City Research Online

City, University of London Institutional Repository

Citation: Li, X., Wang, X., Zhu, R., Ma, Z., Cao, J. & Xue, J-H. (2025). Selectively augmented attention network for few-shot image classification. IEEE Transactions on Circuits and Systems for Video Technology, 35(2), pp. 1180-1192. doi: 10.1109/tcsvt.2024.3480279

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/33977/

Link to published version: https://doi.org/10.1109/tcsvt.2024.3480279

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Selectively augmented attention network for few-shot image classification

Xiaoxu Li, Xiangyang Wang, Rui Zhu, Zhanyu Ma, *Senior Member, IEEE*, Jie Cao, Jing-Hao Xue, *Senior Member, IEEE*

***Abstract*—Few-shot image classification is a challenging task that aims to learn from a limited number of labelled training images a classification model that can be generalised to unseen classes. Two strategies are usually taken to improve the classification performances of few-shot image classifiers: either applying data augmentation to enlarge the sample size of the training set and reduce overfitting, or involving attention mechanisms to highlight discriminative spatial regions or channels. However, naively applying them to few-shot classifiers directly and separately may lead to undesirable results; for example, some augmented images may focus majorly on the background rather than the object, which brings additional noises to the training process. In this paper, we propose a unified framework, the selectively augmented attention (SAA) network, that carefully integrates the best of the two approaches in an end-to-end fashion via a selective best match module to select the most representative images from the augmented training set. The selected images tend to concentrate on the objects with less irrelevant background, which can assist the subsequent calculation of attentions by alleviating the interference from background. Moreover, we design a joint attention module to jointly learn both the spatial and channel-wise attentions. Experimental results on four benchmark datasets showcase the superior classification performance of the proposed SAA network compared with the state-of-the-arts.**

***Index Terms*—Few-shot image classification, Data augmentation, Attention mechanism, Metric-based methods**

I. Introduction

DEEP learning has achieved significant advancements in the field of computer vision, reaching human-level accuracy when the deep model is trained on a substantial amount of labelled data. However, the cost of labelling and training becomes prohibitive for large-scale datasets. Therefore, it is worthwhile to investigate few-shot learning that aims to maintain efficient and accurate classification when there are only few labelled images for training [1].

Metric-based methods that make class membership assignments based on proper metric functions are effective solutions for few-shot image classification [2], [3]. The metric function can be either pre-defined by domain knowledge [4] or properly learnt via training data [5]. ProtoNet [4] is a classical metric-based method, which classifies the test image by calculating its Euclidean distances to the class prototypes, i.e. the simple averages of each class in the support set. DN4 [6] adopts a novel image-to-class metric based on local descriptors, using the local features of samples to learn feature metrics. BSNet [7] uses a dual similarity network as the metric, utilising a combination of two different metrics to learn fewer but more discriminative spatial regions to assist classification. NDPNet [8] uses a feature re-abstraction embedding network that projects local features into the similarity metric learning network, which aims to learn discriminative projection factors. It adopts the Euclidean distance to measure the dissimilarity between constructed features and original features.

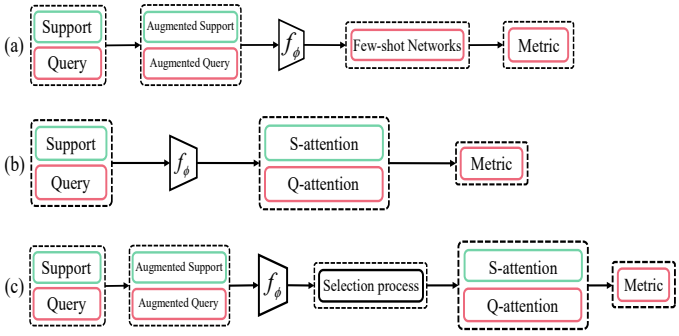
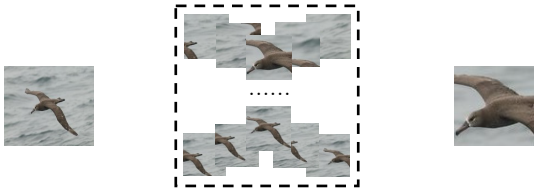


Fig. 1. Three strategies to improve the performance of few-shot image classification. (a) Data augmentation to reduce the overfitting arising from the limited amount of training data. (b) Attention mechanism to assign higher weights to more discriminative features or channels. S-attention and Q-attention denote the attentions for support and query sets, respectively. The attentions can be in either spatial or channel-wise dimensions or both. (c) A unified framework proposed in this paper to carefully integrate both data augmentation and attention mechanism, via a key step of selecting the most representative augmented images to obtain attentions in an end-to-end fashion.



(a) The oiginal image (b) The augmented image set by random cropping (c) The most representative image

Fig. 2. An example of the most representative image selected from the augmented image set.

Corresponding author: Rui Zhu (email: rui.zhu@city.ac.uk)

X. Li, X. Wang, J. Cao are with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China (email: lixiaoxu@lut.etu.cn, 2395792430@qq.com, caoj@lut.edu.cn).

R. Zhu is with the Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London EC1Y 8TZ, UK.

Z. Ma are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (email: mazhanyu@bupt.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (email: jinghao.xue@ucl.ac.uk).

Two strategies are usually implemented to improve the classification performances of few-shot classifiers. First, since there are only few training images for each class, few-shot methods can suffer from overfitting easily. Thus, data augmentation techniques aiming at increasing the sample size and the diversity of the training set are effective solutions to reduce overfitting, as illustrated in Fig. 1(a). Commonly adopted data augmentation methods include image rotation, image flipping, random cropping, noise injection and image mixing [9]. Generative methods such as GAN [10] are also powerful to generate new images. Data augmentation techniques have been well explored in the few-shot setting. For example, Kumar et al. [11] investigate six feature-space data augmentation methods and demonstrate that they can improve classification performance in few-shot setting. Wang et al. [12] propose a novel feature-space augmentation method based on rectified normal distribution by considering the relationship between base and novel classes. However, it is important to note that data augmentation methods can introduce noises and may even carry the risk of overfitting due to inappropriate augmentation, which can impact the final classification performance in a negative way [9]. For example, when creating new images by random cropping, some cropped patches may concentrate mostly on the background rather than the object, which can bring additional noises to the training process. We present one example in Fig. 2, where some random cropped patches focus on the background of water rather than the object of bird.

Second, instead of treating all features equally, identifying and rewarding those features with high discriminative power can boost the accuracy of few-shot image classification. This could be achieved via additional attention mechanisms to obtain attentions that are usually either in the spatial dimension to weigh discriminative spatial features or in the channel dimension to highlight discriminative channels, as illustrated in Fig. 1(b). For instance, Song et al. [13] propose a fusion spatial attention method in both the image space and the embedded space. During the fusion process, different weights are assigned to different positions in the two spaces, and the information is fused and integrated into existing few-shot learning methods. Yan et al. [14] propose a meta learning method by considering spatial attention to locate relevant object regions. They also involve a special task-wise attention mechanism to select similar training data for classification. Some works also design the channel-wise attention mechanisms. For example, Hu et al. [15] propose a squeeze and extraction (SE) block to clarify the interdependence between channels. Lee et al. [16] introduce the task descrepancy maximisation (TDM) module to learn task-wise channel weights. Moreover, in other computer vision areas, such as person re-identification, spatial and channel attentions are learnt jointly [17].

However, to the best of our knowledge, there are no existing methods that carefully combine the two strategies together to further enhance the performance of few-shot image classification. In this paper, we propose a unified framework, the selectively augmented attention (SAA) network, that integrates the strengths of both approaches in an end-to-end fashion. Our motivation is presented in Fig. 1(c). In this framework, after extracting support and query features from the augmented support and query sets, we propose a key step to select the most representative feature from each set and calculate attentions based only on the selected ones. In this way, we aim to alleviate the potential noises brought by the augmented samples and assist the subsequent attention calculation with less contamination in the training data.

In this paper, for illustration purposes, we adopt the random cropping technique to augment the support and query sets, but other augmentation methods or combinations of different augmentation methods can be utilised as well. With random cropping, we obtain patches in random sizes from the original image to enlarge the sample size of the training set. For the selection process, we propose a novel selective best match (SBM) module that can select the features focusing mostly on the objects and eliminate potential background noises. Here each feature corresponds to a specific image passing through the embedding module, and feature selection aims to find the feature corresponding to the most representative image in the augmented set of an original image. Specifically, we obtain the pairwise cosine similarities between all augmented support and query features. From the augmented feature set of each support image, we select the one with the highest mean similarity to all query features as the representative, while for each query image, we select the one that is mostly similar to all support features. In this way, images concentrating on the objects, e.g. with the object in the centre of the image, are usually selected, while those background patches are ignored in the subsequent attention calculation because they tend to be less similar to other images. Fig. 2(c) shows the the most representative image selected by the proposed SBM module, which focuses mostly on the bird's head and beak that are usually the areas helpful to identify the bird species. Moreover, we design a joint attention (JA) module to encourage the model to jointly learn discriminative spatial and channel-wise features. Since only the selected clean features focusing on objects are utilised in the attention mechanism, the obtained attention maps can well capture the spatial regions and channels to discriminatively describe the objects. In such a manner, the SBM and JA modules work together to guide the model to pay more attention to the most crucial features that can largely distinguish between different classes. Experimental results on four benchmark datasets showcase the superior classification performance of our SAA network compared with the state-of-the-arts few-shot learning methods.

To sum up, the contributions of this paper are four-fold:

1) We propose the novel selectively augmented attention (SAA) network for few-shot image classification, which carefully integrates data augmentation with joint spatial and channel-wise attentions to boost the classification performance.
2) To reduce the potential noises brought by the augmented images, we introduce the selective best match (SBM) module that can find the most representative images from the augmented support and query sets. The selected images tend to focus more on the objects rather the irrelevant background.
3) We design the joint attention (JA) module to apply the attention mechanism on both spatial and channel

dimensions based only on the selected features from SBM. In this way, we obtain the adjusted features that can focus mainly on the discriminative regions and channels of objects to assist classification.

4) We conduct extensive experiments on four benchmark datasets to validate the effectiveness of SAA networks. The experimental results demonstrate significant improvement on classification accuracy for few-shot image classification.

The rest of the paper is organised as follows. In section II, we discuss the literature closely related to our work. The technical details of the proposed SAA network are introduced in section III. In section IV, we present extensive experimental results and ablation study to validate the effectiveness of SAA network. Finally, we draw conclusions in section V.

II. Related Work

A. Metric-based few-shot image classification

Metric-based few-shot image classification assigns a test image to its most similar class by a metric function, which can be pre-defined or learnt from the network. Classic metric-based methods include the matching networks (MatchNet) based on the cosine similarities between images [18], the prototypical networks (ProtoNet) based on the Euclidean distances between a query image and class prototypes [4], and the relation network (RelationNet) which learns the metric function via the relation module [19]. The metric function is also learnt in various ways recently. For example, the deep nearest nieghbour neural network (DN4) includes an image-to-class module to compute the consine similarity between a query image and its nearest neighbours in each class [6]. The bi-similarity network (BSNet) calculates two different metrics to capture diverse and discriminative characristics between classes [7].

In this paper, we adopt ProtoNet as the underlying metric-based method to demonstrate the effectiveness of our SAA networks, which carefully integrate both data augmentation and attention mechanism via the proposed novel SBM and JA modules, respectively.

B. Data augmentation for few-shot image classification

Few-shot image classification can easily suffer from overfitting due to the limited number of training samples. Data augmentation is a straightforward pre-processing strategy to increase the training sample size, usually via various transformations or modifications of the existing training data. Kumar et al. [11] demonstrate that the performance for few-shot image classification can be enhanced by data augmentation via a study of six feature-space data augmentation methods. Chu et al. [20] propose a deep reinforcement learning method that can be treated as a learnt data augmentation step to search for different sequences of patches of an image reflecting human glimpse trajectories to recognise an object. Wang et al. [12] adopt the rectified normal distribution for feature-space augmentation that considers the relationship between base and novel classes. Hu et al. [21] fuse the class-irrelevant and class-relevant features to obtain the augmented features. However, the augmentation step can result in noisy training data, which would have harmful impact on the training process.

Different from previous studies to apply existing data augmentation methods to improve classification performance, we aim to properly utilise the augmented images by selecting the most representative ones to reduce the background noises and propose the novel SBM module to achieve this.

C. Attention mechanisms for few-shot image classification

Attention mechanisms are widely applied to address the learning and generalisation challenges when there are limited training data available for few-shot image classification. They can help the model focus on discriminative inter-class information and enhance the generalisation ability to new tasks. Attentions are usually learnt from either the spatial dimension to extract the most informative spatial regions or the channel dimension to identify which channels are more discriminative to assist classification. To learn spatial-wise attention, Song et al. [13] propose to assign different weights to different positions in both the original feature space and the embedding spaces. Yan et al. [14] propose to learn spatial attention and locate relevant object regions via a meta learning approach. Xu et al. [22] propose a dual attention network containing two parallel branches to learn hard attentions that can exploit correlation between fine object parts and soft attentions to learn global aggregated features. To learn channel-wise attention, Lee et al. [16] design the task discrepancy maximisation (TDM) module that can assign task-wise channel weights to capture the most discriminative information. There are studies to fuse both types of attentions in other computer vision areas, such as person re-identification [17].

In our work, we jointly learn the spatial and channel-wise attentions in the novel JA module, based on the selected features from the SBM module.

III. Methodology

In this section, we provide the technical details of the proposed SAA network. The overall structure of the SAA network is presented in section III-B. In sections III-C, III-D and III-E, we introduce the SBM module to select the most representative features, the JA module to learn spatial and channel-wise attentions and the metric module for membership assignment, respectively.

A. Problem formulation

Following literature [4], [6], [8], [23], [24], we employ the classic episodic training strategy for few-shot image classification, where each episode learns from an N-way K-shot task, i.e. the models are trained based on N classes, each with K labelled images. This trained model aims to learn general knowledge from the tasks sampled from the training set and can be easily adapted to the tasks in the test set.

To be more specific, given a dataset $\mathcal{D}$ with a label set $\mathcal{L}$, we randomly split it to three mutually exclusive subsets, the training set $\mathcal{D}_{\text{train}}$, the validation set $\mathcal{D}_{\text{val}}$ and the test set $\mathcal{D}_{\text{test}}$. Note that the label sets of the three subsets form a partition

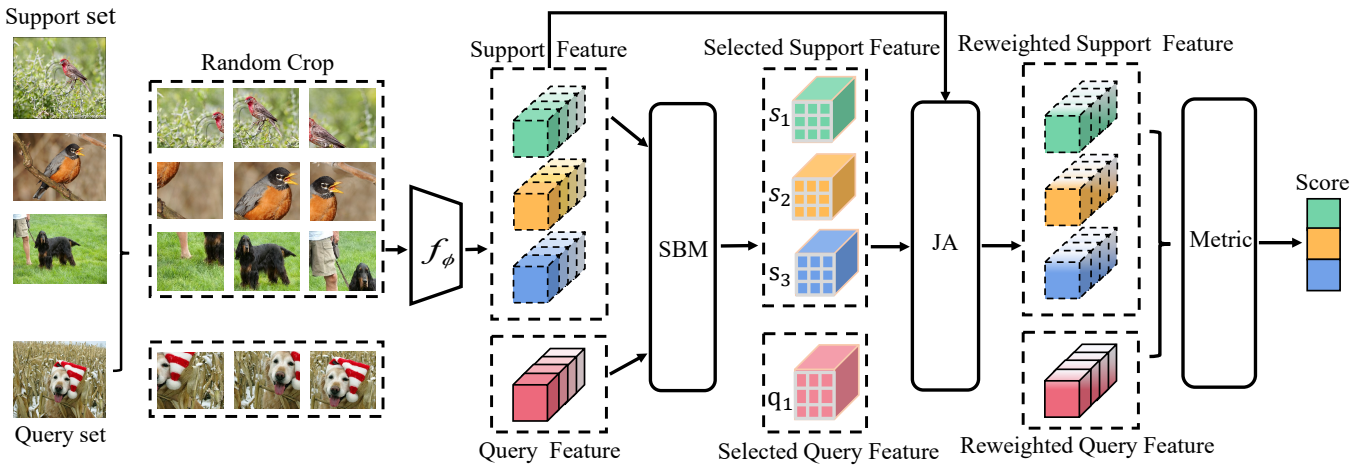


Fig. 3. The overall structure of the proposed SAA network in the three-way one-shot setting. It consists of the random cropping procedure, the embedding module f_ϕ, the SBM module, the JA module and the metric module. The SBM module takes the embedded features of the original support and query sets and their augmented versions as input, calculates the cosine similarities between the support features and query features and selects the most representative features with the highest similarities. Subsequently, the selected features from the SBM module are fed into the JA module to obtain the spatial and channel-wise attentions. We then reweight the original support and query features by the obtained attentions. The reweighted features are utilised in the metric module to calculate the metric scores for classification. The details of the SBM and JA modules are depicted in Figures 4 and 5, respectively.

of $\mathcal{L}$; that is, $\mathcal{L}_{\text{train}}$, $\mathcal{L}_{\text{val}}$ and $\mathcal{L}_{\text{test}}$ are mutually exclusive and their union is $\mathcal{L}$. To form a task for each episode, we randomly select N classes from $\mathcal{L}_{\text{train}}$, and randomly select M images for each of these N classes from $\mathcal{D}_{\text{train}}$. Then, for each selected class, the M images are randomly split to two subsets with K images to form the support set $\mathcal{S}$ and $M-K$ images to form the query set $\mathcal{Q}$. Following the same strategy, we can define tasks for the validation and test sets.

B. The overall structure of the SAA network

To exploit and illustrate the advantages of both data augmentation and attention mechanism in one unified framework, we propose the SAA network with the random cropping to augment the dataset and the joint attention (JA) module to incorporate both spatial and channel-wise attentions. More importantly, to carefully integrate the two data augmentation and attention mechanism, we propose the selective best match (SBM) module to select the most representative support and query features from the augmented features to eliminate potential noises and to facilitate the calculation of attentions that concentrate more on the objects.

The overall structure of SAA networks is depicted in Fig. 3, which consists of the random cropping for data augmentation, the embedding module f_ϕ to extract features, the SBM module to select representative features, the JA module to obtain discriminative spatial features and channels, and finally the metric module to calculate the similarity scores between the reweighted support and query features.

In this paper, we adopt the prototypical network (ProtoNet) [4] as the base metric learning method. To be more specific, we calculate the simple average of each class in the support set as the class prototype, and obtain the Euclidean distances between the query feature and the prototypes. The query image is then assigned to the class with the shortest distance (or say the highest similarity).

Instead of naively taking all augmented features to train the network, we propose to select the most representative ones to reinforce the learning process, especially the attention mechanism, because data augmentation can sometimes provide patches focusing mainly on the nuisance background and this contamination of training data can lead to deficient classification accuracy. To resolve this problem, we propose to carefully integrate the augmentation procedure via the SBM module. By comparing the similarities between the support and query features, the SBM module aims to select the best matched ones with strong focuses on the objects and thus minimise the negative impact brought by the noisy background.

C. The selective best match (SBM) module

To enlarge the sample size of the training set and reduce overfitting, we adopt the random cropping procedure [9] for both support and query sets. That is, for each support or query image, we obtain the corresponding augmented image set with size of R, including the original image. That is, for the support set, the sample size becomes $R \times N \times K$ after generating cropped patches, while for the query set, it becomes $R \times N \times (M-K)$. These samples are then passed through the embedding module f_ϕ, resulting in support features $\mathbf{Z}_i^S \in \mathbb{R}^{R\times c\times h\times w}$ $(i = 1, 2, \ldots, N \times K)$ and $\mathbf{Z}^Q \in \mathbb{R}^{R\times c\times h\times w}$, where c, h and w represent the number of channels, the height and the width of the features, respectively. Note that we provide the example of classifying one query image and $\mathbf{Z}^Q$ represents the augmented image features of one query image rather than the whole query set.

The SBM module takes the embedded features $\mathbf{Z}_i^S$ and $\mathbf{Z}^Q$ as input and aims to select the most representative feature from the augmented set of R features for each image. Specifically, we calculate the pairwise cosine similarities between all

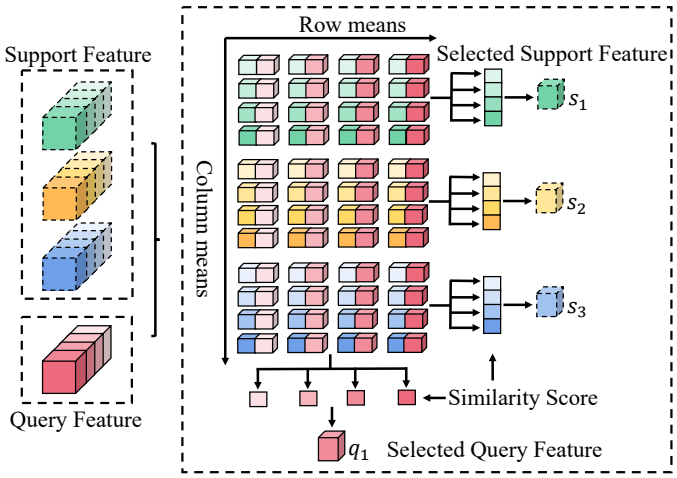


Fig. 4. An illustration of the selective best match (SBM) module in the three-way one-shot setting. The green, yellow and blue tensors on the left-hand side are the augmented feature sets of three support images, respectively, while the red tensor is the augmented feature set of the query image. The small cubes with different shades in each tensor denote the augmented features. For each pair of support and query features, we calculate their cosine similarities and obtain the similarity matrix $\mathbf{S}$. Taking the row means of $\mathbf{S}$, we can find the most representative support features: for the augmented feature set of each support image, we select the feature with the highest mean similarity with all query features. Similarly, we take the column means of $\mathbf{S}$ and for each query image, we select the one with the highest mean similarity to all images in the support set.

support and query features and obtain the cosine similarities

$$s_{ijk} = \frac{(\mathbf{z}_{ij}^S)^T \mathbf{z}_k^Q}{\max(\|\mathbf{z}_{ij}^S\|_2 \cdot \|\mathbf{z}_k^Q\|_2, \epsilon)}, \tag{1}$$

where $\mathbf{z}_{ij}^S$ is the flattened vector of the jth augmented feature of the ith support sample with $i = 1, 2, \ldots, N \times K$ and $j = 1, 2, \ldots, R$; $\mathbf{z}_k^Q$ is the flattened vector of the kth augmented query feature with $k = 1, 2, \ldots, R$; and ϵ is a small nonnegative value to prevent the numerical problem when the denominator is close to zero. These similarities form a three-way tensor of dimensions $(N \times K) \times R \times R$ and we flatten it to obtain the similarity matrix $\mathbf{S} \in \mathbb{R}^{(N \times K \times R) \times R}$, as shown on the right-hand-side of Fig. 4. The rows of $\mathbf{S}$ are the similarities between one support feature and all query features while the columns are those between one query feature and all support features.

The one with the largest similarity to all query features is the selected as the most representative feature for each support image. To achieve this, we calculate the row means of $\mathbf{S}$ and obtain one mean similarity score for each support feature, and for the augmented feature set of each image, we select the one with the largest mean similarity score. Thus, the selected support feature represents the one that is highly correlated to all query features. Similarly, we calculate the column means of $\mathbf{S}$ and obtain one mean similarity score for each query feature, which measures the overall similarity between each query feature and all support features. Again, the query feature with the largest mean score is selected as the representative of all query features.

The output of the SBM module is the selected support features $\hat{\mathbf{Z}}_i^S \in \mathbb{R}^{R \times c \times h \times w}$ and the query features $\hat{\mathbf{Z}}^Q \in \mathbb{R}^{R \times c \times h \times w}$. Note that we repeat the selected features R times to facilitate calculations in latter modules.

D. The joint attention (JA) module

The structure of the joint attention (JA) module is illustrated in Fig. 5, which consists of the following two parts. First, the selected features from the SBM module are utilised to calculate the attentions in both spatial and channel dimensions. Second, the augmented features are then adjusted by the attentions obtained from the first part to assign discriminative features and channels higher weights.

To be more specific, for each class, in the channel dimension, the selected support and query features are passed through an adaptive average pooling layer, resulting in an $N \times c \times 1 \times 1$ tensor at each pixel position. This tensor is flattened into a one-dimensional tensor and sequentially passed through a linear layer, a normalization layer and an activation function layer. Finally, an upsampling function is adopted to restore the shape of the tensor to its initial state. This channel attention mechanism is illustrated in the left branch in Fig. 6. In the spatial direction, a 1×1 convolutional layer is firstly used to reduce the number of channels, which can reduce the number of parameters of the network and the computational complexity. Subsequently, a 3×3 convolutional layer is applied to the feature maps in the $h \times w$ direction, increasing the focus on local regions of the feature maps. Both the 1×1 and 3×3 convolutional layers are followed by a normalization layer and an activation function layer to accelerate network training, reduce model parameters and prevent overfitting. Lastly, the output from the last 1×1 convolutional layer is mapped to $[0, 1]$ using a sigmoid function, as shown in the right branch in Fig. 6. The resulting attentions of the two dimensions are summed together to obtain the final attentions, $\mathbf{W}^S$ and $\mathbf{W}^Q$, for the support and query sets, respectively:

$$\{\mathbf{W}^S, \mathbf{W}^Q\} = g_A(\{\hat{\mathbf{Z}}_i^S\}_{i=1}^{N \times K}, \hat{\mathbf{Z}}^Q), \tag{2}$$

where $g_A(\cdot)$ denotes the attention mechanism in Fig. 6.

Before utilising $\mathbf{W}^S$ and $\mathbf{W}^Q$ to adjust the augmented features, we propose an additional step to further alleviate the impact of the nuisance background. For background positions, their element-wise products with $\mathbf{W}^S$ or $\mathbf{W}^Q$ are close to zeros as they are suppressed by attentions, and they can be eliminated to facilitate classification. To achieve this, we set up a mask matrix with initial values of ones. After passing the products of $\mathbf{W}^S \otimes \exp(\mathbf{Z}_i^S)$ and $\mathbf{W}^Q \otimes \exp(\mathbf{Z}^Q)$ through the softmax function, we record the position information of all 0 elements in the output to identify background positions, replace all corresponding positions in the mask matrix with 0, and then multiply the obtained mask matrix with $\mathbf{Z}_i^S$ or $\mathbf{Z}^Q$ to filter out background.

Finally, the reweighted support and query features are obtained as follows:

$$\mathbf{T}_i^S = \text{softmax}(\frac{\exp(\mathbf{Z}_i^S) \otimes \mathbf{W}^S}{\sqrt{d}}) + \mathbf{M}_i^S \otimes \mathbf{Z}_i^S, \tag{3}$$

$$\mathbf{T}^Q = \text{softmax}(\frac{\exp(\mathbf{Z}^Q) \otimes \mathbf{W}^Q}{\sqrt{d}}) + \mathbf{M}^Q \otimes \mathbf{Z}^Q, \tag{4}$$

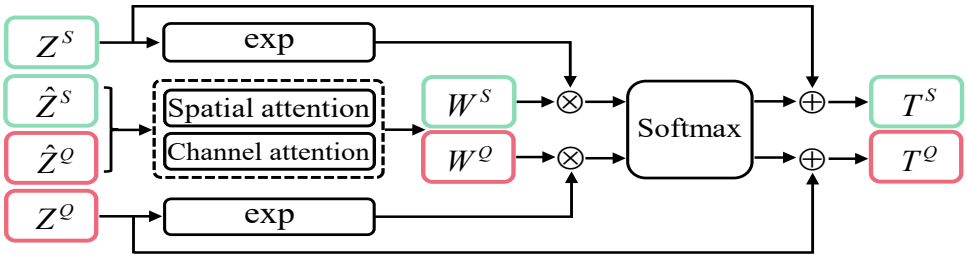


Fig. 5. An illustration of the joint attention (JA) module. $\mathbf{Z}^S$ and $\mathbf{Z}^Q$ are the augmented feature sets of support and query images, respectively. $\hat{\mathbf{Z}}^S$ and $\hat{\mathbf{Z}}^Q$ are the selected representative features by the SBM module, and $\mathbf{W}^S$ and $\mathbf{W}^Q$ are the attentions for the support and query sets, respectively. Only the selected features $\hat{\mathbf{Z}}^S$ and $\hat{\mathbf{Z}}^Q$ are utilised to obtain attentions $\mathbf{W}^S$ and $\mathbf{W}^Q$, which are then used to reweight the original features $\mathbf{Z}^S$ and $\mathbf{Z}^Q$. The details of the spatial and channel attention mechanisms are depicted in Figure 6.

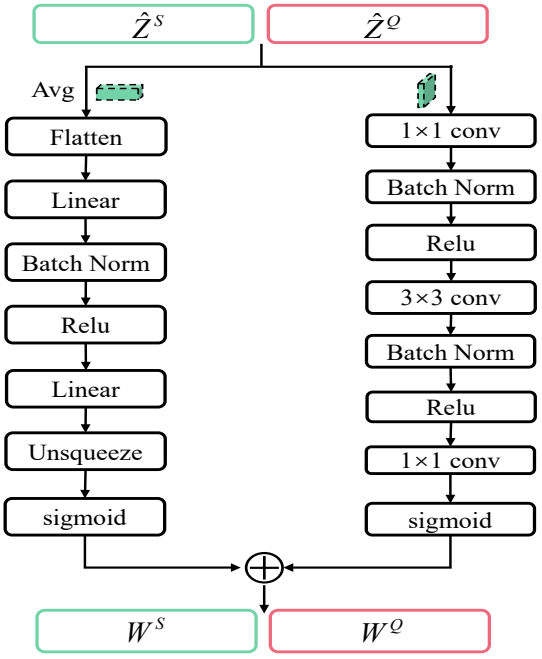


Fig. 6. An illustration of the spatial attention (right branch) and channel attention (left branch) mechanisms in the JA module.

where $\mathbf{M}_i^S \in \mathbb{R}^{R\times c\times h\times w}$ and $\mathbf{M}^Q \in \mathbb{R}^{R\times c\times h\times w}$ are the masks for the support and query features, respectivley, $\mathbf{T}^Q \in \mathbb{R}^{R\times c\times h\times w}$ are the reweighted support and query features, respectively, and d is the scaling parameter. In this way, $\mathbf{T}_i^S$ and $\mathbf{T}^Q$ carry the information of the discriminative spatial regions and channels for better classification, with background information filtered out through the SBM module and the mask operation.

E. The metric module

In the metric module, we assess the differences between the prototype of the original augmented features and the reweighted query features, as well as those between the reweighted prototype and the original query features.

Specifically, we calculate

$$d_{\mathbf{T}^Q\rightarrow\boldsymbol{\mathcal{P}}_n^S} = \|\mathbf{T}^Q - \boldsymbol{\mathcal{P}}_n^S\|_2^2, \tag{5}$$

and

$$d_{\mathbf{Z}^Q\rightarrow\boldsymbol{\mathcal{Q}}_n^S} = \|\boldsymbol{\mathcal{Q}}_n^S - \mathbf{Z}^Q\|_2^2, \tag{6}$$

where $\mathbf{Z}^Q$ and $\mathbf{T}^Q$ are the original and reweighted query features, respectively, while $\boldsymbol{\mathcal{P}}_n^S$ and $\boldsymbol{\mathcal{Q}}_n^S$ are the original and reweighted prototypes of the nth class, respectively. Finally, the total distance between the query and the support set of the nth class is obtained by taking the weighted sum of the two differences:

$$d_n^Q = \lambda_1 d_{\mathbf{T}^Q\rightarrow\boldsymbol{\mathcal{P}}_n^S} + \lambda_2 d_{\mathbf{Z}^Q\rightarrow\boldsymbol{\mathcal{Q}}_n^S}, \tag{7}$$

where λ_1 and λ_2 are learnable weights, and both of them are initialised as 0.5.

To train the network, we follow ProtoNet and minimise the cross-entropy loss:

$$L = -\sum_{l=1}^{N(M-K)} \log\left(P(\hat{y}_l = y_l \mid \mathbf{x}_l^Q)\right), \tag{8}$$

with

$$P\left(\hat{y} = n \mid \mathbf{x}^Q\right) = \frac{e^{-\delta d_n^Q}}{\sum_{n'\in[1,N]} e^{-\delta d_{n'}^Q}}, \tag{9}$$

where $\mathbf{x}_l^Q$ is the lth query image with label y_l and δ parameterises the softmax function.

TABLE I
THE DETAILS OF THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS, INCLUDING THE SAMPLE SIZE, THE NUMBER OF CLASSES AND THE TRAINING\VALIDATION\TEST SPLIT.

	Sample size	Number of classes	Training\valid.\test split
CUB	11,788	200	100\50\50
Dogs	20,580	120	60\30\30
Aircraft	10,000	100	50\25\25
Flower	8,189	102	51\25\26

IV. EXPERIMENTS

A. Datasets

In the experiments, we choose four benchmark datasets, the CUB, Aircraft, Dogs and Flower datasets. The CUB dataset [33] consists of 200 categories with a total of 11,788 images. The Aircraft dataset [34] contains 10,000 aircraft images with four-level hierarchical notations: model, variant, family and manufacturer. The Dogs dataset [35] includes 28,580 annotated images of 120 dog breeds from around the

TABLE II
FIVE-WAY FEW-SHOT CLASSIFICATION ACCURACIES ON THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE CONV-4 BACKBONE STRUCTURE. THE BEST RESULTS ARE IN BOLD.

Model	5-Way Accuracy (%)							
	CUB		*Dogs*		*Aircraft*		*Flowers*	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [4]	61.64 ± 0.23	70.23 ± 0.15	37.59 ± 0.79	48.19 ± 0.91	50.90 ± 0.22	71.65 ± 0.15	59.23 ± 0.23	79.97 ± 0.16
MatchingNet [18]	60.06 ± 0.88	74.57 ± 0.73	46.10 ± 0.86	59.79 ± 0.72	58.47 ± 0.28	70.90 ± 0.11	63.89 ± 0.90	77.46 ± 0.59
RelationNet [19]	63.94 ± 0.92	77.87 ± 0.64	47.35 ± 0.88	66.20 ± 0.74	62.04 ± 0.91	82.48 ± 0.49	65.44 ± 0.95	83.45 ± 0.52
DN4 [6]	56.45 ± 0.89	80.41 ± 0.58	39.08 ± 0.76	69.81 ± 0.69	64.41 ± 0.91	83.48 ± 0.49	65.15 ± 0.94	78.86 ± 0.56
Baseline++ [25]	62.36 ± 0.84	79.08 ± 0.61	44.49 ± 0.70	64.48 ± 0.66	52.38 ± 0.83	70.62 ± 0.60	65.54 ± 0.84	80.63 ± 0.58
DeepEMD [26]	64.08 ± 0.50	80.55 ± 0.71	46.73 ± 0.49	65.74 ± 0.63	66.37 ± 0.23	80.76 ± 0.14	63.41 ± 0.21	76.93 ± 0.17
BSNet(D&C) [7]	62.84 ± 0.95	81.39 ± 0.56	43.42 ± 0.86	71.49 ± 0.68	56.61 ± 1.09	70.80 ± 0.81	66.60 ± 1.04	80.42 ± 0.75
SAML [27]	65.35 ± 0.65	78.47 ± 0.41	45.46 ± 0.36	59.65 ± 0.51	-	-	-	-
MixtFSL [28]	53.61 ± 0.88	73.24 ± 0.75	43.96 ± 0.77	64.43 ± 0.68	44.89 ± 0.75	62.81 ± 0.73	67.01 ± 0.90	**84.50 ± 0.62**
LRPABN [29]	46.81 ± 0.73	71.82 ± 0.99	45.72 ± 0.75	60.94 ± 0.66	-	-	-	-
ProtoNet [4]+TDM [16]	65.90 ± 0.49	76.96 ± 0.50	44.73 ± 0.48	53.24 ± 0.51	53.36± 0.90	73.07 ± 0.33	64.16 ± 0.52	81.72 ± 0.53
Ours	**67.26 ± 0.51**	**81.49 ± 0.34**	**54.13 ± 0.63**	**71.67 ± 0.37**	**71.11 ± 0.50**	**84.39 ± 0.29**	**67.14 ± 0.52**	84.28 ± 0.35

TABLE III
FIVE-WAY FEW-SHOT CLASSIFICATION ACCURACIES ON THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE RESNET-12 BACKBONE STRUCTURE.

Model	5-Way Accuracy (%)							
	CUB		Dogs		Aircraft		Flowers	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [4]	68.48 ± 0.50	76.48 ± 0.28	62.90 ± 0.51	77.95 ± 0.34	59.48 ± 0.52	77.87 ± 0.33	68.53 ± 0.52	84.30 ± 0.35
MatchingNet [18]	71.87 ± 0.85	85.08 ± 0.57	65.12 ± 0.35	80.50 ± 0.42	82.20 ± 0.80	88.99 ± 0.50	74.57 ± 0.69	87.61 ± 0.55
RelationNet [19]	70.09 ± 0.46	84.38 ± 0.16	59.27 ± 0.79	79.10 ± 0.37	74.20 ± 1.04	86.62 ± 0.55	69.51 ± 1.01	86.84 ± 0.56
BSNet(P&C) [7]	69.61 ± 0.92	81.28 ± 0.64	63.58 ± 0.96	79.10 ± 0.77	63.02 ± 0.93	79.00 ± 0.75	70.21 ± 1.06	84.78 ± 0.82
DeepEMD [26]	71.11 ± 0.31	86.30 ± 0.19	67.59 ± 0.30	83.13 ± 0.20	73.30 ± 0.29	88.37 ± 0.17	70.00 ± 0.35	83.63 ± 0.26
MixtFSL [28]	67.86 ± 0.94	82.18 ± 0.66	67.26 ± 0.90	82.05 ± 0.56	60.55 ± 0.86	77.57 ± 0.69	72.60 ± 0.91	86.52 ± 0.65
BlockMix [30]	75.16 ± 0.69	87.62 ± 0.35	67.87 ± 0.33	82.26 ± 0.45	-	-	-	-
QSFormer [31]	75.26 ± 0.17	86.42 ± 0.19	68.87 ± 0.72	83.56 ± 0.45	76.45 ± 0.56	88.97 ± 0.55	**75.24 ± 0.25**	87.81 ± 0.60
ProtoNet [4]+TDM [16]	71.32 ± 0.37	80.12 ± 0.34	64.73 ± 0.48	78.19 ± 0.36	61.39 ± 0.76	79.09 ± 0.56	70.24 ± 0.64	85.22 ± 0.30
Ours	**75.57 ± 0.48**	**88.03 ± 0.29**	**70.32 ± 0.50**	**84.61 ± 0.32**	**85.70 ± 0.40**	**91.80 ± 0.19**	74.22 ± 0.49	**90.19 ± 0.28**

world. The Flowers dataset [36] consists of images of flowers from 102 different categories with a total of 8,189 images.

We randomly split each dataset to a training set, a validation set and a test set in the ratio of $2:1:1$. The sample size of each dataset, the number of classes and the split of classes for the three subsets are shown in Table I. All images in the four datasets are resized to 84×84.

B. Implementation details

All of our experiments are performed using PyTorch on NVIDIA 3090Ti GPUs. We conduct experiments on two backbone architectures: Conv-4 and ResNet-12. The architectures of Conv-4 and ResNet-12 are the same as those used in previous works [37], [38]. In addition, the proposed attention module in this paper is shared during the training process. For both backbone architectures, we train the models under 10-way 1-shot and 5-shot settings and conduct tests under 5-way 1-shot and 5-shot settings.

During the training phase, the Adam optimizer is used with a learning rate of 0.003, without decay or scheduling. The 1-shot models are trained for 300 epochs, with 200 tasks per epoch, resulting in a total of 60,000 tasks. The 5-shot models are trained for 200 epochs, with 200 tasks per epoch, resulting in a total of 40,000 tasks. During the test phase, we evaluate the models by testing 15 query samples per class in each episode, and report the average accuracy with a 95% confidence interval.

For random cropping to augment the datasets, we generate 20 randomly cropped images for each image, with a randomly chosen ratio of cropping size from 0.3, 0.5 and 0.7. Regardless of the sizes of the cropped images, they are converted to 84×84 as input to the embedding module.

C. Comparison with the state-of-the-arts

To validate the effectiveness of the proposed SAA network, we compare it with several classic methods for few-shot image classification, such as ProtoNet [4], Baseline++ [25], MatchingNet [18], and RelationNet [19], as well as some state-of-the-arts methods, including SAML [27], LRPABN [29], MTL [39], DN4 [6], BSNet [7], DeepEMD [26], MixFSL [28], TDM [16], BlockMix [30], QSFormer [31]. The classification performances of these methods on the four benchmark datasets can be found in Tables II and III.

For the Conv-4 backbone, our proposed SAA outperforms other methods on all four datasets, except for the 5-shot scenario on the Flowers dataset where SAA is slightly worse than MixFSL. For the ResNet-12 backbone, SAA can beat all methods on all datasets. The results demonstrate that

1

TABLE IV

COMPARING THE SELECTIVELY AUGMENTATION (SA) METHOD WITH TWO CLASSIC DATA AUGMENTATION METHODS, CUTOUT AND RANDOM FLIP, ON THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE RESNET-12 BACKBONE STRUCTURE. THE BEST RESULTS ARE IN BOLD.

Model	5-Way Accuracy (%)							
	CUB		*Dogs*		*Aircraft*		*Flowers*	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
SENet [15]	69.98 ± 0.50	81.53 ± 0.29	63.79 ± 0.50	79.21 ± 0.29	62.86 ± 0.51	79.69 ± 0.30	70.55 ± 0.50	84.25 ± 0.29
SENet+Cutout	63.24 ± 0.50	80.16 ± 0.32	61.28 ± 0.49	76.20 ± 0.34	66.72 ± 0.50	82.63 ± 0.25	65.37 ± 0.52	83.76 ± 0.36
SENet+Random flip	71.23 ± 0.50	83.07 ± 0.33	65.97 ± 0.49	81.72 ± 0.31	68.78 ± 0.49	83.52 ± 0.32	72.48 ± 0.49	84.12 ± 0.37
SENet+SA	**72.15 ± 0.49**	**84.65 ± 0.30**	**67.21 ± 0.50**	**82.53 ± 0.30**	**69.32 ± 0.50**	**84.57 ± 0.30**	**72.76 ± 0.50**	**87.61 ± 0.30**
TDM [16]	70.96 ± 0.49	83.59 ± 0.30	65.75 ± 0.50	81.74 ± 0.29	64.12 ± 0.50	81.34 ± 0.30	71.28 ± 0.51	86.45 ± 0.29
TDM+Cutout	68.42 ± 0.51	82.74 ± 0.32	62.82 ± 0.50	81.77 ± 0.37	69.95 ± 0.48	83.31 ± 0.30	66.58 ± 0.49	84.72 ± 0.31
TDM+Random flip	72.17 ± 0.49	85.87 ± 0.29	67.54 ± 0.50	83.15 ± 0.30	71.91 ± 0.49	86.03 ± 0.31	72.64 ± 0.51	87.14 ± 0.30
TDM+SA	**73.85 ± 0.50**	**86.73 ± 0.30**	**68.34 ± 0.49**	**83.46 ± 0.30**	**73.56 ± 0.49**	**86.75 ± 0.30**	**73.53 ± 0.50**	**88.39 ± 0.30**
CTX [32]	70.58 ± 0.50	82.85 ± 0.30	64.17 ± 0.50	80.67 ± 0.30	63.79 ± 0.49	80.82 ± 0.30	71.45 ± 0.50	85.76 ± 0.30
CTX+Cutout	68.17 ± 0.49	82.67 ± 0.30	64.73 ± 0.48	78.94 ± 0.35	68.27 ± 0.50	83.52 ± 0.32	67.54 ± 0.51	82.73 ± 0.31
CTX+Random flip	72.98 ± 0.50	84.19 ± 0.32	66.57 ± 0.50	82.95 ± 0.36	70.48 ± 0.50	85.06 ± 0.28	72.91 ± 0.52	87.26 ± 0.30
CTX+SA	**74.10 ± 0.50**	**85.92 ± 0.30**	**67.52 ± 0.50**	**83.48 ± 0.30**	**72.85 ± 0.50**	**85.62 ± 0.29**	**73.38 ± 0.50**	**88.43 ± 0.30**

TABLE V

THE CLASSIFICATION ACCURACIES OF FIVE-WAY FEW-SHOT CLASSIFICATION FOR THE ABLATION STUDY OF THE SBM AND JA MODULES ON THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE RESNET-12 BACKBONE STRUCTURE.

	Crop	*SBM*	*JA*	5-Way Accuracy (%)							
				CUB		Dogs		Aircraft		Flowers	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
(a)	×	×	×	68.48 ± 0.50	76.48 ± 0.28	62.90 ± 0.51	77.95 ± 0.34	59.48 ± 0.52	77.87 ± 0.33	68.53 ± 0.52	84.30 ± 0.35
(b)	✓	×	×	68.19 ± 0.51	83.08 ± 0.31	61.19 ± 0.49	79.04 ± 0.32	61.46 ± 0.57	78.55 ± 0.32	67.30 ± 0.48	85.57 ± 0.36
(c)	✓	✓	×	71.40 ± 0.48	84.53 ± 0.29	63.43 ± 0.50	80.24 ± 0.30	78.80 ± 0.45	85.96 ± 0.23	70.00 ± 0.50	87.35 ± 0.28
(d)	×	×	✓	70.42 ± 0.50	85.92 ± 0.29	61.52 ± 0.53	82.18 ± 0.31	62.66 ± 0.53	78.86 ± 0.33	70.11 ± 0.51	87.44 ± 0.31
Ours	✓	✓	✓	**75.57 ± 0.48**	**88.03 ± 0.29**	**70.32 ± 0.50**	**84.61 ± 0.32**	**85.70 ± 0.40**	**91.80 ± 0.19**	**74.22 ± 0.49**	**90.19 ± 0.28**

the proper integration of data augmentation and attention mechanism in SAA can substantially improve the classification performance for few-shot image classification. It is also obvious that SAA can provide more impressive improvements on classification accuracies for 1-shot scenarios than 5-shot scenarios. This may suggest that the few-shot setting with extremely limited labelled training images can benefit more from data augmentation with careful image selection.

Furthermore, to compare our selectively augmentation (SA) method, i.e. random cropping + SBM, with other traditional data augmentation methods, e.g. cutout and random flip, we evaluate their improvements on the classification accuracies of three state-of-the-art methods, SENet [15], TDM [16] and CTX [32] in Table IV. Cutout randomly crops 0.3 of the image area around a randomly generated centre point. Random flip randomly flips the image horizontally or vertically. For cutout and random flip, we generate five augmented images per original image to augment the support set. Clearly, not all augmentation methods can improve performance; for example, cutout largely decreases classification accuracies for all datasets, except for Aircraft. While random flip can enhance the performance for all scenarios, SA provides the best improvments in accuracy.

In summary, the SAA network shows superior classification performance compared with the state-of-the-arts. Moreover, the SA part alone is readily to be used as a data augmentation method to improve the performance of existing few-shot classifiers.

D. Ablation Study

To evaluate the impacts of the new modules in the SAA network, we conduct a series of ablation experiments based on the ResNet-12 backbone structure on the four datasets.

1) The effectiveness of the SBM and JA modules: We implement the following four scenarios to demonstrate the effectiveness of the data augmetation step and the two new modules and record the results in Table V: (a) we exclude data augmentation and the two modules, which is equivalent to the original ProtoNet; (b) we only keep random cropping for data augmentation; (c) we remove the JA module and keep the SBM module to select the most representative features from the augmented set to calculate the class prototypes and the metrics to assign test images; and (d) we remove data augmentation and the SBM module and directly apply the JA module to all original images to obtain spatial and channel-wise attentions.

Clearly, scenario-(a) provides the worst classification results for most cases. In addition, naively applying data augmentation as in scenario-(b) cannot always improve the classification accuracy compared with scenario-(a), especially for 1-shot settings. However, by utilising the SBM module in scenario-(c), we observe substantial increases in classification accuracies compared with scenario-(b), especially for the Aircraft dataset, which demonstrate the effectiveness of the proper selection

TABLE VI
THE CLASSIFICATION ACCURACIES OF FIVE-WAY FEW-SHOT CLASSIFICATION FOR THE ABLATION STUDY OF THE TWO DIMENSIONS (JA_S FOR THE CHANNEL DIMENSION AND JA_C FOR THE CHANNEL DIMENSION) IN THE JA MODULE ON THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE RESNET-12 BACKBONE STRUCTURE.

	SBM	*JA_S*	*JA_C*	5-Way Accuracy (%)							
				CUB		Dogs		Aircraft		Flowers	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
(a)	✓	×	✓	72.66 ± 0.48	86.76 ± 0.28	68.78± 0.51	83.97 ± 0.30	83.37 ± 0.43	89.64 ± 0.23	71.05 ± 0.49	89.78 ± 0.26
(b)	✓	✓	×	72.38 ± 0.49	86.29 ± 0.28	68.43 ± 0.50	83.11 ± 0.30	83.75 ± 0.41	89.92 ± 0.21	72.06 ± 0.50	88.56 ± 0.27
Ours	✓	✓	✓	**75.57 ± 0.48**	**88.03 ± 0.29**	**70.32 ± 0.50**	**84.61 ± 0.32**	**85.70 ± 0.40**	**91.80 ± 0.19**	**74.22 ± 0.49**	**90.19 ± 0.28**

TABLE VII
THE IMPACT OF THE CROPPING METHODS ON THE CLASSIFICATION ACCURACY FOR THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE RESNET-12 BACKBONE STRUCTURE.

Model	5-Way Accuracy (%)							
	CUB		*Dogs*		*Aircraft*		*Flowers*	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Random cropping	**75.57 ± 0.48**	88.03 ± 0.29	**70.32 ± 0.50**	**84.61 ± 0.32**	**85.70 ± 0.40**	**91.80 ± 0.19**	74.22 ± 0.49	90.19 ± 0.28
Centre cropping	74.83 ± 0.21	**88.46 ± 0.35**	70.17 ± 0.34	84.42 ± 0.30	83.75 ± 0.41	90.64 ± 0.23	**74.63 ± 0.48**	**90.56 ± 0.27**
Batch cropping	74.57 ± 0.46	87.60 ± 0.27	69.31 ± 0.50	82.93 ± 0.37	82.45 ± 0.45	89.50 ± 0.23	72.78 ± 0.49	88.99 ± 0.29

TABLE VIII
THE IMPACT OF THE ATTENTION MECHANISM ON THE CLASSIFICATION ACCURACY FOR THE CUB, DOGS, AIRCRAFT AND FLOWERS DATASETS UNDER THE RESNET-12 BACKBONE STRUCTURE.

Model	5-Way Accuracy (%)							
	CUB		*Dogs*		*Aircraft*		*Flowers*	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Ours	**75.57 ± 0.48**	**88.03 ± 0.29**	**70.32 ± 0.50**	**84.61 ± 0.32**	**85.70 ± 0.40**	**91.80 ± 0.19**	**74.22 ± 0.49**	**90.19 ± 0.28**
Ours+TDM [16]	74.67 ± 0.47	86.34 ± 0.28	69.82 ± 0.50	84.23 ± 0.30	84.48 ± 0.51	90.27 ± 0.31	73.84 ± 0.50	89.42 ± 0.30
Ours+CTX [32]	74.60 ± 0.51	87.34 ± 0.30	68.79 ± 0.49	84.12 ± 0.28	84.41 ± 0.50	89.96 ± 0.30	73.63 ± 0.51	89.35 ± 0.31

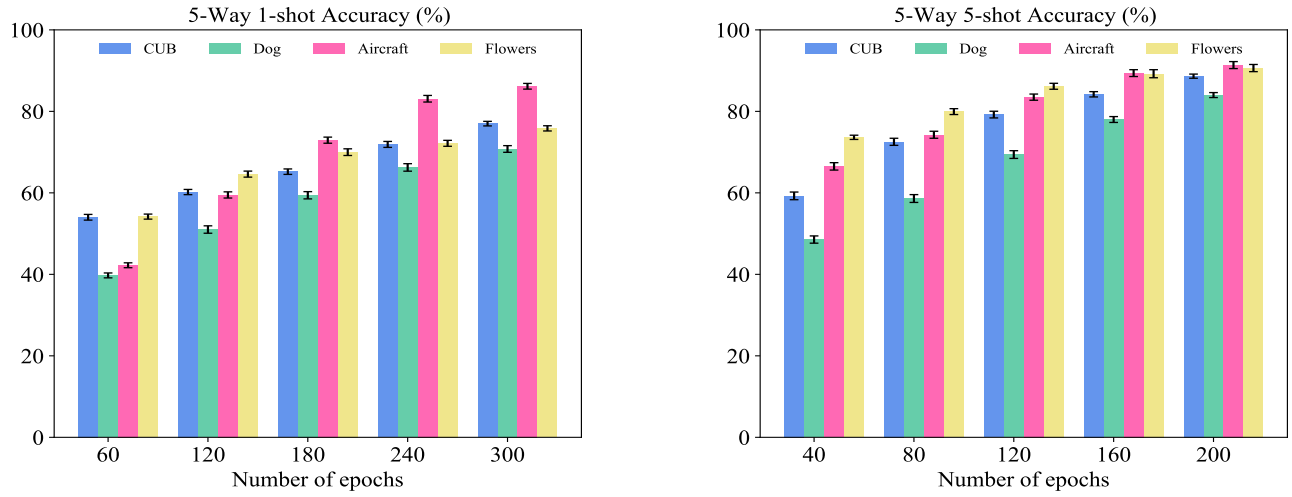


Fig. 7. The impact of the number of epochs on the validation classification accuracy for the CUB, Aircraft, Dogs and Flowers datasets.

of augmented features. Moreover, only using the attention module in scenario-(d) can improve the classification performance compared with scenarios-(a) and (b), but cannot always beat scenario-(c) on all datasets, which leads to the following two conclusions. First, reweighting the discriminative spatial regions and channels can bring benefits to the classification task. Second, using a carefully selected training set on a simple model, e.g. scenario-(c), can provide competitive and some-

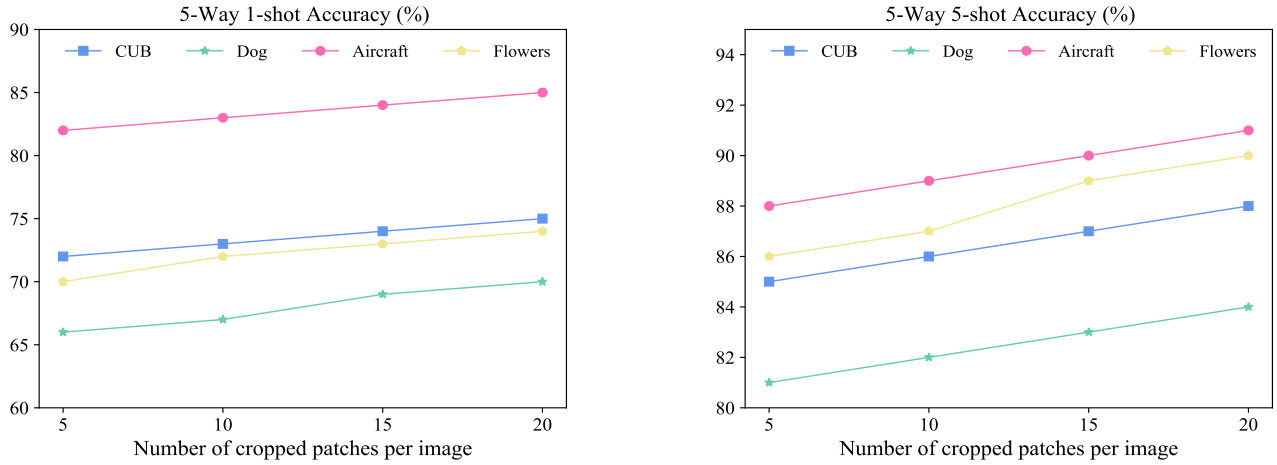


Fig. 8. The impact of the number of cropped patches per image on the validation classification accuracy for the CUB, Aircraft, Dogs and Flowers datasets.

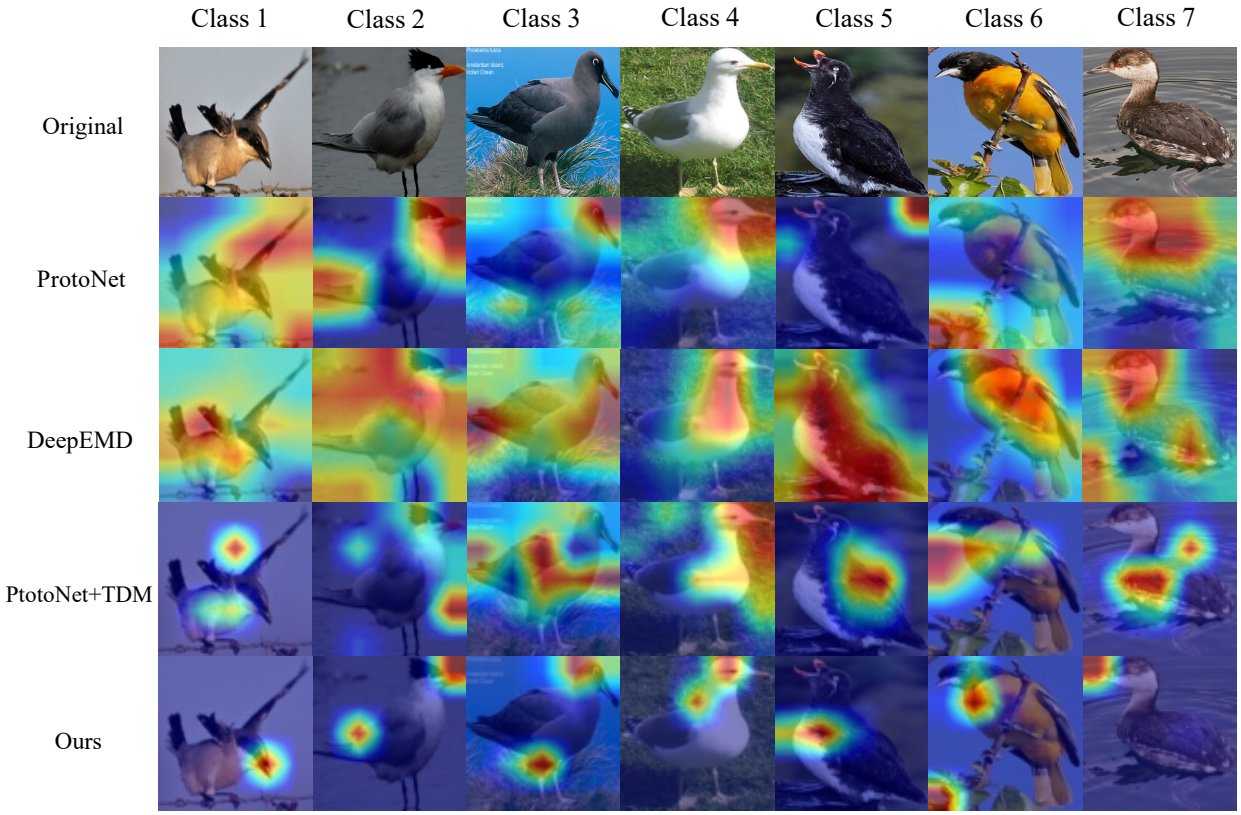


Fig. 9. Visualisations of discriminative regions of example images in the CUB dataset for the ProtoNet, DeepEMD, ProtoNet+TDM and our proposed method.

times better classification accuracies than using models with more complex structures, e.g. scenario-(d). Finally, utilising both modules can achieve the highest classification accuracies on all datasets.

Furthermore, in Table VI, we compare the classification performances of only involving attentions in one dimension, i.e. either spatial-wise or channel-wise. It is obvious that using the spatial-wise or channel-wise attentions can provide competitive classification accuracies, and using both has the best results.

2) The impact of the data augmentation methods: To verify the impact of the data augmentation methods on the test accuracy, we conduct experiments on two additional cropping methods, centre cropping and batch cropping, in Table VII. Note that since SBM aims to select the best matched local areas across the support and query images and eleminate background noise, data augmentation methods such as image fipping and rotation are not considered here, because the

background is still included in the augmentations. In the experiments, we tailor centre cropping for SBM by cropping the centre area with the cropping ratio randomly selected from [0.3,1]. For batch cropping, we divide the image into nine non-overlapping patches.

It is obvious that random cropping can provide the best or competitive classification accuracies for all datasets and K-shot settings. Batch cropping usually performs worse than the other two methods, because the non-overlapping patches may segment the important features to different patches. With the effect of SBM, only one patch of an image is selected for subsequent analysis, and thus batch cropping cannot always capture a sufficient amount of discriminative features for classification. On the other hand, centre cropping can sometimes perform better than random cropping depending on the characteristics of the images. For example, for the Flowers dataset, the discriminative regions are mostly located in the centre of the images, e.g. small petals or stamens. Hence centre cropping can provide better patches for SBM to select. However, for datasets such as Dogs and Aircraft, the discriminative regions are usually not in the centre of the image, e.g. heads of dogs or wings of aircraft, and thus centre cropping cannot capture sufficient discriminative information for classification. From the above analysis, we can conclude that data augmentation method has to be properly chosen to work together with SBM.

3) The impact of the attention methods: To evaluate the effectiveness of the proposed JA module, we replace it by two state-of-the-art attention mechanisms, the task discrepancy maximisation (TDM) module [16] and the CrossTransformers (CTX) [32], and report the classification accuracies in Table VIII. Obviously, only re-weighting the channel dimension in TDM and the spatial dimensions in CTX cannot beat the classification performance of using the JA module that can re-weight both spatial and channel dimensions.

4) The impact of the number of epochs: We also study the impact of the number of epochs on the classification accuracy for all four benchmark datasets. There is an obvious upward trend of classification accuracy when the number of epochs increases, and the trend starts to slow down when the number of epochs reaches 300 and 200 for the 1-shot and 5-shot scenarios, respectively.

5) The impact of the number of cropped patches per image: Here we test the impact of the number of cropped patched per image on the test accuracy in Fig. 8. Clearly, as the number increases, the test accuracy also increases, which makes sense because the more patches from different locations, the higher the likelihood that the key features are presented in one patch. An increase in the number of patches may contribute to further improvement in test accuracy, but with a burden on the computational resources.

E. Visualisations of discriminative regions and attention maps

In this section, we visualise the discriminative regions of example images of seven classes in the CUB dataset obtained by ProtoNet, ProtoNet+TDM, DeepEMD and our proposed SAA network in Fig. IV-C. ProtoNet can focus on some discriminative regions but also involve nuisance backgrounds, while DeepEMD usually includes most of the object. ProtoNet+TDM can provide better focusing areas compared with the previous two methods. Clearly, our proposed SAA network show the best precise focus on the most discriminave regions, e.g. beaks, eyes, bellies and legs.

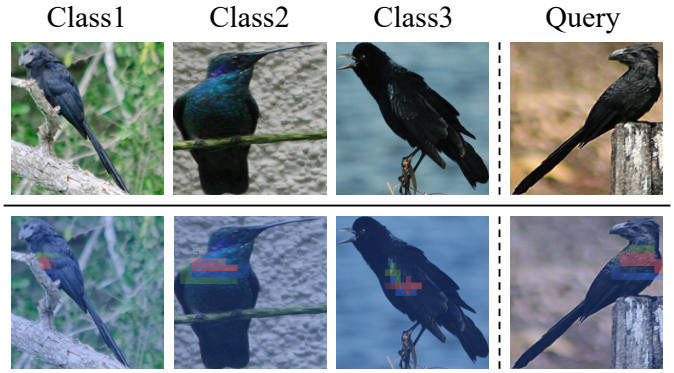


Fig. 10. Visualisations of the attention maps obtained in the JA module. The highest attentions are labelled in the original images of the examples in the CUB dataset. Red areas receive the highest attentions, followed by the green and blue areas.

In addition, in Fig. 10, we visually assess the effectiveness of the attention maps obtained in the JA module, i.e. $\mathbf{W}^S$ and $\mathbf{W}^Q$ in equation (2). We label the attention maps with large values in the original images to highlight the areas receiving high weights. The red patches receive the highest weights, followed by blue and green areas. It is obvious that the query weights $\mathbf{W}^Q$ identify the neck and wings of the query bird as the most important body parts for classification, while similar parts in the support classes are also selected by the support weights $\mathbf{W}^S$.

F. Computational complexity

TABLE IX
THE NUMBER OF PARAMETERS AND FLOPS OF DN4, DEEPEMD, FRN AND SAA. THE SMALLER THE TWO METRICS, THE HIGHER THE COMPUTATIONAL EFFICIENCY.

	Backbone	params.(M)	FLOPs.(G)
DN4 [6]	ResNet-12	16.7	68.48
DeepEMD [26]	ResNet-12	15.2	37.29
ProtoNet [4]+TDM [16]	ResNet-12	19.5	79.96
SAA(Ours)	ResNet-12	14.5	29.69

Finally, we compare the computational complexity of SAA with three state-of-the-art methods, DN4 [6], DeepEMD [26] and ProtoNet [4]+TDM [16], in terms of the number of parameters and the FLOPs in Table IX. It is clear that SAA has the least number of parameters and the FLOPs compared with the state-of-the-art competitors, while shows superior classification performance over them.

V. CONCLUSION

In this paper, for few-shot image classification, we propose a unified framework, the selectively augmented attention (SAA)

network, that carefully integrate data augmentation and attention mechanism in an end-to-end fashion. We propose the new SBM module to select the most representative images that focus more on the objects and less on the background to alleviate potential noises in the augmented set. Moreover, we propose the novel JA module to learn the spatial and channel-wise attentions together on the selected features to identify discriminative spatial regions and channels. Extensive experiments and ablation study demonstrate the superior classification performance of the SAA network and the effectiveness of the new modules. Besides the application to few-shot image classification, the appropriate modifications of SAA also have potential to solve other few-shot learning problems, such as few-shot object detection [40], [41], few-shot semantic segmentation [42] and few-shot video object segmentation [43], as well as zero-shot image classification [44]. Given the observation that cropping methods have impact on the classification performance of SAA, in the future, we aim to develop a more intelligent data augmentation methodology that can capture local regions focusing more on the objects. Moreover, instead of using the simple cosine similarity in SBM, more sophisticated while efficient network modules can be designed in the future to learn the similarity between support and query patches.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62176110, U19B2036, 62225601, Beijing Natural Science Foundation Project under Grant Z200002, the Program of Youth Innovative Research Team of BUPT under Grant 2023QNTD02, the Key Research and Development Program of Gansu Province under Grant 22YF7GA130, Hong-liu Distinguished Young Talents Foundation of Lanzhou University of Technology and the Royal Society under International Exchanges Award IEC\NSFC\201071.

REFERENCES

[1] W. Y. Chen, Y. C. Liu, Z. Kira, Y. C. F. Wang, and J. B. Huang, "A closer look at few-shot classification," in *arXiv preprint arXiv:1904.04232*, 2019.

[2] X. Li, X. Yang, Z. Ma, and J.-H. Xue, "Deep metric learning for few-shot image classification: A review of recent developments," *Pattern Recognition*, p. 109381, 2023.

[3] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1091–1102, 2020.

[4] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[5] F. Zhou, L. Zhang, and W. Wei, "Meta-generating deep attentive metric for few-shot classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6863–6873, 2022.

[6] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7260–7268.

[7] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, and J.-H. Xue, "BSNet: Bi-similarity network for few-shot fine-grained image classification," *IEEE Transactions on Image Processing*, 2021.

[8] W. Zhang, X. Liu, Z. Xue, Y. Gao, and C. Sun, "NDPNet: A novel non-linear data projection network for few-shot fine-grained image classification," *arXiv preprint arXiv:2106.06988*, 2021.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, 2019.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, 2014.

[11] V. Kumar, H. Glaude, C. de Lichy, and W. Campbell, "A closer look at feature space data augmentation for few-shot intent classification," *arXiv preprint arXiv:1910.04176*, 2019.

[12] H. Wang, S. Tian, Y. Fu, J. Zhou, J. Liu, and D. Chen, "Feature augmentation based on information fusion rectification for few-shot image classification," *Scientific Reports*, 2023.

[13] H. Song, B. Deng, M. Pound, E. Özcan, and I. Triguero, "A fusion spatial attention approach for few-shot learning," *Information Fusion*, 2022.

[14] S. Yan, S. Zhang, X. He *et al.*, "A dual attention network with semantic embedding for few-shot learning," in *AAAI*, vol. 33, 2019, pp. 9079–9086.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] S. Lee, W. Moon, and 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.

[17] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[18] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[19] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[20] W.-H. Chu, Y.-J. Li, J.-C. Chang, and Y.-C. F. Wang, "Spot and learn: A maximum-entropy patch sampler for few-shot image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6251–6260.

[21] Z. Hu, L. Shen, S. Lai, and C. Yuan, "Task-adaptive feature disentanglement and hallucination for few-shot classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] S.-L. Xu, F. Zhang, X.-S. Wei, and J. Wang, "Dual attention networks for few-shot fine-grained recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2911–2919.

[23] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," *Advances in neural information processing systems*, vol. 34, pp. 24 581–24 592, 2021.

[24] J. Wu, D. Chang, A. Sain, X. Li, Z. Ma, J. Cao, J. Guo, and 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, no. 3, 2023, pp. 2821–2829.

[25] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *International Conference on Learning Representations*, 2019.

[26] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213.

[27] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8460–8469.

[28] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Mixture-based feature space learning for few-shot image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9041–9051.

[29] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1666–1680, 2020.

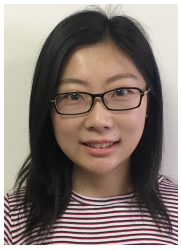
[30] H. Tang, Z. Li, Z. Peng, and J. Tang, "BlockMix: Meta regularization and self-calibrated inference for metric-based meta-learning," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 610–618.

[31] X. Wang, X. Wang, B. Jiang, and B. Luo, "Few-shot learning meets transformer: Unified query-support transformers for few-shot classifica-

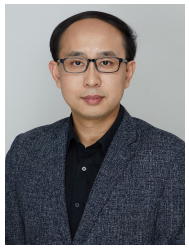
tion," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
[32] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: spatially-aware few-shot transfer," 2020.
[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.
[34] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
[35] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *CVPR*, 2011.
[36] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," *arXiv preprint arXiv:1803.00676*, 2018.
[37] P. Chikontwe, S. Kim, and S. H. Park, "CAD: Co-adapting discriminative features for improved few-shot classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 554–14 563.
[38] D. Wertheimer, L. Tang, and 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.
[39] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 403–412.
[40] M. Köhler, M. Eisenbach, and H.-M. Gross, "Few-shot object detection: a comprehensive survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[41] J. Yao, L. Han, G. Guo, Z. Zheng, R. Cong, X. Huang, J. Ding, K. Yang, D. Zhang, and J. Han, "Position-based anchor optimization for point supervised dense nuclei detection," *Neural Networks*, vol. 171, pp. 159–170, 2024.
[42] H. Gao, J. Xiao, Y. Yin, T. Liu, and J. Shi, "A mutually supervised graph attention network for few-shot segmentation: the perspective of fully utilizing limited samples," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
[43] N. Liu, K. Nan, W. Zhao, Y. Liu, X. Yao, S. Khan, H. Cholakkal, R. M. Anwer, J. Han, and F. S. Khan, "Multi-grained temporal prototype learning for few-shot video object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 862–18 871.
[44] D. Cheng, G. Wang, B. Wang, Q. Zhang, J. Han, and D. Zhang, "Hybrid routing transformer for zero-shot learning," *Pattern Recognition*, vol. 137, p. 109270, 2023.

Xiaoxu Li received the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2012. She is currently an Associate 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. She is also a member of the China Computer Federation.

Xiangyang Wang obtained a Bachelor's degree in Electronic Information Engineering from Hunan City College, China in 2021. She is currently a graduate student at Lanzhou University of Technology. Her research interests include machine learning and small sample learning.



Rui Zhu received the Ph.D. degree in statistics from University College London in 2017. She is a Senior Lecturer in Statistics in the Faculty of Actuarial Science and Insurance, City St George's, University of London. Her research interests include machine learning, computer vision and interdisciplinary applications in actuarial science. She serves as the Associate Editor for IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology and Neurocomputing.



Zhanyu Ma (Senior Member, IEEE) received the PhD degree in electrical engineering from the KTH-Royal Institute of Technology, Sweden, in 2011. Since 2019, he has been a professor with the Beijing University of Posts and Telecommunications, Beijing, China. From 2012 to 2013, he was a postdoctoral research fellow with the School of Electrical Engineering, KTH-Royal Institute of Technology. From 2014 to 2019, he has been an associate professor with the Beijing University of Posts and Telecommunications, Beijing, China. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing, and data mining.

Jie Cao received the M.E. degree from Xi'an Jiaotong University, China, in 1994. She is currently a Professor and a Vice President of the Lanzhou University of Technology. Her research interests include machine learning, pattern recognition, speech and speaker recognition, information fusion, and computer vision.

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.