**	**94.14**
Methods	Flowers				
	1-shot	3-shot	5-shot	7-shot	9-shot
FRN [15]	81.86	90.44	92.66	93.78	94.39
FRN+ITAM	**82.96**	**91.61**	**93.88**	**95.02**	**95.69**

were added to three different metric-based few-shot learning methods: ProtoNet [9], TDM [43], and FRN [15]. As shown in Table 8, TAM, despite not being applied to fine-grained image classification, is effective in enhancing the classification performance in most cases. This suggests that the cross-dimension interaction is effective in capturing discriminative representations at fine details. The proposed ITAM consistently leads to (mostly the best) performance gains over the baseline method, highlighting its suitability for fine-grained image classification and its potential to be integrated with other metric-based few-shot learning methods.

The impact of the number of shots. Table 9 reports the classification accuracy under different numbers of shots on the CUB and Flowers datasets, while fixing the training settings as described in Section 4.2. It is evident that on the CUB and Flowers datasets, the accuracy of ITAM compared with FRN on 1/3/5/7/9-shot tasks continues to improve.

4.5. Visualization analysis

The previous sections quantitatively demonstrated that the proposed FRN+ITAM outperforms the baseline FRN. This section illustrates how attention influences feature learning and its impact on feature reconstruction, and consequently, classification.

Visualization of feature maps. Since the goal of FRN is to minimize the reconstruction error, the magnitude of local features in a feature map

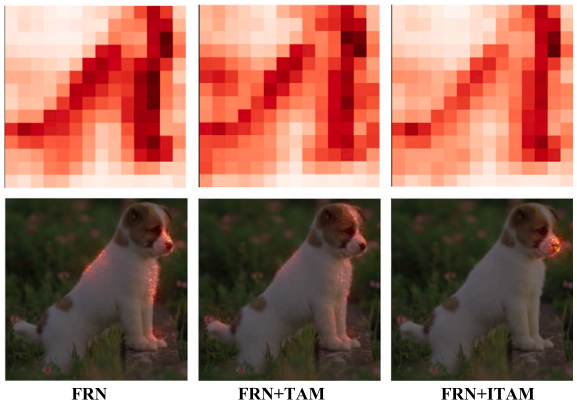


Fig. 4. Investigation into feature map used for feature reconstruction, i.e., the features extracted by the ResNet-18 backbone (shown in the column of *FRN*) and the features enhanced by TAM and ITAM (*FRN + TAM* and *FRN + ITAM*, resp.). Top: Heatmaps of L_2-norm computed over the channel dimension of the feature map. Warmer color indicates larger value. Bottom: Saliency maps for the top 5 % spatial positions with the highest L_2-norm, overlaid on the raw image.

influences the focus of reconstruction: if local features of a query image all have similar magnitudes, then each local feature will be equally recovered; conversely, if some local features have much higher magnitudes compared to others, then these features will have larger influence on the reconstruction error and thus will be paid more attention. The top row of Fig. 4 presents heatmaps that visualize the L_2-norm of local features. More specifically, for FRN, L_2-norm is computed for the feature vector at each spatial position in the embedded features $\hat{x}$, resulting in a 2D L_2-norm map. For FRN + TAM and FRN + ITAM, L_2-norm is computed over the enhanced features $\tilde{x}$. By comparing these three heatmaps, we see that in FRN + ITAM, the L_2-norm values exhibit greater contrast, with fewer spatial positions having high L_2-norm values. This indicates that feature reconstruction in FRN + ITAM is more focused, prioritizing a smaller subset of local features with high importance.

The bottom row of Fig. 4 displays saliency maps for the top 5 % spatial positions with the highest L_2-norm, overlaid on the original image. Red pixels in these maps indicate areas of the image corresponding to local features with the largest L_2-norm values. The saliency maps validate that FRN + ITAM focuses on the correct regions of the input image, i.e., the dog's nose in this example.

Visualization of feature reconstruction. According to the classification mechanism of FRN, it assigns a query image to the class with the smallest reconstruction error, so FRN + ITAM of higher accuracy implies a smaller reconstruction error. However, it is unclear what drives the decrease in the reconstruction error. To answer this question, we randomly select one sample from each dataset for feature reconstruction visualization. An inverted ResNet-12 is trained as a decoder to map features to the original image. In Fig. 5, the leftmost column shows the original images of four datasets, where each row shows one support image. The second leftmost column shows the images generated by the inverted ResNet-12 from the corresponding support features, where the features were extracted by using the ResNet-12 backbone. It can be seen that the inverted ResNet-12 can recover the original images from features to a good extent. The third and fourth leftmost columns show the original query images and the corresponding query features. The rightmost three columns show the query features reconstructed from support features, with features learned by using FRN, FRN + TAM and FRN + ITAM.

From the three columns of reconstructed query features, it can be observed that all three methods effectively reconstruct the coarse details of the target of interest. However, due to the interference of background information and other noise, FRN pays limited attention to the locally important region of the sample, resulting in some fine details being missed, such as the tire of the car and the face of the dog. It also generates some artifacts, such as those observed around the outline of the bird. When incorporating TAM into FRN, some fine details are recovered, but the artifact issue still persists. The proposed method, FRN + ITAM, recovers the finest details and generates the fewest artifacts, indicating that by considering the full interaction across three dimensions, it can effectively extract discriminative features, facilitating feature reconstruction.

Visualization of feature discriminability. To supplement Fig. 1, Fig. 6 shows the heatmaps of the Flowers, Dogs and Cars datasets, generated by using Grad-CAM [61]. The figures show that regions corresponding to the target object receive high attention in all methods. However, FRN also focuses on some background regions, such as the grass in the Dogs case and trees in the Cars case. By allowing interaction between

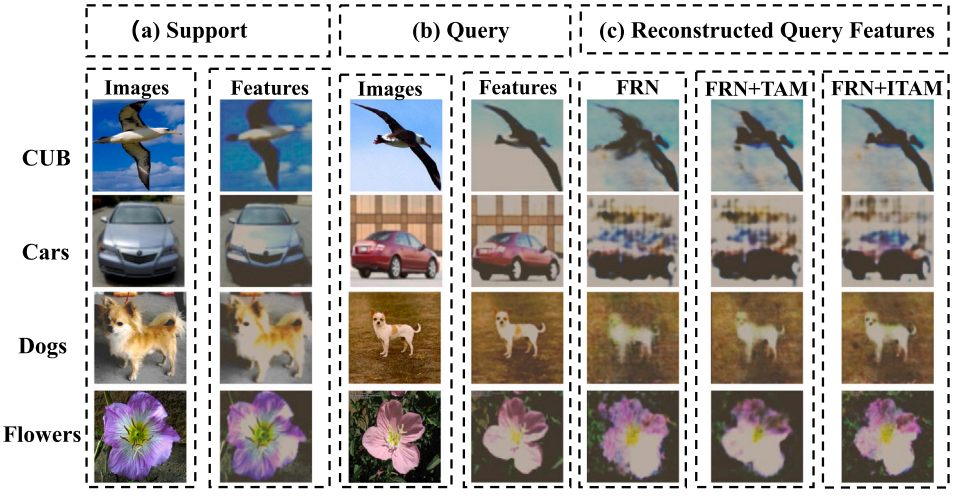


Fig. 5. Visualization of feature reconstruction. Panel (a) shows 4 original support images and 4 recovered support images generated by using the features after feature extraction. Panel (b) shows 4 original and recovered query images. Panel (c) shows images generated by using the reconstructed query features under FRN, FRN + TAM, and FRN + ITAM, respectively. It can be seen that the reconstructed images by using the proposed method (FRN + ITAM) contain fewer artifacts and more fine-grained details than those from FRN and FRN + TAM.

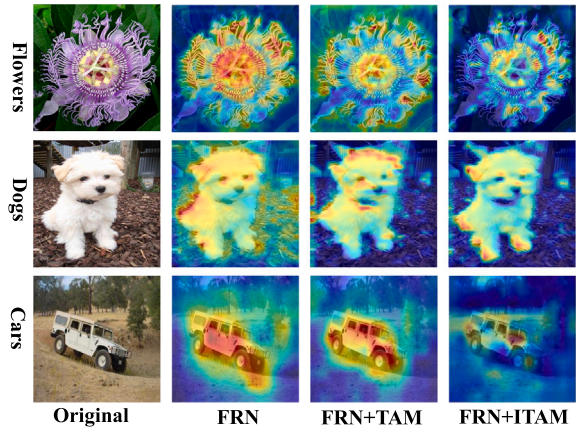


Fig. 6. Heatmaps of the Flowers, Dogs and Cars datasets, generated by using Grad-CAM [61]. Warmer color indicates more important regions. Compared with FRN and FRM + TAM, FRN + ITAM places more focused and compact attention on key feature regions.

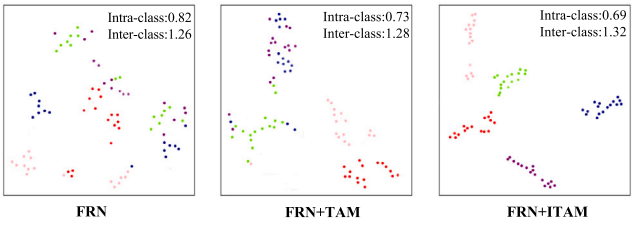


Fig. 7. t-SNE visualizations of feature embeddings learned by FRN, FRN + TAM, and FRN + ITAM on the CUB dataset. The intra-class and inter-class Euclidean distances are displayed in the top-right corner of each subfigure. ResNet-12 was used as backbone.

the three dimensions, the influence from the background is reduced, as demonstrated by FRN + TAM. Further increasing the interaction, i.e., FRM + ITAM, leads the method to place more focused and compact attention on key feature areas, highlighting fewer but more discriminative regions.

Fig. 7 presents the visualization of feature embeddings learned by FRN, FRN + TAM, and FRN + ITAM on the CUB dataset, where t-SNE [62] was employed to reduce the dimensionality of high-dimensional features. Each point in the figure represents a query sample from a randomly selected episode, and different colors indicate different classes. As shown in the figure, in FRN, samples from the same class fail to stay close together (such as those represented by green, pink, and black), indicating low intra-class compactness; samples from different classes locate closely (such as those represented by red and purple), indicating small inter-class separability. FRN + TAM improves the feature representations to a certain extent, with some classes (such as red and pink) well separated from others. However, samples of the same class are still rather scattered. In contrast, FRN + ITAM substantially improves the embedding space: samples from the same class are positioned more closely together (such as the green and black classes), while samples from different classes are more clearly separated. This demonstrates the enhancement of intra-class compactness and inter-class separability, which plays a crucial role in improving the performance of few-shot fine-grained classification tasks.

To quantitatively evaluate the quality of feature embeddings, intra-class and inter-class Euclidean distances were calculated, with the results displayed in the top-right corner of each subfigure. The intra-class distance is calculated as the average Euclidean distance among feature embeddings within the same class, further averaged across all five classes. A smaller intra-class distance indicates that samples from the same class locate more closely in the feature space. The inter-class distance is calculated as the average Euclidean distance between each sample and all samples from different classes. A larger inter-class distance suggests that samples from different classes are more widely separated in the feature space. From FRN to FRN + TAM and FRN + ITAM, the intra-class distance continuously decreases (from 0.82 to 0.69), and the inter-class distance steadily increases (from 1.26 to 1.32). These results suggest that considering feature interaction can help learn more discriminative feature embeddings.

Table 10
Comparison of computational complexity. A smaller number of parameters and fewer FLOPs indicate better efficiency.

Methods	Backbone	Model complexity	
		Params. (M)	FLOPs (G)
FRN	ResNet-12	12.42	704.60
FRN + TAM	ResNet-12	12.42	704.73
FRN + ITAM	ResNet-12	12.42	704.73

4.6. Comparison of computational complexity

Besides classification accuracy, the computational cost is also crucial from a practical perspective. Table 10 reports the number of parameters in the model and floating-point operation (FLOPs) for FRN, FRN + TAM, and FRN + ITAM. The FLOPs are measured during the training phase in 10-way 5-shot tasks with 15 query samples per class using the CUB-200-2011 dataset. The results clearly suggest that ITAM incurs little extra computational complexity.

5. Conclusion

This paper extends the application of triplet attention to the domain of fine-grained few-shot image classification, which encodes spatial and channel information and captures the interactions between spatial and channel dimensions when computing attention weights. It further enhances the attention mechanism, referred to as ITAM, to establish the full interactions between all three dimensions. ITAM incurs negligible computational overhead and can be readily integrated into metric-based few-shot learning methods. Experiments on four fine-grained image datasets demonstrate that the proposed approach achieves state-of-the-art performance and exhibits high flexibility.

This work focuses solely on few-shot fine-grained image classification. However, considering the efficacy, efficiency, and ease of integration with other methods, the proposed method has potential for other few-shot image classification tasks, particularly those involving subtle differences between images, such as medical image classification. In the future, we will also explore extending ITAM to capture the complex relationships between different modalities, such as image and text, and improving cross-modal fusion.

CRediT authorship contribution statement

Xiaoxu Li: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Shaoying Xue:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Jiyang Xie:** Validation, Methodology, Investigation, Data curation. **Xiaochen Yang:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Zhanyu Ma:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization. **Jing-Hao Xue:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the Beijing Natural Science Foundation Project No. L242025, in part by the Royal Society under International Exchanges Award IEC\NSFC\201071, IEC\NSFC\211131, in part by the National Natural Science Foundation of China (NSFC) under Grant 62176110, 62225601, U23B2052, in part by the Key Talent Program of Gansu Province under Grant 2025RCXM002, the S&T Program of Hebei under Grant SZX2020034, and Hong-Liu Distinguished Young Talents Foundation of Lanzhou University of Technology, China.

Data availability

All datasets are publicly available.

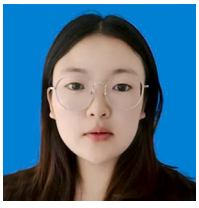
References

[1] X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, Q. Li, Few-shot learning for domain-specific fine-grained image classification, IEEE Trans. Ind. Electron. 68 (4) (2020) 3588–3598.
[2] Y. Ge, Y. Guo, S. Das, M.A. Al-Garadi, A. Sarker, Few-shot learning for medical text: a review of advances, trends, and opportunities, J. Biomed. Inform. 144 (2023) 104458.
[3] Y. Song, T. Wang, P. Cai, S.K. Mondal, J.P. Sahoo, A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities, ACM Comput. Surv. 55 (13s) (2023) 1–40.
[4] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: a review of recent developments, Pattern Recognit. 138 (2023) 109381.
[5] H. Gharoun, F. Momenifar, F. Chen, A. Gandomi, Meta-learning approaches for few-shot learning: a survey of recent advances, ACM Comput. Surv. 56 (12) (2024) 1–41.
[6] J. Ren, C. Li, Y. An, W. Zhang, C. Sun, Few-shot fine-grained image classification: a comprehensive review, AI 5 (1) (2024) 405–425.
[7] D. Misra, T. Nalamada, A.U. Arasanipalai, Q. Hou, Rotate to attend: convolutional triplet attention module, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3139–3148.
[8] O. Vinyals, C. Blundell, T.P. Lillicrap, K. Kavukcuoglu, D. Wierstra, Matching networks for one shot learning, Adv. Neural Inf. Process. Syst. (2016).
[9] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Adv. Neural Inf. Process. Syst. 30 (2017).
[10] J. Xie, F. Long, J. Lv, Q. Wang, P. Li, Joint distribution matters: deep brownian distance covariance for few-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7972–7981.
[11] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H. Torr, T.M. Hospedales, Learning to compare: relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[12] Z. Sun, W. Zheng, M. Wang, SLTRN: sample-level transformer-based relation network for few-shot classification, Neural Netw. 176 (2024) 106344.
[13] J. Liu, Z. Ji, Y. Pang, Y. Yu, NTK-guided few-shot class incremental learning, IEEE Trans. Image Process. (2024).
[14] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, J. Luo, Revisiting local descriptor based image-to-class measure for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7260–7268.
[15] D. Wertheimer, L. Tang, B. Hariharan, Few-shot classification with feature map reconstruction networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8012–8021.
[16] B. Zhang, J. Yuan, B. Li, T. Chen, J. Fan, B. Shi, Learning cross-image object semantic relation in transformer for few-shot fine-grained image classification, in: Proceedings of the ACM International Conference on Multimedia, 2022, pp. 2135–2144.
[17] H. Tang, C. Yuan, Z. Li, J. Tang, Learning attention-guided pyramidal features for few-shot fine-grained recognition, Pattern Recognit. 130 (2022) 108792.
[18] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, J.-H. Xue, Locally-enriched cross-reconstruction for few-shot fine-grained image classification, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7530–7540.
[19] Z. Sun, W. Zheng, P. Guo, KLSANET: key local semantic alignment network for few-shot image classification, Neural Netw. 178 (2024) 106456.
[20] Z. Ji, Z. Wang, X. Liu, Y. Yu, Y. Pang, J. Han, Frequency-spatial complementation: unified channel-specific style attack for cross-domain few-shot learning, IEEE Trans. Image Process., 2025.
[21] X. Wang, Z. Ji, Y. Pang, Y. Yu, A cognition-driven framework for few-shot class-incremental learning, Neurocomputing 600 (2024) 128118.
[22] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, J.-H. Xue, BSNET: bi-similarity network for few-shot fine-grained image classification, IEEE Trans. Image Process. 30 (2021) 1318–1331.
[23] C. Zhang, Y. Cai, G. Lin, C. Shen, DeepEMD: few-shot image classification with differentiable earth mover's distance and structured classifiers, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12200–12210.
[24] C. Liang, S. Bai, Found missing semantics: supplemental prototype network for few-shot semantic segmentation, Comput. Vis. Image Underst. 249 (2024) 104191.
[25] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.
[26] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, GCNET: non-local networks meet squeeze-excitation networks and beyond, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[27] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.
[28] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.
[29] Y. Liu, Z. Shao, Y. Teng, N. Hoffmann, NAM: normalization-based attention module, in: NEURIPS 2021 Workshop on ImageNet: Past, Present, and Future, 2021.
[30] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, Z. Huang, Efficient multi-scale attention module with cross-spatial learning, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2023, pp. 1–5.
[31] W. Xu, Y. Wan, Ela: efficient local attention for deep convolutional neural networks, 2024, arXiv preprint arXiv:2403.01123.
[32] D. Han, T. Ye, Y. Han, Z. Xia, S. Pan, P. Wan, S. Song, G. Huang, Agent attention: on the integration of softmax and linear attention, in: European Conference on Computer Vision, Springer, 2024, pp. 124–140.
[33] S. Gupta, A.K. Tripathi, Flora-NET: integrating dual coordinate attention with adaptive kernel based convolution network for medicinal flower identification, Comput. Electron. Agric. 230 (2025) 109834.
[34] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD birds-200-2011 Dataset, Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.
[35] H.-J. Ye, H. Hu, D.-C. Zhan, F. Sha, Few-shot learning via embedding adaptation with set-to-set functions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8808–8817.
[36] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision Workshops, 2013, pp. 554–561.
[37] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel dataset for fine-grained image categorization: Stanford dogs, in: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[38] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008, pp. 722–729.
[39] K. Lee, S. Maji, A. Ravichandran, S. Soatto, Meta-learning with differentiable convex optimization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10657–10665.
[40] Y. Tian, Y. Wang, D. Krishnan, J.B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need?, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 266–282.
[41] Y. Chen, Z. Liu, H. Xu, T. Darrell, X. Wang, Meta-baseline: exploring simple meta-learning for few-shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9062–9071.
[42] D. Kang, H. Kwon, J. Min, M. Cho, Relational embedding for few-shot classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[43] S. Lee, W. Moon, J.-P. Heo, Task discrepancy maximization for fine-grained few-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5331–5340.
[44] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, X. He, Learning to affiliate: mutual centralized learning for few-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14411–14420.
[45] J. Wu, D. Chang, A. Sain, X. Li, Z. Ma, J. Cao, J. Guo, Y.-Z. Song, Bi-directional feature reconstruction network for fine-grained few-shot image classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 2821–2829.
[46] Z. Zha, H. Tang, Y. Sun, J. Tang, Boosting few-shot fine-grained recognition with background suppression and foreground alignment, IEEE Trans. Circuits Syst. Video Technol. (2023).
[47] X. Wang, X. Wang, B. Jiang, B. Luo, Few-shot learning meets transformer: unified query-support transformers for few-shot classification, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7789–7802.
[48] Z.-X. Ma, Z.-D. Chen, L.-J. Zhao, Z.-C. Zhang, X. Luo, X.-S. Xu, Cross-layer and cross-sample feature optimization network for few-shot fine-grained image classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 4136–4144.
[49] H. Zhu, Z. Gao, J. Wang, Y. Zhou, C. Li, Few-shot fine-grained image classification VIA multi-frequency neighborhood and double-cross modulation, IEEE Trans. Multimed. (2024).
[50] X. Li, X. Wang, R. Zhu, Z. Ma, J. Cao, J.-H. Xue, Selectively augmented attention network for few-shot image classification, IEEE Trans. Circuits Syst. Video Technol. (2024).
[51] X. Li, L. Wang, R. Zhu, Z. Ma, J. Cao, J.-H. Xue, SRML: structure-relation mutual learning network for few-shot image classification, Pattern Recognit. 168 (2025) 111822.

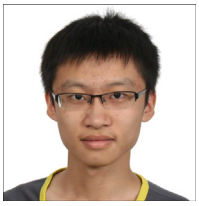
[52] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C.F. Wang, J.-B. Huang, A closer look at few-shot classification, in: International Conference on Learning Representations, 2019.
[53] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, H. Hu, Negative margin matters: understanding margin in few-shot classification, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 438–455.
[54] B. Munjal, A. Flaborea, S. Amin, F. Tombari, F. Galasso, Query-guided networks for few-shot fine-grained classification and person search, Pattern Recognit. 133 (2023) 109049.
[55] Z. Wang, P. Duan, Y. Rong, SRCPT: spatial reconstruction contrastive pretext task for improving few-shot image classification, in: Proceedings of the International Conference on Machine Learning and Computing, 2024, pp. 424–432.
[56] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[57] J. Li, Y. Wen, L. He, SCCONV: spatial and channel reconstruction convolution for feature redundancy, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6153–6162.
[58] Z. Chen, H. Wang, S. Zhang, F. Zhong, Dual-attention network for few-shot segmentation, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2022, pp. 2210–2214.
[59] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[60] R. Hou, H. Chang, B. Ma, S. Shan, X. Chen, Cross attention network for few-shot classification, Adv. Neural Inf. Process. Syst. 32 (2019).
[61] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[62] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

Author biography

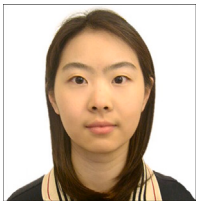
Xiaoxu Li received her Ph.D. degree from Beijing University of Posts and Telecommunications in 2012. She is a Professor with the School of Computer and Communication, Lanzhou University of Technology. Her research interests include machine learning fundamentals with a focus on applications in image and video understanding.



Shaoying Xue received her B.E. degree in Computer Science and Technology from Tianshui Normal University in 2022 and M.Eng. degree in Computer Technology at Lanzhou University of Technology in 2025. Her research interests include machine learning, computer vision, and few-shot learning.



Jiyang Xie received the B.E. degree in information engineering and the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications (BUPT), China, in 2017 and 2022, respectively. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in image processing, data mining, and deep learning.



Xiaochen Yang received the Ph.D. degree in statistical science from University College London, London, in 2020. She is currently a Senior Lecturer with the School of Mathematics and Statistics, University of Glasgow, Glasgow, U.K. Her research interests include statistical classification, machine learning, and medical image analysis. She is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology.

Zhanyu Ma is currently a Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2019. He received the Ph.D. degree in electrical engineering from KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he was a Postdoctoral Research Fellow with the School of Electrical Engineering, KTH. He has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, from 2014 to 2019. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing. He is a Senior Member of IEEE.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a Professor of Statistical Pattern Recognition in the Department of Statistical Science, University College London. His research interests include statistical pattern recognition, machine learning, and computer vision. He received the Best Associate Editor Award of 2021 from the IEEE Transactions on Circuits and Systems for Video Technology, and the Outstanding Associate Editor Award of 2022 from the IEEE Transactions on Neural Networks and Learning Systems.