Query-aware Cross-mixup and Cross-reconstruction for Few-shot Fine-grained Image Classification

Zhimin Zhang, Dongliang Chang, Rui Zhu, Xiaoxu Li, Zhanyu Ma, *Senior Member, IEEE*, Jing-Hao Xue, *Member, IEEE*

***Abstract*—Few-shot fine-grained image classification is prominent but challenging in computer vision, aiming to distinguish sub-classes under the same parent class but with only a few labeled support samples. Data augmentation techniques were explored to address the few-shot issue, but they often fail to mitigate the bias between support and query samples. Therefore, in this paper we propose a query-aware cross-mixup and cross-reconstruction method to address both few-shot and fine-grained issues. Specifically, in the training phase, we randomly select query samples and mix them with the support samples from the same class to augment the support set. This first strategy ensures the augmented support set query-aware within each sub-class. Then, we reconstruct both query samples and support samples from both original and cross-mixed support samples, thus leveraging both cross-reconstruction and self-reconstruction to enhance classification. This second strategy, enabling the reconstruction also query-aware, further mitigates the bias between support and query samples, leading to more reliable generalization. We evaluate our proposed method on four widely used few-shot fine-grained image classification datasets, and experimental results demonstrate its effectiveness in achieving the state-of-the-art classification performance.**

***Index Terms*—Data augmentation, Few-shot image classification, Fine-grained image classification.**

I. Introduction

FINE-GRAINED image classification is an important topic in computer vision and pattern recognition. It is particularly challenging due to the extremely similar sub-classes that yield minimal inter-class variance [1], [2]. Concurrently, variations in pose, age, and background, etc. within each sub-class lead to substantial intra-class variance [3], [4]. These two factors make the fine-grained image classification task highly challenging. With the advancements in deep learning, there have been significant strides in image classification [5]–[13]. However, these achievements usually depend on large sample sizes, which may not be available in many practices. As a result, few-shot fine-grained image classification has emerged.

Current techniques to enhance few-shot image classification include metric learning, transfer learning, and data augmentation. Among them, data augmentation, aiming at augmenting the support set, has garnered significant attention in recent years due to its simplicity and plug-and-play nature [14], [16]–[20]. Xu et al. [14] propose a feature disentanglement framework (VFD) that can provide augmented features with randomly sampled intra-class variations. Li et al. [16] develop adversarial feature hallucination networks hallucinating diverse and discriminative features. Zhao et al. [17] present mirror mapping networks to generate the common features for augmentation based on textual descriptions and knowledge graph. However, as illustrated in Figure 1, they still fall short in mitigating the bias between support and query samples, which can be crucial for correctly classifying fine-grained samples. This inadequacy underscores a key challenge for few-shot fine-grained image classification [1], [21], [22].

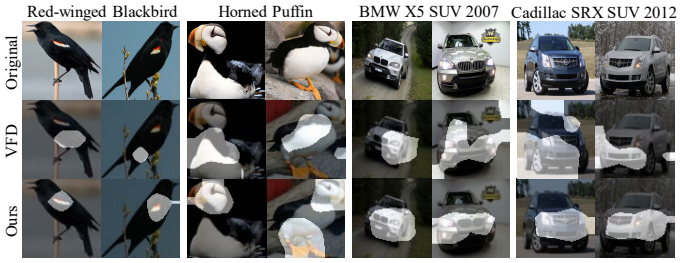


Figure 1. Visualization of the regions of interest (i.e. the bright areas) identified by VFD [14] and our method, using Grad-CAM [15] on samples from the CUB-200-2011 dataset. Compared with VFD, our method can identify discriminative regions more accurately.

Therefore, in this paper we propose a query-aware cross-mixup and cross-reconstruction method to address both few-shot and fine-grained issues. Specifically, in the training phase, we randomly select query samples and mix them with the support samples from the same class to augment the support set. This first strategy ensures the augmented support set query-aware within each sub-class. Then, we reconstruct both query samples and support samples from both original and cross-mixed support samples, thus leveraging both cross-reconstruction and self-reconstruction to enhance classification. This second strategy, enabling the reconstruction also query-aware, further mitigates the bias between support and query samples, leading to more reliable generalization.

In sum, our novelties and contributions are three-fold:

1) We propose a data augmentation strategy called query-aware cross-mixup that generates new support samples with imported information from the query samples

Corresponding author: Dongliang Chang; Xiaoxu Li.

Z. Zhang and X. Li are with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China.

D. Chang is with the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: changdongliang@tsinghua.edu.cn

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

Z. Ma is with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.

J.-H. Xue is with the Department of Statistical Science, University College London, London, WC1E 6BT, U.K.

within the same class. This strategy ensures the augmented support set query-aware within each sub-class, hence better generalization for fine-grained sub-classes.
2) We develop a strategy leveraging both cross-reconstruction and self-reconstruction to reconstruct both query samples and support samples from both original and cross-mixed support samples. This strategy, enabling the reconstruction also query-aware, further mitigates the bias between support and query samples, leading to a more reliable classification.
3) Experimental results on four widely used few-shot fine-grained image classification datasets demonstrate the proposed method's effectiveness in achieving the state-of-the-art classification performance.

II. Related Work

A. Few-shot Fine-grained Image Classification

Few-shot fine-grained image classification aims to distinguish sub-classes within a parent class using only a few labeled samples. Recently, some methods have achieved significant progress [1], [23]–[26]. Li et al. [23] propose a bi-similarity network (BSNet) that utilizes two different similarity measures to improve the performance with small model complexity. Xu et al. [1] introduce a dual attention network, which hard-attention and soft-attention branches jointly learn global and local features to classify fine-grained data. Lee et al. [25] present the task discrepancy maximization module (TDM) for fine-grained few-shot classification, which learns task-specific channel weights. Zha et al. [26] develop a two-stage background suppression and foreground alignment framework.

Different from them, we use query-aware data cross-augmentation and feature cross-reconstruction to align query samples with support samples, thus improving the few-shot fine-grained image classification performance.

B. Few-shot Learning with Data Augmentation

Data augmentation aims to increase sample information to assist feature learning and improve the generalization ability of the model [20], [27]–[29]. For few-shot learning, Wang et al. [30] optimize both the meta-learner and the data generator to generate additional training samples. Gidaris et al. [31] propose a self-supervised data augmentation method via rotation at different angles. Phoo et al. [32] introduce a representation learning method that allows few-shot learners to leverage coarsely-labeled data before evaluation. Zhang et al. [33] proposed a hierarchical tree structure-aware method to generate multiple groups of augmented images. Ma et al. [34] proposed partner-assisted learning with supervised contrastive learning.

Different from these methods, our data augmentation approach is query-aware cross-mixup. It ensures the augmented support set query-aware within each sub-class, hence offering better generalization for fine-grained sub-classes.

C. Few-shot Learning with Feature Reconstruction

Recently, some feature reconstruction approaches have achieved excellent results in few-shot learning [25], [35]–[38]. Wertheimer et al. [35] propose feature map reconstruction networks, which use ridge regression to reconstruct query sample features from support sample features, alleviating metric bias. Li et al. [37] propose a locally-enriched cross-reconstruction network (LCCRN) to extract more discriminative local representations. Sun et al. [38] introduce an $l_{2,1}$-norm regularization to guide feature reconstruction towards semantically rich target regions. Wu et al. [36] introduce a bi-reconstruction mechanism to simultaneously accommodate for inter-class and intra-class variations.

Different from these methods, we develop a strategy leveraging both cross-reconstruction and self-reconstruction to reconstruct both query samples and support samples and thus alleviate metric bias.

III. The Proposed Method

A. Problem Formulation

Given a dataset $\mathcal{D} = \{(x_i, y_i), y_i \in \mathcal{L}\}_{i=1}^{N}$, following the setting in [23], we divide it into three parts: $\mathcal{D}_{train} = \{(x_i, y_i),\ y_i \in \mathcal{L}_{train}\}_{i=1}^{N_{train}}$, $\mathcal{D}_{val} = \{(\bar{x}_i, \bar{y}_i),\ \bar{y}_i \in \mathcal{L}_{val}\}_{i=1}^{N_{val}}$, $\mathcal{D}_{test} = \{(x_i^*, y_i^*),\ y_i^* \in \mathcal{L}_{test}\}_{i=1}^{N_{test}}$, where $\mathcal{D}_{train} \bigcap \mathcal{D}_{val} \bigcap \mathcal{D}_{test} = \emptyset$. We train the model on the $\mathcal{D}_{train}$, validate it on $\mathcal{D}_{val}$ to select appropriate hyperparameters, and finally use $\mathcal{D}_{test}$ to evaluate the performance of the trained model. In the C-way K-shot few-shot setting, we randomly select C classes from the training set, with M samples randomly selected from each class. Among them, K samples form the support set $\mathcal{S} = \{(x_i, y_i), y_i \in \mathcal{L}_{train}\}_{i=1}^{C\times K}$, and the rest $M-K$ samples form the query set $\mathcal{Q} = \{(x_j, y_j), y_j \in \mathcal{L}_{train}\}_{j=1}^{C\times(M-K)}$; $\mathcal{S}$ and $\mathcal{Q}$ together form a task $\mathcal{T}$ in training. Similarly, we construct tasks $\bar{\mathcal{T}}$ for validation and $\mathcal{T}^*$ for test.

B. Overview of the Proposed Method

As shown in Figure 2, our method consists of four modules: feature embedding module f_θ, cross-mixup module, feature calibration module, and Euclidean metric module.

In the training phase, firstly, a meta-task with support set S and query set Q is input into the cross-mixup module, where each support sample is randomly mixed with a query sample from the same class. Then, the original and new support samples, as well as the query samples, are fed into feature embedding module f_θ to produce S_c, S_c^+ and Q, respectively. Then, they enter the feature calibration module, containing two branches that use two support sets, S_c and S_c^+, to produce for S_c and Q four reconstructed feature maps, two for each: $\tilde{S}_c$, $\tilde{S}_c^+$, $\tilde{Q}_c$, and $\tilde{Q}_c^+$. Finally, the original features and four reconstructed features enter the Euclidean module to calculate four corresponding distances and make the final decision of classification.

In the test phase, we only use the original support set to reconstruct the query sample, and use the distance between the reconstructed query feature $\tilde{Q}_c$ and the original query feature Q for the final classification.

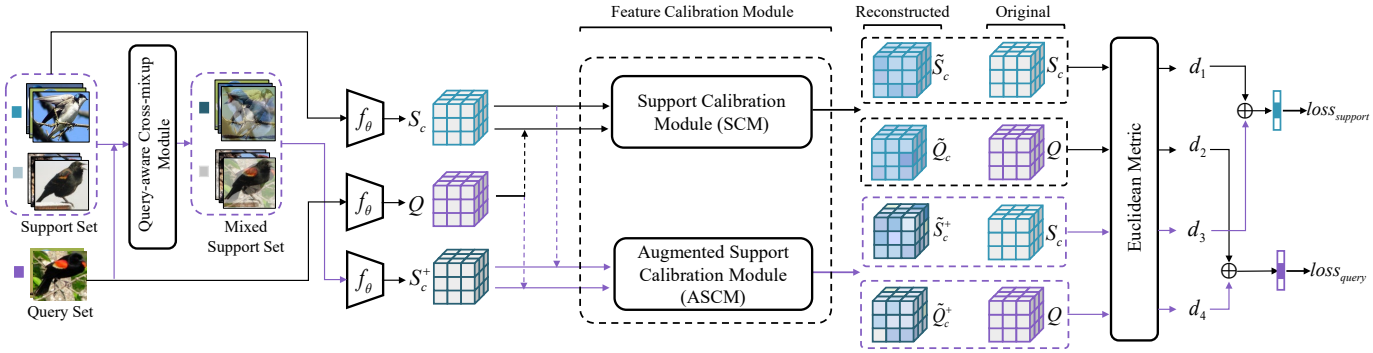


Figure 2. Diagram of the proposed method. It consists of four modules: feature extractor f_θ, cross-mixup module, feature calibration module, and Euclidean metric module. The black line shows the process of original support samples' participation the in model training, while the purple line is for the newly generated support samples.

Algorithm 1 Cross-mixup

Input: Support set S, query set Q, C-way, K-shot.
Output: Set S^+ of augmented support samples.

$\delta \sim \text{Uniform}(0, 1)$
for class $c = 1, \ldots, C$ **do**
  For each support sample S_{ck}, $k = 1, \ldots, K$, randomly select a query image Q_{ck} from Q_c of class c
  $S^+_{ck} = \delta \times S_{ck} + (1 - \delta) \times Q_{ck}$
return S^+

C. Cross-mixup Module

As summarized in Algorithm 1, our query-aware cross-mixup strategy is very simple. For each support sample S_{ck} from class c in a support set S, we randomly select a query image Q_{ck} from Q_c, where Q_c is the set of the query samples also from class c, and then linearly combine these two samples with a random weight to generate a new support sample. This strategy, although simple, ensures the augmented support set query-aware within each sub-class, hence better diversity and generalization for fine-grained sub-classes.

D. Feature Calibration Module and Euclidean Metric Module

To fully exploit the augmented support samples, we design a feature calibration module, which as shown in Figure 2 contains two modules for using self-reconstruction and cross-reconstruction to reconstruct four feature maps. The support calibration module uses the original support features S_c to reconstruct two feature maps, $\tilde{S}_c$ for support samples S_c and $\tilde{Q}_c$ for query samples Q_c, respectively. In contrast, the augmented support calibration module uses the augmented data S^+_c to reconstruct two feature maps, $\tilde{S}^+_c$ for support samples S_c and $\tilde{Q}^+_c$ for query samples Q_c, respectively.

For convenience of presentation, we shall describe the four reconstructions together with the Euclidean metric module in more detail as follows.

1) Reconstruct query from support: $S_c \to \tilde{Q}_c$: We express the feature of the support set and the feature of the query sample in the form of matrix. That is, let matrix $S_c \in \mathbb{R}^{KR \times d}$ represent the support set feature of class c, where K is the shot number, $R = H \times W$, and d is the number of channels. In the meantime, let matrix $Q_j \in \mathbb{R}^{R \times d}$ represent the matrix of the j-th query sample.

As with [35], we use ridge regression to estimate a weight matrix $M_w \in \mathbb{R}^{R \times KR}$, such that $Q_j \approx M_w S_c$:

$$\tilde{M}_w = \arg\min_{M_w} \|Q_j - M_w S_c\|^2 + \lambda_1 \|M_w\|^2, \quad (1)$$

where λ_1 is the penalty weight. The ridge regression has a closed-form solution:

$$\tilde{M}_w = Q_j S_c^\top (S_c S_c^\top + \lambda_1 I)^{-1}, \quad (2)$$

where $I \in \mathbb{R}^{KR \times KR}$ is the identity matrix. Therefore, the cross-reconstructed $\tilde{Q}_j$ from S_c can be expressed as

$$\tilde{Q}_{cj} = \gamma_1 \tilde{M}_w S_c = \gamma_1 Q_j (S_c^\top S_c + \lambda_1 I)^{-1} S_c^\top S_c, \quad (3)$$

where λ_1 and γ_1 can be designed to improve the stability of training by setting

$$\lambda_1 = \frac{KR}{d} e^{\alpha_1}, \quad \gamma_1 = e^{\beta_1}, \quad (4)$$

where α_1 and β_1 are learnable parameters.

Then, Q_j and its cross-reconstructed feature map $\tilde{Q}_{cj}$ for class c are input into the Euclidean metric module to calculate their distance (d_2 in Figure 2) as

$$d_2 = d_{cj} = \left\| Q_j - \tilde{Q}_{cj} \right\|^2, \quad (5)$$

2) Reconstruct query from augmented support: $S^+_c \to \tilde{Q}^+_c$: Similarly, we can use ridge regression to get cross-reconstructed query feature map $\tilde{Q}^+_{cj}$ from the augmented support feature maps S^+_c, and its distance (d_4 in Figure 2) from Q_j as

$$\tilde{Q}^+_{cj} = \gamma_2 Q_j (S^{+\top}_c S^+_c + \lambda_2 I)^{-1} S^{+\top}_c S^+_c, \quad (6)$$

$$d_4 = d^+_{cj} = \left\| Q_j - \tilde{Q}^+_{cj} \right\|^2. \quad (7)$$

3) Reconstruct support from support: $S_c \to \tilde{S}_c$: When self-reconstructing the i-th feature map S_i from the feature maps in S_c, we can use the following formula:

$$\tilde{S}_{ci} = \gamma_3 S_i (S_c^\top S_c + \lambda_3 I)^{-1} S_c^\top S_c, \quad (8)$$

and its distance (d_1 in Figure 2) from S_i:

$$d_1 = d_{ci} = \left\| S_i - \tilde{S}_{ci} \right\|^2. \quad (9)$$

4) Reconstruct support from augmented support: $S_c^+ \rightarrow \tilde{S}_c^+$: Similarly, we can reconstruct the i-th support map S_i from the augmented support feature maps S_c^+ to obtain $\tilde{S}_{ci}^+$ and calculate its distance (d_3 in Figure 2) from S_i as

$$\tilde{S}_{ci}^+ = \gamma_4 S_i (S_c^{+\top} S_c^+ + \lambda_4 I)^{-1} S_c^{+\top} S_c^+, \tag{10}$$

$$d_3 = d_{ci}^+ = \left\| S_i - \tilde{S}_{ci}^+ \right\|^2. \tag{11}$$

E. Loss Functions

1) Query loss: Considering both the distances in Eq.(5) and Eq.(7), we can obtain the probability of predicting the j-th query sample into class c as

$$P(y_j = c|Q_j) = \frac{e^{-\xi_1(d_{cj}+d_{cj}^+)}}{\sum_{c' \in C} e^{-\xi_1(d_{c'j}+d_{c'j}^+)}}, \tag{12}$$

where ξ_1 is a learnable temperature factor.

Then the cross-entropy loss $loss_{query}$ for classifying query samples can be expressed as

$$loss_{query} = -\frac{1}{M-K} \sum_{j=0}^{M-K} \mathbf{y}_j^\top \log(P(\mathbf{y}_j|Q_j)), \tag{13}$$

where $M-K$ is the number of query samples, $\mathbf{y}_j$ is the one-hot vector and $P(\mathbf{y}_j|Q_j)$ is the vector of predicted probabilities.

2) Support loss: Similarly, considering both the distances in Eq.(9) and Eq.(11), we can obtain the probability of predicting the i-th support sample into class c as

$$P(y_i = c|S_i) = \frac{e^{-\xi_2(d_{ci}+d_{ci}^+)}}{\sum_{c' \in C} e^{-\xi_2(d_{c'i}+d_{c'i}^+)}}, \tag{14}$$

and the cross-entropy loss $loss_{support}$ for classifying support samples as

$$loss_{support} = -\frac{1}{K} \sum_{i=0}^{K} \mathbf{y}_i^\top \log(P(\mathbf{y}_i|S_i)), \tag{15}$$

where K is shot number, and y_i is the one-hot vector.

3) Auxiliary loss: In addition, we follow [35] to use an auxiliary loss to make the support classes orthogonal to each other and increase the distance between classes:

$$loss_{aux} = \sum_{i \in C} \sum_{j \in C, j \neq i} \left\| \hat{S}_i \hat{S}_j^T \right\|, \tag{16}$$

where $\hat{S}$ is the normalized support sample feature.

4) Total loss: Finally, we use the total $Loss$ for model training:

$$Loss = loss_{support} + loss_{query} + loss_{aux}. \tag{17}$$

F. Inference

In the test phase, we do not use the augmented support set. For every test image, only the reconstruction of query from support $S_c \rightarrow \tilde{Q}$ is conducted and the distance in the form of Eq.(5) to calculate the the prediction probabilities:

$$P(y_j = c|Q_j) = \frac{e^{d_{cj}}}{\sum_{c' \in C} e^{d_{c'j}}}. \tag{18}$$

The query sample will be classified into the class with the highest probability.

IV. Experimental Analysis

A. Datasets

To evaluate the effectiveness of our method, we use four fine-grained benchmark datasets: *CUB-200-2011*, *Flowers*, *Stanford-Cars*, and *FGVC-Aircraft*. For each dataset, follow the setting in [23], we divide them into the training set $\mathcal{D}_{train}$, the validation set $\mathcal{D}_{val}$, and the test set $\mathcal{D}_{test}$. All images in the four datasets are resized to 84×84.

CUB-200-2011 (CUB) [52] contains 11,788 images of 200 bird species. We divide it into a training set with 100 classes, a validation set with 50 classes, and a test set with 50 classes.

Flowers [53] consists of 102 categories of common flowers and each category consists of 40 to 256 images. We randomly divide this dataset into a training set with 51 classes, a validation set with 26 classes, and a test set with 25 classes.

Stanford-Cars (Cars) [54] contains 16,185 images of 196 classes of cars. We randomly select 130 classes to form the training set, 17 classes for the validation set, and 49 classes for the test set.

FGVC-Aircraft (Aircraft) [55] contains 10,000 images of aircraft spanning 100 aircraft models. We randomly select 50 classes to form a training set, 25 classes for a validation set, and 25 classes for a test set.

In addition, for a comprehensive evaluation, we also test our method on three coarse-grained datasets: *mini-ImageNet* [49], *tiered-ImageNet* [56], and *FC-100* [57].

mini-ImageNet [49] consists of 100 categories, each category with 600 images. Following [58], we divide the dataset into 64 classes for training, 16 classes for validation, and 20 classes for testing.

tiered-ImageNet [56] consists of 351 categories for training, 97 classes for validation, and 160 for testing.

FC100 [57] is extracted from the CIFAR-100 dataset, with the training set of 60 categories, the validation set of 20 categories, and the test set of 20 categories.

B. Implementation Details

Follow the setting in [42], we adopt two widely-used backbones: ResNet-12 [59], [60] and ResNet-18 [5], [61].

However, we do not completely adopt the ResNet-18 set of [5], but modify it based on ResNet-12. Our ResNet-18 has four layers, and the first two layers each contains two residual blocks. There is only one residual block for each of the last two layers, and each residual block contains three convolution layers of 3×3 convolution kernels. Each convolution layer is followed by a batch normalization layer. Only after the first bath normalization layer, there is a ReLU nonlinear activation layer, and each residual block has a 2×2 max pooling layer at the end. In this setting, the input is of $3 \times 84 \times 84$ dimension and the output is of $640 \times 5 \times 5$ dimension.

The initial learning rate is set to 0.1. After every 400 epochs, the learning rate decreases by a factor of 10. The weight decay is set to 5e-4. We train ResNet-12 and ResNet-18 backbones in the 10-way 5-shot setting for 1,200 epochs. In addition, we verify the performance of the model every 20 epochs in training, preserving the best model parameters.

Table I
EVALUATION OF 5-WAY CLASSIFICATION ACCURACY ON FOUR FINE-GRAINED DATASETS USING THE RESNET-12 BACKBONE. WE REPRODUCED THE COMPARISON METHOD UNDER THE SAME SETTINGS AND DATASETS USING THEIR OPEN SOURCE CODE. ∗ DENOTES THE CLASSIFICATION ACCURACY ORIGINAL FROM THEIR ORIGINAL PAPER.

Model	CUB		Flowers		Cars		Aircraft	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet (NIPS-17) [39]	79.64 ± 0.20	91.15 ± 0.11	75.41 ± 0.22	89.46 ± 0.14	82.29 ± 0.20	93.11 ± 0.10	86.57 ± 0.18	93.51 ± 0.09
RelationNet (CVPR-18) [40]	63.94 ± 0.92	77.87 ± 0.64	69.51 ± 1.01	86.84 ± 0.56	69.67 ± 1.01	84.29 ± 0.68	74.20 ± 1.04	86.62 ± 0.55
Baseline++ (ICLR-19) [10]	64.62 ± 0.98	81.15 ± 0.61	69.03 ± 0.92	85.72 ± 0.63	67.92 ± 0.92	84.17 ± 0.58	74.51 ± 0.90	88.06 ± 0.44
DeepEMD (CVPR-20) [41]	71.11 ± 0.31	86.30 ± 0.19	70.00 ± 0.35	83.62 ± 0.26	73.30 ± 0.29	88.37 ± 0.17	69.86 ± 0.30	85.17 ± 0.28
MiXFSL (ICCV-21) [42]	67.87 ± 0.94	82.18 ± 0.66	72.60 ± 0.91	86.52 ± 0.65	58.15 ± 0.87	80.54 ± 0.63	60.55 ± 0.86	77.57 ± 0.69
VFD (ICCV-21) [14]	79.12 ± 0.83	91.48 ± 0.39	76.20 ± 0.92	89.90 ± 0.53	77.52 ± 0.85	90.76 ± 0.46	76.88 ± 0.85	88.77 ± 0.46
FRN (CVPR-21) [35]	83.16 ± 0.19	92.59 ± 0.11	81.07 ± 0.20	92.52 ± 0.11	86.48 ± 0.18	94.78 ± 0.08	87.53 ± 0.18	93.98 ± 0.09
RENet (ICCV-21) [43]	79.49 ± 0.44	91.11 ± 0.24	79.91 ± 0.42	92.33 ± 0.22	79.66 ± 0.44	91.95 ± 0.22	82.04 ± 0.41	90.50 ± 0.24
DeepBDC (CVPR-22) [44]	79.71 ± 0.44	92.54 ± 0.22	81.10 ± 0.49	93.25 ± 0.23	81.92 ± 0.40	96.12 ± 0.15	85.92 ± 0.41	94.62 ± 0.16
TDM (CVPR-22) [25]	82.41 ± 0.19	92.37 ± 0.10	82.85 ± 0.19	93.60 ± 0.10	86.91 ± 0.17	96.11 ± 0.07	88.35 ± 0.17	94.36 ± 0.08
HelixFormer (MM-22) [45]	81.66 ± 0.30	91.83 ± 0.17	-	-	79.40 ± 0.43	92.26 ± 0.15	-	-
BiFRN (AAAI-23) [36]	82.90 ± 0.19	93.11 ± 0.10	80.30 ± 0.20	92.30 ± 0.11	87.80 ± 0.16	96.49 ± 0.06	87.05 ± 0.18	93.78 ± 0.09
TFD* (TCSVT-23) [19]	84.08 ± 0.81	92.54 ± 0.39	-	-	-	-	-	-
BSFA (TCSVT-23) [26]	83.88 ± 0.44	90.76 ± 0.26	74.48 ± 0.54	86.05 ± 0.36	**88.93 ± 0.38**	95.20 ± 0.20	87.85 ± 0.35	94.93 ± 0.14
LCCRN (TCSVT-23) [37]	82.71 ± 0.19	93.48 ± 0.10	**84.12 ± 0.18**	**94.77 ± 0.09**	87.27 ± 0.18	96.01 ± 0.06	86.78 ± 0.18	95.09 ± 0.07
EFRN* (TCSVT-23) [38]	84.55 ± 0.19	93.46 ± 0.10	-	-	-	-	-	-
QSFormer* (TCSVT-23) [46]	75.44 ± 0.29	86.30 ± 0.19	-	-	-	-	-	-
IDEAL-clean (TPAMI-23) [47]	77.56 ± 0.86	88.87 ± 0.51	74.39 ± 0.93	87.29 ± 0.61	74.02 ± 0.89	89.98 ± 0.50	61.37 ± 0.92	82.51 ± 0.55
C2-Net (AAAI-24) [48]	83.37 ± 0.42	92.20 ± 0.23	80.86 ± 0.46	91.54 ± 0.27	84.81 ± 0.42	92.61 ± 0.23	87.98 ± 0.39	93.96 ± 0.20
Ours	**84.56 ± 0.18**	**94.21 ± 0.09**	83.52 ± 0.19	94.51 ± 0.09	87.51 ± 0.17	**97.11 ± 0.06**	**88.38 ± 0.16**	**95.10 ± 0.07**

Table II
EVALUATION OF 5-WAY CLASSIFICATION ACCURACY ON FOUR FINE-GRAINED DATASETS USING THE RESNET-18 BACKBONE. WE REPRODUCED THE COMPARISON METHOD UNDER THE SAME SETTINGS AND DATASETS USING THEIR OPEN SOURCE CODE. ∗ DENOTES THE CLASSIFICATION ACCURACY ORIGINAL FROM THEIR ORIGINAL PAPER.

Model	CUB		Flowers		Cars		Aircraft	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet (NIPS-16) [49]	72.88 ± 0.89	85.25 ± 0.57	76.07 ± 0.82	87.46 ± 0.51	75.03 ± 0.95	87.02± 0.56	82.84 ± 0.81	88.77 ± 0.54
ProtoNet (NIPS-17) [39]	78.20 ± 0.21	90.73 ± 0.11	75.82 ± 0.22	90.47 ± 0.13	84.16 ± 0.19	94.02 ± 0.09	85.81 ± 0.19	93.66 ± 0.09
Baseline++ (ICLR-19) [10]	65.67 ± 0.95	81.53 ± 0.58	67.90 ± 0.96	84.34 ± 0.62	67.41 ± 0.99	85.50 ± 0.58	75.92 ± 0.88	88.13 ± 0.47
Neg-margin (ECCV-20) [50]	72.51 ± 0.82	89.25 ± 0.43	76.34 ± 0.89	90.83 ± 0.47	76.04 ± 0.81	93.06 ± 0.38	77.40 ± 0.86	90.92 ± 0.39
FRN (CVPR-21) [35]	83.40 ± 0.19	92.69 ± 0.10	81.22 ± 0.21	92.33 ± 0.11	87.63 ± 0.17	95.35 ± 0.08	87.89 ± 0.18	93.96 ± 0.09
RENet (ICCV-21) [43]	77.14 ± 0.47	90.59 ± 0.27	76.81 ± 0.49	89.13 ± 0.30	80.33 ± 0.44	91.63 ± 0.23	82.95 ± 0.42	90.51 ± 0.23
DeepBDC (CVPR-22) [44]	83.65 ± 0.40	94.18 ± 0.17	80.65 ± 0.48	93.28 ± 0.24	85.57 ± 0.39	96.36 ± 0.15	87.45 ± 0.39	94.97 ± 0.15
TDM (CVPR-22) [25]	83.25 ± 0.19	92.98 ± 0.10	82.31 ± 0.20	93.46 ± 0.11	87.69 ± 0.17	96.06 ± 0.07	87.91 ± 0.17	94.28 ± 0.08
BiFRN (AAAI-23) [25]	82.86 ± 0.19	93.24 ± 0.10	80.44 ± 0.20	93.11 ± 0.10	88.29 ± 0.16	96.80 ± 0.06	87.73 ± 0.17	94.16 ± 0.09
LCCRN (TCSVT-23) [37]	82.74 ± 0.19	93.55 ± 0.10	83.58 ± 0.18	94.87 ± 0.08	86.24 ± 0.18	96.34 ± 0.07	86.95 ± 0.18	95.06 ± 0.07
QGN* (PR-23) [51]	83.82	91.22	-	89.9	-	91.3	-	92.0
Ours	**85.22 ± 0.18**	**94.47 ± 0.09**	**84.12 ± 0.18**	**94.91 ± 0.08**	**88.35 ± 0.16**	**97.15 ± 0.05**	**89.00 ± 0.16**	**95.41 ± 0.07**

In the test phase, we report the average classification accuracies with 95% confidence intervals of 10,000 randomly generated tasks on the test sets under the standard 5-way 1-shot and 5-way 5-shot settings.

C. Comparison with State-of-the-Art Methods

The classification accuracies of our method and the state-of-the-art methods using the ResNet-12 and ResNet-18 backbones are listed in Table I and Table II, respectively. We reproduce the results of all state-of-the-art methods with the same training settings using their open-source code. The proposed method achieves the best performance in most cases for CUB, Cars and Aircraft data using both backbones. For the Flowers dataset, although our method performs the second best using the ResNet-12 backbone in Table I, it is the best using the ResNet-18 backbone in Table II.

This can be ascribed to the fact that the proposed method uses the query samples to generate augmented support samples and models the diverse similarities between the two types of support samples and the query samples within the same sub-

class, alleviating the large intra-class variance and improving the classification accuracy.

To further verify the statistical significance of the superior performance of our method, we perform the one-tailed paired t-test to compare the 5-shot accuracies of our method with those of state-of-the-art methods in Tables I and II, and report the results in Table III. In this hypothesis test, we have null hypothesis $H_0 : \mu_{\text{Ours}} - \mu_* \leq 0$ and alternative hypothesis $H_1 : \mu_{\text{Ours}} - \mu_* > 0$, where μ is the mean accuracy of a method and $*$ denotes the state-of-the-art methods. When the p-value is less than 0.05, we reject the null hypothesis and conclude that our method is significantly better than the state-of-the-art methods. In Table III, ✓ denotes $p < 0.05$ while ✗ denotes $p \geq 0.05$. We can observe in both tables that our method is significantly better than most state-of-the-art methods, except for LCCRN on Flowers and Aircraft datasets.

Table III
THE p-VALUES OF THE ONE-TAILED PAIRED t-TEST ($H_1 : \mu_{\text{OURS}} - \mu_* > 0$), CALCULATED BASED ON THE 5-WAY 5-SHOT CLASSIFICATION ACCURACIES ON FOUR FINE-GRAINED DATASETS USING RESNET-12 BACKBONE IN TABLE I AND USING RESNET-18 BACKBONE IN TABLE II. THE SIGNIFICANCE LEVEL IS 0.05. NOTATION: "✓" : $p < 0.05$ AND "✗" : $p \geq 0.05$.

Dataset	In Table I: Ours *vs.* *					
	ProtoNet	FRN	BiFRN	BSFA	LCCRN	C2-Net
CUB	✓	✓	✓	✓	✓	✓
Flowers	✓	✓	✓	✓	✗	✓
Cars	✓	✓	✓	✓	✓	✓
Aircraft	✓	✓	✓	✓	✗	✓

Dataset	In Table II: Ours *vs.* *			
	ProtoNet	FRN	BiFRN	LCCRN
CUB	✓	✓	✓	✓
Flowers	✓	✓	✓	✗
Cars	✓	✓	✓	✓
Aircraft	✓	✓	✓	✓

D. Ablation Studies

1) The impact of different branches of reconstruction: Table IV shows the impact of the two branches and the inner four reconstructions on classification performance. The first row represents the results that reconstruct the query and support feature maps by only using the original support features (SCM), the second row shows the results that reconstruct the query and support feature maps by only using the augmented samples (ASCM), and the third row is the proposed method including both SCM and ASCM. It is clear that the best performance is reached when both branches are used. The middle row of Table IV shows the case of eliminating all support self-reconstruction, and the results also verify that the support self-reconstruction strategy improves the model's classification performance. The lower part of Table IV displays the classification accuracies of removing one of the four reconstructions, and the results show that using all four achieves the best performance.

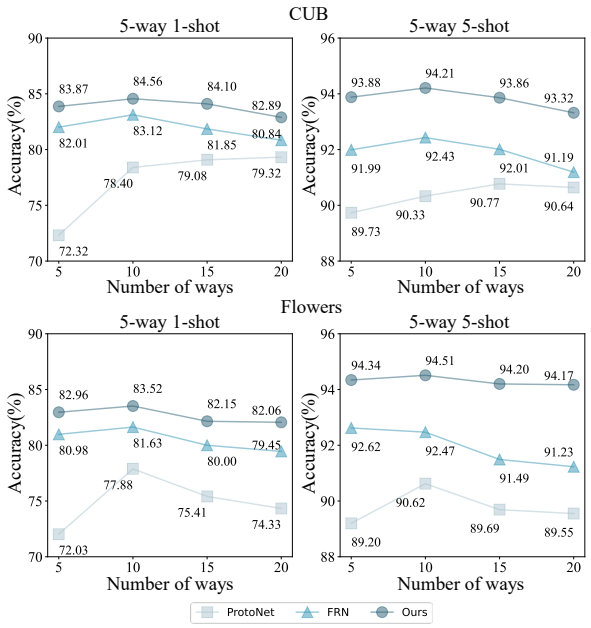


Figure 3. The effect of the number of ways on classification accuracy. We employ the C-way 5-shot training approach and evaluate using the 5-way 1-shot and 5-way 5-shot settings on CUB and Flowers with the ResNet-12 backbone.

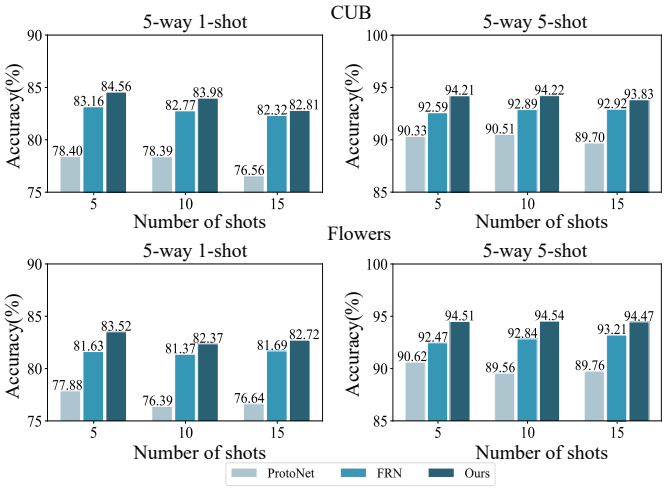


Figure 4. The effect of the number of shots on classification accuracy. We employ the 10-way K-shot training approach and evaluate using the 5-way 1-shot and 5-way 5-shot settings on CUB and Flowers with the ResNet-12 backbone.

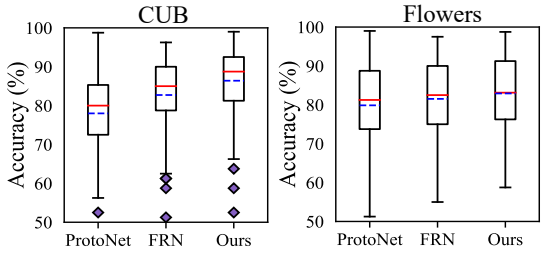


Figure 5. The boxplots of the test classification accuracies of ProtoNet, FRN and our method (Ours) across 100 randomly sampled tasks. The models are trained by the 10-way 5-shot setting and evaluated by 5-way 1-shot test tasks.

Table IV
ABLATION STUDY OF DIFFERENT RECONSTRUCTION TASKS ON THREE FINE-GRAINED DATASETS USING THE 5-WAY SETTING AND RESNET-12 BACKBONE.

SCM		ASCM		CUB		Flowers		Cars	
$S_c \to \tilde{Q}_c$	$S_c \to \tilde{S}_c$	$S_c^+ \to \tilde{Q}_c^+$	$S_c^+ \to \tilde{S}_c^+$	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
✓		✗		83.50 ± 0.19	93.19 ± 0.10	83.14 ± 0.19	94.06 ± 0.10	86.48 ± 0.18	95.85 ± 0.07
✗		✓		83.98 ± 0.18	93.76 ± 0.09	82.20 ± 0.20	93.61 ± 0.10	85.71 ± 0.18	95.89 ± 0.07
✓		✓		84.56 ± 0.18	**94.21 ± 0.09**	**83.52 ± 0.19**	**94.51 ± 0.09**	**87.51 ± 0.17**	**96.58 ± 0.07**
✓	✗	✓	✗	83.69 ± 0.19	94.08 ± 0.10	82.11 ± 0.20	93.48 ± 0.10	87.34 ± 0.17	96.51 ±0.06
✗	✓	✓	✓	**84.73 ± 0.18**	94.14 ± 0.09	82.81 ± 0.20	94.03 ± 0.10	86.62 ± 0.18	95.99 ± 0.07
✓	✗	✓	✓	84.23 ± 0.18	94.02 ± 0.09	82.63 ± 0.19	94.15 ± 0.09	86.42 ± 0.18	96.43 ± 0.07
✓	✓	✗	✓	84.19 ± 0.18	93.76 ± 0.09	82.66 ± 0.20	93.93 ± 0.10	86.13 ± 0.18	95.58 ± 0.08
✓	✓	✓	✗	84.34 ± 0.18	94.08 ± 0.09	83.05 ± 0.19	94.21 ± 0.09	86.07 ± 0.18	96.39 ± 0.07

Table V
ABLATION STUDY OF DIFFERENT DATA AUGMENTATION METHODS ON THREE FINE-GRAINED DATASETS USING THE 5-WAY SETTING AND RESNET-12 BACKBONE.

	CUB		Flowers		Cars	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Ours w/ Cutout [62]	83.06 ± 0.19	92.89 ± 0.10	82.52 ± 0.19	94.09 ± 0.09	87.29 ± 0.18	95.77 ± 0.07
Ours w/ Mixup [63]	82.88 ± 0.19	92.90 ± 0.10	81.04 ± 0.20	92.40 ± 0.11	85.60 ± 0.18	95.10 ± 0.08
Ours w/ Cutmix (query-aware) [64]	84.22 ± 0.18	93.82 ± 0.09	83.26 ± 0.19	94.62 ± 0.09	86.93 ± 0.18	96.56 ± 0.06
Ours w/ Cross-mixup	**84.56 ± 0.18**	**94.21 ± 0.09**	**83.52 ± 0.19**	**94.51 ± 0.09**	**87.51 ± 0.17**	**96.58 ± 0.07**

Table VI
CLASSIFICATION ACCURACY OF CROSS-DOMAIN TASKS USING TWO DIFFERENT BACKBONES.

Training ⟶Testing	Method	Backbone	1-shot	5-shot
Flowers ⟶CUB	ProtoNet [39]	ResNet-12	40.16 ± 0.19	56.22 ± 0.19
	FRN [35]		43.34 ± 0.20	58.34 ± 0.20
	LCCRN [37]		44.12 ± 0.18	62.91 ± 0.19
	Ours		**47.30 ± 0.20**	**65.27 ± 0.19**
	ProtoNet [39]	ResNet-18	39.64 ± 0.19	56.58 ± 0.19
	FRN [35]		45.72 ± 0.21	60.18 ± 0.20
	LCCRN [37]		44.53 ± 0.19	63.42 ± 0.19
	Ours		**47.41 ± 0.20**	**64.37 ± 0.19**
Cars ⟶Aircraft	ProtoNet [39]	ResNet-12	29.98 ± 0.14	45.13 ± 0.15
	FRN [35]		30.87 ± 0.14	40.04 ± 0.14
	LCCRN [37]		40.07 ± 0.16	59.76 ± 0.16
	Ours		**41.86 ± 0.17**	**62.38 ± 0.16**
	ProtoNet [39]	ResNet-18	29.85 ± 0.14	43.71 ± 0.15
	FRN [35]		30.01 ± 0.14	38.84 ± 0.14
	LCCRN [37]		33.14 ± 0.14	46.57 ± 0.14
	Ours		**40.46 ± 0.16**	**60.41 ± 0.16**

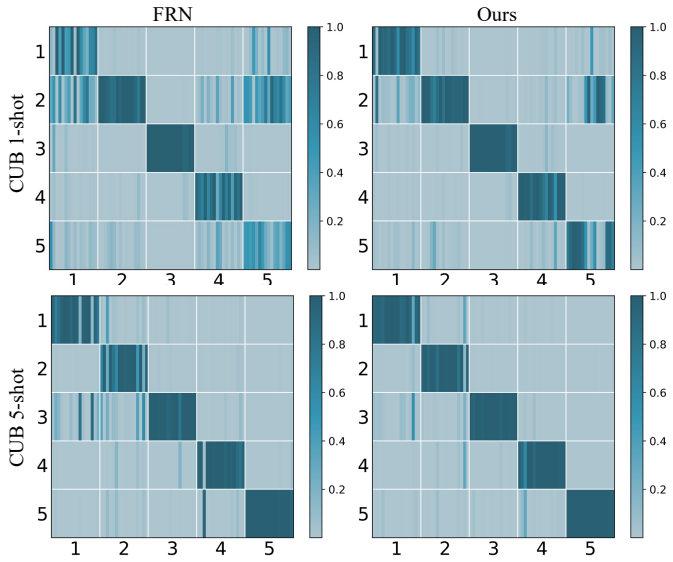


Figure 6. Visualization of the predicted probabilities of FRN and the proposed method (Ours) on the CUB dataset. In the confusion matrices, each block contains 16 bars representing the predicted probabilities of 16 randomly selected test query images. The darker the bars, the higher the predicted probability.

2) The impact of different data augmentation methods: In Table V, we compare the proposed cross-mixup augmentation method with three related augmentation methods, cutout [62], mixup [63] and cutmix [64]. Cutout randomly masks out squared regions of images, mixup creates augmented samples by convex combinations of training samples, while cutmix cuts and pastes patches in training samples. In this experiment, cutout and mixup are applied to the same classes of the support set only. Cutout is revised to fit the image size of 84×84, and crop is performed with the 0.25 ratio of the width of the image. Cutmix is performed to cut and paste the patches of query samples with support samples from the same classes, which makes the method query-aware.

Clearly, our proposed cross-mixup is the best augmentation

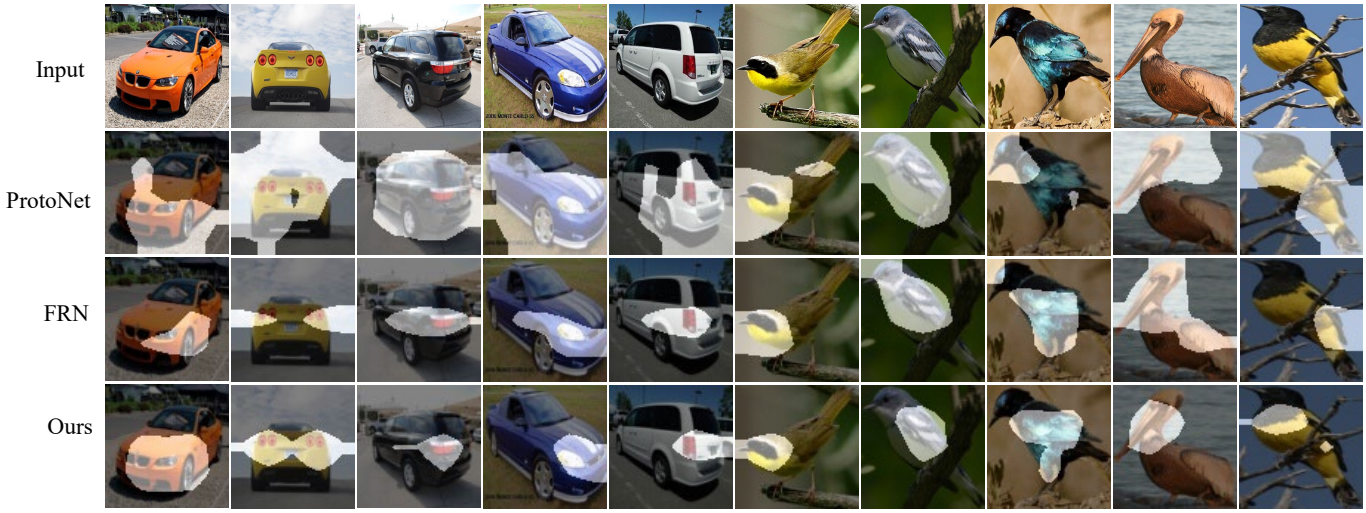


Figure 7. Visualization of the discriminative regions captured by ProtoNet, FRN and our method (Ours). Our method can identify the most delicate and discriminative regions to classify fine-grained classes.

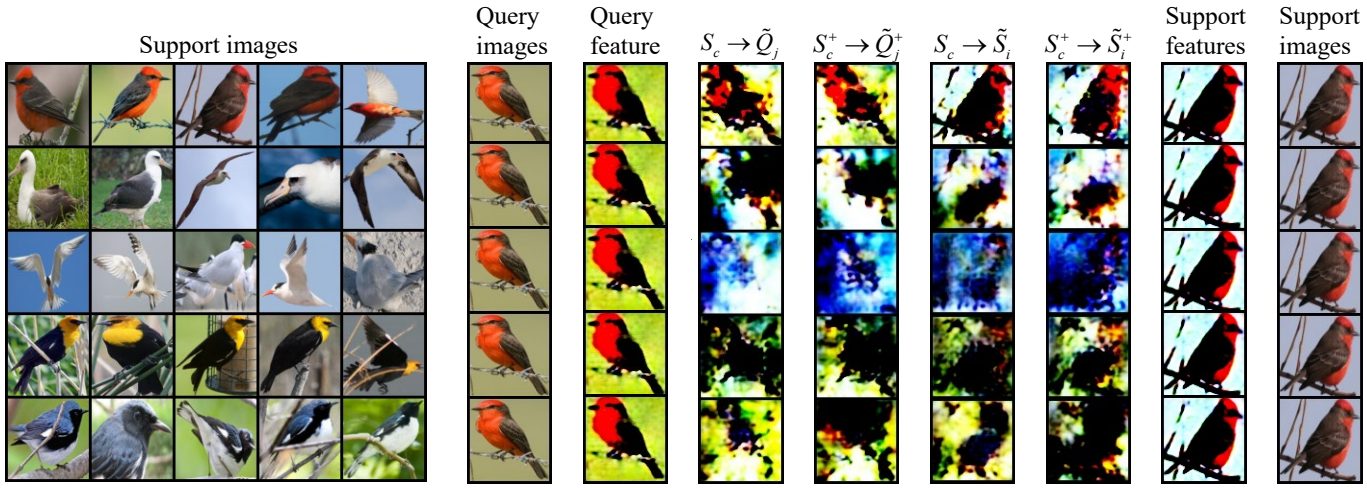


Figure 8. Visualization of the reconstructed query and support images from the CUB dataset of the four reconstruction tasks in our method. The reconstructed images from the SCM and ASCM branches can provide complementary details.

method. Compared with mixup, cross-mixup substantially increases the classification accuracies of all datasets, demonstrating the importance of query-awareness. In addition, higher accuracies of cross-mixup over query-aware cutmix suggest that the linear mixture strategy is better than the simple cut-and-paste strategy for fine-grained image classification.

3) The impact of the numbers of ways and shots: In order to further analyze the impact of the numbers of ways and shots on the model performance, we compare the test classification accuracies of ProtoNet, FRN and our method in Figures 3 and 4, respectively. The results of two test settings, 5-way 1-shot and 5-shot, are reported for the CUB and Flowers data.

In Figure 3, when the number of ways increases from 5 to 10, the overall classification performance shows an upward trend. However, when the number of ways exceeds 10, the classification accuracy of FRN and our method declines. In Figure 4, 5 shots tend to provide the best classification accuracy. Nonetheless, our method is superior over the other two methods for all number of ways and shots.

E. Distribution of Classification Accuracy

In Figure 5, we show the boxplots of the classification accuracies of ProtoNet [39], FRN [35] and our method for 100 randomly selected test tasks. The three models are trained by the 10-way 5-shot setting and evaluated by 5-way 1-shot test tasks. In the boxplots, the red line is the median and the blue dotted is the mean. We can clearly observe that our method performs better than the other two methods with higher means and medians.

F. Cross-Domain Performance

In Table VI, we further compare the performance of our method with the most relevant methods, FRN [35] and LC-CRN [37], on two cross-domain tasks, where the models

Table VII
EVALUATION OF 5-WAY CLASSIFICATION ACCURACY ON TREE COARSE-GRAINED DATASETS USING THE RESNET-12 BACKBONE.

	mini-ImageNet		tiered-ImageNet		FC-100	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
FRN [35]	**66.40 ± 0.19**	83.43 ± 0.13	**70.93 ± 0.22**	**85.70 ± 0.15**	41.05 ± 0.18	55.52 ± 0.18
Ours	63.90 ± 0.19	**84.08 ± 0.12**	70.52 ± 0.22	84.97 ± 0.15	**41.59 ± 0.18**	**57.09 ± 0.18**

are trained and test on two different datasets. Our method can achieve the best classification accuracies for all cases, demonstrating its superior generalization ability.

G. Qualitative Analysis via Visualization

1) Visualizing the predicted probabilities: In Figure 6, we provide the visualization of the predicted probabilities for FRN and the proposed method in the training settings of 5-way 1-shot and 5-way 5-shot on the CUB dataset. The predicted probabilities are calculated according to Eq.(12) on 16 randomly sampled query images for each class. In each plot, the diagonal blocks represent the probabilities of the correct prediction while the off-diagonal blocks represent those of the wrong predictions. The darker the bars, the higher the predicted probability. It is obvious that our method can provide more correct predictions on the diagonals and less wrong predictions in the off-diagonal blocks.

2) Visualizing the discriminative regions: The discriminative regions captured by ProtoNet, FRN and our method are visualized in Figure 7. ProtoNet tends to include most of the object and irrelevant background as discriminative features, while FRN can focus more on the targets with less background involved. Our method can provide the most delicate discriminative regions; for example, the head and rear lights of cars and the heads, beaks and wings of birds.

3) Visualizing the reconstructed images of the four reconstruction tasks: In Figure 8, we visualize the reconstructed images obtained by the four reconstruction tasks, and the following two conclusions can be drawn. First, the images reconstructed by images from the same class are better than those reconstructed by images from different classes, showing evidence of using the reconstruction error as a metric for classification. Second, the reconstruction by augmented support features complement the details ignored by the original support feature reconstruction, e.g., the branch of $S_c^+ \rightarrow \tilde{Q}_c^+$ reconstructs the query feature better.

H. Performance on Coarse-grained Datasets

In Table VII, we evaluate the classification accuracies of our method against FRN on three coarse-grained datasets. Both methods are trained using the 10-way 5-shot setting and evaluated in 5-way 1-shot and 5-way 5-shot scenarios. While our method outperforms FRN on the FC-100 dataset, it performs worse or comparably to FRN on mini-ImageNet and tiered-ImageNet, unlike its superior performance on fine-grained datasets. Coarse-grained data normally have more diverse scenes and coarse-grained classes than fine-grained data. However, while our cross-mixup can enhance subtle discriminative regions between fine-grained sub-classes, it may also mix up sub-classes within a coarse-grained class while creating the cross-mixed support samples during the training phase, hence falls short in coarse-grained tasks during the test phase that only original support samples can be used.

I. Evaluation of Model Efficiency

Table VIII
COMPARISON OF MODEL EFFICIENCY.

Method	FLOPs (G)	Params (K)
FRN [35]	1127.36	12424.32
RENet [43]	1469.49	12659.53
TDM [25]	1409.20	12424.32
BiFRN [36]	1446.14	16116.48
LCCRN [37]	2832.85	25005.95
C2-NET [48]	1440.81	18486.09
Ours	1761.50	12424.32

We compare the model efficiency in a 10-way 5-shot setting and use THOP to obtain the FLOPs and parameters for each model. As shown in Table VIII, our method has fewer parameters, but it has more FLOPs due to the augmented support samples for cross-reconstruction and self-reconstruction.

V. CONCLUSION

In this paper, we propose a data augmentation method, called query-aware cross-mixup. Unlike traditional mix-up methods that combine samples from different classes in the support set, the proposed method randomly selects samples from the query set and mixes them with support samples from the same class, to augment the support set and encourages the model to learn fine-grained feature representation. In addition, we develop a strategy to leverage both cross-reconstruction and self-reconstruction to mitigate the bias between support and query samples for a better generalization. Extensive experiment results on four widely used few-shot fine-grained image datasets demonstrate the superior classification performance of the proposed method to the state-of-the-art methods.

We note two limitations of our method. First, as indicated in Table VIII, our method has relatively high FLOPs. This suggests some room in computational efficiency for our method to improve. Second, as shown in Table VII, our method, particularly designed for fine-grained datasets, does not perform so superior on coarse-grained datasets as on fine-grained datasets. This suggests a comprehensive extension of our method to both fine-grained and coarse-grained data. It is our future work to address these two limitations.

ACKNOWLEDGEMENT

This work was partly supported by the National Nature Science Foundation of China (Grants 62176110, 62463015, 62225601,U23B2052,62406171), the Key Research and Development Program of Gansu Province, China under Grant 22YF7GA130, S&T Program of Hebei, China under Grant SZX2020034, Hong-Liu Distinguished Young Talents Foundation of Lanzhou University of Technology, in part by the Beijing Natural Science Foundation Project under Grant L242025, in part by the Youth Innovative Research Team of BUPT under Grant 2023YQTD02, in part by the China Postdoctoral Science Foundation No. 2023M741961, and in part by the Postdoctoral Fellowship Program of CPSF No. GZB20240359, and the Royal Society under International Exchanges Award IEC\NSFC\201071.

REFERENCES

[1] Shu-Lin Xu, Faen Zhang, Xiu-Shen Wei, and Jianhua Wang. Dual attention networks for few-shot fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pages 2911–2919, 2022.

[2] Ruyi Ji, Jiaying Li, Libo Zhang, Jing Liu, and Yanjun Wu. Dual transformer with multi-grained assembly for fine-grained visual classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):5009–5021, 2023.

[3] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo 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 (CVPR)*, pages 7260–7268, 2019.

[4] Huaxi Huang, Junjie Zhang, Jian Zhang, Jingsong Xu, and Qiang Wu. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *IEEE Transactions on Multimedia*, 23:1666–1680, 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[6] Jinhui Tang, Xiangbo Shu, Guo-Jun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1662–1674, 2016.

[7] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Zechao Li, Yu-Gang Jiang, and Shuicheng Yan. Image classification with tailored fine-grained dictionaries. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(2):454–467, 2016.

[8] Meng Pang, Yiu-Ming Cheung, Risheng Liu, Jian Lou, and Chuang Lin. Toward efficient image representation: Sparse concept discriminant matrix factorization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3184–3198, 2018.

[9] Jinhui Tang, Xiangbo Shu, Zechao Li, Yu-Gang Jiang, and Qi Tian. Social anchor-unit graph regularized tensor completion for large-scale image retagging. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2027–2034, 2019.

[10] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representation (ICLR)*, 2019.

[11] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9062–9071, 2021.

[12] Shell Xu Hu, Da Li, Jan Stuhmer, Minyoung Kim, and Timothy M. Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9077, 2022.

[13] Meng Pang, Binghui Wang, Mang Ye, Yiu-Ming Cheung, Yintao Zhou, Wei Huang, and Bihan Wen. Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2024.

[14] Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8812–8821, 2021.

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[16] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13470–13479, 2020.

[17] Jiabao Zhao, Xin Lin, Jie Zhou, Jing Yang, Liang He, and Zhaohui Yang. Knowledge-based fine-grained classification for few-shot learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[18] Chia-Ching Lin, Hsin-Li Chu, Yu-Chiang Frank Wang, and Chin-Laung Lei. Joint feature disentanglement and hallucination for few-shot image classification. *IEEE Transactions on Image Processing*, 30:9245–9258, 2021.

[19] Zixuan Hu, Li Shen, Shenqi Lai, and Chun Yuan. Task-adaptive feature disentanglement and hallucination for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3638–3648, 2023.

[20] Shuai Shao, Yan Wang, Bin Liu, Weifeng Liu, Yanjiang Wang, and Baodi Liu. Fads: Fourier-augmentation based data-shunting for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):839–851, 2024.

[21] Huaxi Huang, Junjie Zhang, Jian Zhang, Qiang Wu, and Jingsong Xu. Compare more nuanced: Pairwise alignment bilinear network for few-shot fine-grained learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 91–96, 2019.

[22] Xin Sun, Hongwei Xv, Junyu Dong, Huiyu Zhou, Changrui Chen, and Qiong Li. Few-shot learning for domain-specific fine-grained image classification. *IEEE Transactions on Industrial Electronics*, 68(4):3588–3598, 2020.

[23] Xiaoxu Li, Jijie Wu, Zhuo Sun, Zhanyu Ma, Jie Cao, and Jing-Hao Xue. Bsnet: Bi-similarity network for few-shot fine-grained image classification. *IEEE Transactions on Image Processing*, 30:1318–1331, 2020.

[24] Huaxi Huang, Junjie Zhang, Litao Yu, Jian Zhang, Qiang Wu, and Chang Xu. Toan: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):853–866, 2021.

[25] SuBeen Lee, WonJun Moon, and Jae-Pil Heo. Task discrepancy maximization for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5331–5340, 2022.

[26] Zican Zha, Hao Tang, Yunlian Sun, and Jinhui Tang. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3947–3961, 2023.

[27] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2019.

[28] Xiaofeng Zhang, Zhangyang Wang, Dong Liu, Qifeng Lin, and Qing Ling. Deep adversarial data augmentation for extremely low data regimes. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):15–28, 2020.

[29] Weiqiu Wang, Zhicheng Zhao, Pingyu Wang, Fei Su, and Hongying Meng. Attentive feature augmentation for long-tailed visual recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5803–5816, 2022.

[30] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7278–7286, 2018.

[31] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8059–8068, 2019.

[32] Cheng Perng Phoo and Bharath Hariharan. Coarsely-labeled data for better few-shot transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9052–9061, 2021.

[33] Min Zhang, Siteng Huang, Wenbin Li, and Donglin Wang. Tree structure-aware few-shot image classification via hierarchical aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–470, 2022.

[34] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-assisted learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10573–10582, 2021.

[35] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] Jijie Wu, Dongliang Chang, Aneeshan Sain, Xiaoxu Li, Zhanyu Ma, Jie Cao, Jun Guo, and Yi-Zhe Song. Bi-directional feature reconstruction network for fine-grained few-shot image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 2821–2829, 2023.

[37] Xiaoxu Li, Qi Song, Jijie Wu, Rui Zhu, Zhanyu Ma, and Jing-Hao Xue. Locally-enriched cross-reconstruction for few-shot fine-grained image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7530–7540, 2023.

[38] Jiaxing Sun, Xiaobo Shen, and Quansen Sun. Efficient feature reconstruction via l 2, 1-norm regularization for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7452–7465, 2023.

[39] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the Advances in neural information processing systems (NIPS)*, volume 30, 2017.

[40] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208, 2018.

[41] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua 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 (CVPR)*, pages 12203–12213, 2020.

[42] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Mixture-based feature space learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9041–9051, 2021.

[43] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8822–8833, 2021.

[44] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua 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 (CVPR)*, pages 7972–7981, 2022.

[45] Bo Zhang, Jiakang Yuan, Baopu Li, Tao Chen, Jiayuan Fan, and Botian Shi. Learning cross-image object semantic relation in transformer for few-shot fine-grained image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2135–2144, 2022.

[46] Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7789–7802, 2023.

[47] Yuexuan An, Hui Xue, Xingyu Zhao, and Jing Wang. From instance to metric calibration: A unified framework for open-world few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9757–9773, 2023.

[48] Zhen-Xiang Ma, Zhen-Duo Chen, Li-Jun Zhao, Zi-Chao Zhang, Xin Luo, and Xin-Shun 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 (AAAI)*, volume 38, pages 4136–4144, 2024.

[49] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the Advances in neural information processing systems (NIPS)*, volume 29, 2016.

[50] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 438–455, 2020.

[51] Bharti Munjal, Alessandro Flaborea, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided networks for few-shot fine-grained classification and person search. *Pattern Recognition*, 133:109049, 2023.

[52] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[53] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[54] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 554–561, 2013.

[55] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013.

[56] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *ICCR*, 2018.

[57] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Proceedings of the Advances in neural information processing systems (NIPS)*, volume 31, 2018.

[58] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representation (ICLR)*, 2017.

[59] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 266–282, 2020.

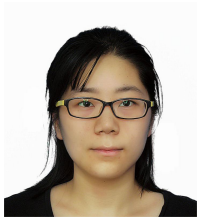
[60] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10657–10665, 2019.

[61] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei 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 (CVPR)*, pages 8808–8817, 2020.

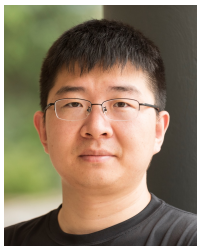
[62] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552, 2017.

[63] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017.

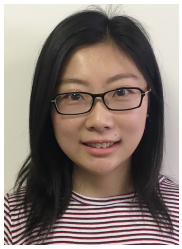
[64] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.



Zhimin Zhang received her B.E. degree in Things of Internet Engineering from Anyang Normal University, China, in 2017. She is currently pursuing the Ph.D. degree in Lanzhou University of Technology. Her research interests include machine learning and few-shot learning.



Dongliang Chang is currently a Postdoctoral Research Fellow in the Department of Automation at Tsinghua University, Beijing, China, starting in 2023. He received his Ph.D. degree in Information and Communication Engineering from Beijing University of Posts and Telecommunication, China, in 2023. His research interest lies at the intersection of deep learning and computer vision, with a specific focus on fine-grained visual understanding.



Rui Zhu received the Ph.D. degree in statistics from University College London in 2017. She is a Senior Lecturer in the Faculty of Actuarial Science and Insurance, City, University of London. Her research interests include machine learning and its applications in image quality assessment, hyperspectral image analysis and actuarial science.

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

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 In stitute 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.

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.