Journal Pre-proof

A simple scheme to amplify inter-class discrepancy for improving few-shot fine-grained image classification

Xiaoxu Li, Zijie Guo, Rui Zhu, Zhayu Ma, Jun Guo, Jing-Hao Xue

PII: S0031-3203(24)00487-4
DOI: https://doi.org/10.1016/j.patcog.2024.110736
Reference: PR 110736

To appear in: *Pattern Recognition*

Received date : 22 October 2023
Revised date : 12 May 2024
Accepted date : 26 June 2024

Please cite this article as: X. Li, Z. Guo, R. Zhu et al., A simple scheme to amplify inter-class discrepancy for improving few-shot fine-grained image classification, *Pattern Recognition* (2024), doi: https://doi.org/10.1016/j.patcog.2024.110736.

A simple scheme to amplify inter-class discrepancy for improving few-shot fine-grained image classification

Xiaoxu Li[a], Zijie Guo[a], Rui Zhu[b,*], Zhayu Ma[c], Jun Guo[c], Jing-Hao Xue[d]

[a]*School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China.*
[b]*Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, London EC1Y 8TZ, UK*
[c]*Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.*
[d]*Department of Statistical Science, University College London, London WC1E 6BT, UK*

Abstract

Few-shot image classification is a challenging topic in pattern recognition and computer vision. Few-shot fine-grained image classification is even more challenging, due to not only the few shots of labelled samples but also the subtle differences to distinguish subcategories in fine-grained images. A recent method called task discrepancy maximisation (TDM) can be embedded into the feature map reconstruction network (FRN) to generate discriminative features, by preserving the appearance details through reconstructing the query image and then assigning higher weights to more discriminative channels, producing the state-of-the-art performance for few-shot fine-grained image classification. However, due to the small inter-class discrepancy in fine-grained images and the small training set in few-shot learn-

*Corresponding author. Tel.: +44(0)1227 827008

Email addresses: `lixiaoxu@lut.edu.cn` (Xiaoxu Li), `gzj18801586376@163.com` (Zijie Guo), `rui.zhu@city.ac.uk` (Rui Zhu), `mazhanyu@bupt.edu.cn` (Zhayu Ma), `guojun@bupt.edu.cn` (Jun Guo), `jinghao.xue@ucl.ac.uk` (Jing-Hao Xue)

ing, the training of FRN+TDM can result in excessively flexible boundaries between subcategories and hence overfitting. To resolve this problem, we propose a simple scheme to amplify inter-class discrepancy and thus improve FRN+TDM. To achieve this aim, instead of developing new modules, our scheme only involves two simple amendments to FRN+TDM: relaxing the inter-class score in TDM, and adding a centre loss to FRN. Extensive experiments on five benchmark datasets showcase that, although embarrassingly simple, our scheme is quite effective to improve the performance of few-shot fine-grained image classification. The code is available at `https://github.com/Airgods/AFRN.git`.

Keywords: Few-shot learning, fine-grained image classification, metric-based methods.

---

1. Introduction

Few-shot fine-grained image classification is a challenging task that draws wide attention in the pattern recognition and computer vision communities. Although deep neural networks learnt from a large amount of labelled training data can provide impressive image classification performances, few-shot learning that trains a model with little labelled data for each class remains difficult. Moreover, the fine-grained setting brings further challenges, as each class is divided to a large number of subcategories, which makes the inter-class discrepancy even smaller and the classification task much harder.

Metric-based methods are effective for few-shot learning [1]. They aim to learn a metric function to measure the similarities/dissimilarities between different classes and assign the test image to the class with the highest similarity or lowest dissimilarity. For example, the prototypical networks (ProtoNet) proposed by Snell et al. [2] adopt the average of features of all images from the same class in the support set as the prototype of that class, and assign the query image to the class with the shortest Euclidean distances from the class prototypes. Recent works enhance ProtoNet by generating more representative prototypes [3]. The matching networks (MatchingNet) [4] utilise a bidirectional LSTM network to map the support set and an attention mechanism-based LSTM to map the query set, and adopt the cosine similarity as the metric function. In addition to the common metric functions, Zhang et al. [5] propose a new metric function EMD, which assigns different weights to different positions of the image and calculates the best matching between the image blocks of the support set and the query set to represent their similarities. To maintain feature discriminability, Nguyen et al. [6] propose the square root of the sum of the Euclidean distance and the norm distance as the metric function. Similarities between images can also be measured via a properly structured neural network [7].

However, when the high similarities between subclasses are not carefully considered, metric-based methods can fail to classify fine-grained images. Thus it is crucial to extract features with strong discriminative power to distinguish the ultra-fine differences between subclasses. Li et al. introduce the bi-similarity network (BSNet) with two similarity metrics to learn such discriminative features [8]. Huang et al. propose the low-rank pairwise aligned bilinear network (LRPABN), which utilises bilinear pooling operations to distinguish support and query images [9]. Huang et al. also propose the targeted alignment network (TOAN), which can increase the inter-class variation by extracting discriminative fine-grained features while reducing intra-class variation by matching support and query features [10].

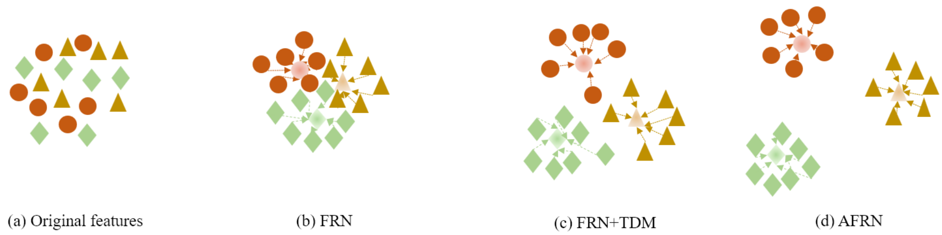

Figure 1: An illustration of the motivation of the adaptive feature map reconstruction network (AFRN). The solid circles, triangles and diamonds represent the instances from three classes, respectively, and the transparent circle, triangle and diamond represent the corresponding prototypes of the three classes, respectively. In (a), we depict a challenging classification task, with severe overlapping between the three classes in the original features space. This challenge is partially resolved by FRN in (b), because the appearance details of images are well preserved by reconstruction, which potentially makes the embedded features more discriminative. In (c), TDM is incorporated to FRN to assign high weights to channels with strong discriminative abilities, and thus the classes become more separable. Finally, in (d), AFRN further improves FRN+TDM by amplifying the inter-class discrepancy, and thus the three classes can be more easily distinguished.

There is a problem in many previous metric-based learning algorithms that the input to the metric function has to be reshaped to vectors, resulting in deficient spatial information. To resolve this problem, Wertheimer et al. [11] propose a novel metric-based classification mechanism, feature map reconstruction networks (FRN), for few-shot learning. FRN predicts the membership of the query image by reconstructing the query feature map via the pooled support features of each class. The idea behind FRN is that the query feature map is expected to be well reconstructed by the support

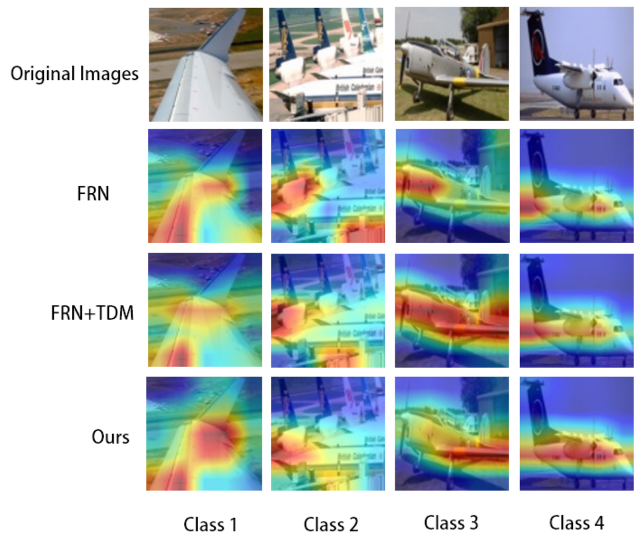

Figure 2: Examples of the features captured by FRN, FRN+TDM and AFRN on four subclasses of airplanes. Apparently, FRN focuses on the objects as well as the nuisance background. Involving TDM in FRN makes the features more discriminative and the focus on background is reduced slightly. In comparison, our AFRN can identify the most discriminative features to distinguish the subclasses with the least focus on the background.

features from the correct class with the smallest reconstruction error. Hence, through the reconstruction process, FRN can well preserve the appearance details of the images.

However, in FRN, all channels are treated equally with the same weights, without stressing the different importance of different channels. Hence, Lee et al. [12] propose the task discrepancy maximisation (TDM) module to identify channels with high discriminative power and assign higher weights to these channels to improve the classification results of few-shot methods, such as FRN, for fine-grained images. TDM produces channel weights for both support and query sets via the support attention module (SAM) and the query attention module (QAM), respectively. SAM provides class-wise channel weights to highlight the discriminative channels to distinguish between classes, while QAM provides object-wise channel weights to focus more on the object-relevant channels. Lee et al. [12] demonstrate that by incorporating TDM to FRN, namely FRN+TDM, a state-of-the-art performance of few-shot fine-grained image classification can be achieved.

However, due to the small inter-class discrepancy omnipresent in fine-grained images and the small training set in the setting of few-shot learning, FRN+TDM can produce excessively flexible boundaries between subcategories and hence overfitting. To resolve this problem, we propose a simple scheme to amplify inter-class discrepancy and thus improve FRN+TDM. To this end, instead of developing new modules to further enhance the extraction of discriminative features, our scheme only involves two simple amendments to FRN+TDM: relaxing the inter-class score in TDM, and adding a centre loss to FRN. We name the network incorporating our scheme to FRN+TDM the adaptive feature map reconstruction network (AFRN).

The centre loss [13] aims to achieve intra-class compactness by penalising the distance between the learnt features and their corresponding class centres, which is vital to distinguish subclasses with high similarity normally occurring in fine-grained image classification. In Figure 1, we illustrate the motivation of AFRN by a challenging classification of three overlapping classes, which is typical in fine-grained image classification with small inter-class discrepancy. By involving the centre loss in AFRN, we expect that the three classes can be intra-class more compact and thus inter-class more separated to make the classification easier. Moreover, in Figure 2, we demonstrate one real-data example of the discriminative features extracted by FRN, FRN+TDM and AFRN on four subclasses of airplanes. The original FRN focuses on the airplanes as well as the nuisance backgrounds; incorporating TDM can improve this situation with less focus on the backgrounds; while, in comparison, AFRN can identify the most discriminative features with the least focus on the backgrounds. For instance, in class 2, the background in the lower right corner is least highlighted in our method.

More importantly, we observe that FRN+TDM can produce excessively flexible boundaries between subcategories and thus overfitting, as the inter-class score in TDM to measure the discrepancy between classes is the Euclidean distance between one class and its *nearest* class. Such an inter-class score can result in extremely flexible classification boundaries for fine-grained images and thus overfitting to the seen classes in the training set. In few-shot fine-grained learning, this problem is severer, because in the test phase, few-shot learning aims to classify the novel set with completely different classes from those in the training set. Thus we propose to relax the inter-class score in TDM simply to the Euclidean distance between one class and its *furthest* class, to mitigate the potential overfitting to a large extent. This amendment makes the original TDM module a relaxed TDM module.

In summary, the main contributions of our work are as follows.

- We propose AFRN, a simple scheme to amplify inter-class discrepancy and thus improve the few-shot fine-grained image classification. Our scheme only involves two simple amendments to FRN+TDM: relaxing the inter-class score in TDM, and adding a centre loss to FRN.
- By relaxing the inter-class score in TDM, we are able to remarkably mitigate the negative impact, from the overfitting to the seen training set of fine-grained subclasses, on the inference of unseen novel classes in the few-shot learning setting.
- By incorporating the guidance of the centre loss to FRN, we are able to enhance the discriminative power of the learnt features for fine-grained image classification, through enlarging the omnipresent subtle distances between fine-grained subclasses.
- The experiments on five benchmark fine-grained datasets demonstrate that our scheme, although very simple, is quite effective to improve the performance of few-shot fine-grained image classification.

The rest of the paper is organised as follows. In section 2, we discuss the literature that is closely related to our work. The technical details of FRN+TDM and AFRN are presented in section 3. In section 4, we demonstrate the superior classification performances of AFRN through extensive experimental results and ablation studies. Lastly, we draw conclusions in section 5.

2. Related Work

2.1. Metric-based few-shot methods for image classification

Metric-based few-shot methods aim to learn discriminative feature embeddings that can be well generalized to new classes based on a predefined or a learnt distance metric, such as Euclidean distance [2], cosine distance [14], hyperbolic distance [15], or distance parameterized by neural networks [16]. MatchingNet [4] adopts the cosine similarity to assign the label of the query image. ProtoNet [2] calculates prototypes as the average features of each class in the support set and assign the query image to the nearest class prototype by Euclidean distance. Instead of using a predefined metric, RelationNet [16, 17] utilises a neural network to compute the nonlinear similarities between different samples. Moreover, Satorras and Estrach propose to utilise graph neural networks to measure the similarities between images [18]. A large amount of work has also been done to extend the metric-based methods for fine-grained images. For example, BSNet involves two similarity metrics to learn discriminative features [8] and LRPABN adopts bilinear pooling operations [9].

2.2. Feature alignment-based few-shot methods for image classification

Feature alignment methods usually aim to align the object positions between the support and query sets to improve the classification performance [19]. CrossTransformers (CTX) [20] utilises the transformer-based network to explore the spatially-correlated features and calculate the similarity between two images. A more recent transformer-based method is QSFormer [21], which effectively learns consistent representations of the support and query sets via the global sample transformer and the local patch transformer. Dynamic meta-filer (DMF) [22] considers both channel-wise and spatial-wise alignments by neural ordinary differential equation. Relational embedding network (RENet) utilises the self-correlational representation (SCR) module and the cross-correlational attention (CCA) module, where the SCR module transforms the basic feature maps into self-correlational tensors and extracts structural patterns, while the CCA module calculates the cross-correlations between images and generates common attention between them. FRN [11] aligns the features maps of the query image and the support set via reconstructing the query image based on the pooled support features, where the ridge regression-based reconstruction with close-form solutions makes the process efficient. Besides the L_2 norm adopted in FRN, Sun et al. [23] propose to utilise the $L_{2,1}$ norm for feature reconstruction. To alleviate overfitting of the reconstruction-based methods, Li et al. [24] propose the self-reconstruction network that can diversify the query features by reconstructing the query features by themselves.

3. Methodology

3.1. Problem definition

Few-shot learning aims to learn discriminative knowledge from a small amount of labelled data to classify test instances from new tasks. In few-shot learning, the dataset is usually divided into a base set $\mathcal{D}_{\mathcal{B}}$, a validation set $\mathcal{D}_{\mathcal{V}}$ and a novel set $\mathcal{D}_{\mathcal{N}}$, where the classes of the three subsets do not intersect. Few-shot learning learns from the tasks on $\mathcal{D}_{\mathcal{B}}$ to classify instances of new tasks on $\mathcal{D}_{\mathcal{N}}$. The instances in $\mathcal{D}_{\mathcal{V}}$ assist to find the best model during the training process. In this paper, we follow the classic setting of N-way K-shot, i.e. the model is trained by the support set, $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^{N\times K}$, of N classes with K instances each class, and evaluated on the query set of the same classes in $\mathcal{S}$, $\mathcal{Q} = \{\mathbf{x}_j, y_j\}_{j}^{N\times q}$, of N classes with q instances each class. The classification performance of the trained model is finally tested on $\mathcal{D}_{\mathcal{N}}$ with its average classification accuracy as the performance measure.

3.2. FRN+TDM

In metric-based few-shot learning methods, reshaping feature maps to feature vectors as input to metric function can lead to loss of spatial details. FRN [11] aims to resolve this problem by reconstructing every location of the query feature map by the pooled support features from each class through ridge regression. The class membership of the query instance is then assigned based on the reconstruction error. However, in FRN, all channels are treated equally with the same weights, which cannot stress the regions with high discriminative abilities. To identify the discriminative regions, the TDM module can be embedded in the FRN framework.

Specifically, TDM [12] takes the features extracted from the embedding module to calculate the task-wise channel weight vector $\boldsymbol{\beta}_n$ of the nth class as a linear combination of the support weight vector $\boldsymbol{\beta}_n^S$ and the query weight vector $\boldsymbol{\beta}^Q$:

$$\boldsymbol{\beta}_n = \alpha\boldsymbol{\beta}_n^S + (1-\alpha)\boldsymbol{\beta}^Q \in \mathbb{R}^C, \tag{1}$$

where $\alpha \in [0, 1]$ is a hyper-parameter. $\boldsymbol{\beta}_n^S$ and $\boldsymbol{\beta}^Q$ are obtained from the support attention module (SAM) and the query atttention module (QAM), respectively, based on the task-wise intra-class scores $\mathbf{r}_n^{\text{intra}}$ and inter-class scores $\mathbf{r}_n^{\text{inter}}$.

The input to SAM is the prototype of each class $\boldsymbol{\mathcal{P}}_n \in \mathbb{R}^{H\times W\times C}$, i.e. the average of all support set instances in the nth class. The cth element of $\mathbf{r}_n^{\text{intra}}$ is then calculated as

$$r_{n,c}^{\text{intra}} = \frac{1}{H\times W}||\boldsymbol{\mathcal{P}}_{n,c} - \mathbf{M}_n||_2^2, \tag{2}$$

where H and W are the height and width of the feature maps, C is the number of channels, $\boldsymbol{\mathcal{P}}_{n,c} \in \mathbb{R}^{H\times W}$ is the cth channel of the nth prototype and $\mathbf{M}_n \in \mathbb{R}^{H\times W}$ is the average of the channels in $\boldsymbol{\mathcal{P}}_n$, i.e. $\mathbf{M}_n = \frac{1}{C}\sum_{c=1}^{C}\boldsymbol{\mathcal{P}}_{n,c}$. Thus $\mathbf{r}_n^{\text{intra}}$ measures the dispersion of the channels in the prototype of each class. On the contrary, the cth element of $\mathbf{r}_n^{\text{inter}}$ involves information from different classes:

$$r_{n,c}^{\text{inter}} = \frac{1}{H\times W}\min_{1\leq l\leq N, n\neq l}||\boldsymbol{\mathcal{P}}_{n,c} - \mathbf{M}_l||_2^2, \tag{3}$$

where $\mathbf{M}_l$ denotes the mean spatial features of the lth class. It is clear that $r_{n,c}^{\text{inter}}$ measures the difference between each channel and its closest mean spatial features of a different class. Finally, we obtain $\boldsymbol{\beta}_n^S$ as

$$\boldsymbol{\beta}_n^S = \eta(g^{\text{inter}}(\mathbf{r}_n^{\text{inter}})) + (1-\eta)(g^{\text{intra}}(\mathbf{r}_n^{\text{intra}})), \tag{4}$$

where g^{inter} and g^{intra} are fully-connected blocks and $\eta \in [0, 1]$. We adopt the same structure for g as in [12].

Since the labels of query images are unknown, only the intra-class score is involved in QAM:

$$r_{Q,c}^{\text{intra}} = \frac{1}{H \times W} ||\boldsymbol{\mathcal{P}}_{Q,c} - \mathbf{M}_Q||_2^2, \tag{5}$$

where $\boldsymbol{\mathcal{P}}_{Q,c}$ is the cth channel of the query feature maps and $\mathbf{M}_Q$ is the mean of all channels of $\boldsymbol{\mathcal{P}}_Q$. Then, $\boldsymbol{\beta}^Q$ is calculated as

$$\boldsymbol{\beta}^Q = g^{\text{Q}}(\mathbf{r}_Q^{\text{intra}}), \tag{6}$$

where g^{Q} is a fully-connected block with the same structure as g^{inter} and g^{intra}. By substituting equation(4) and equation(6) to equation(1), we obtain the task-wise weights $\boldsymbol{\beta}_n$.

In FRN+TDM, suppose the pooled support features of the nth class is $\mathbf{S}_n \in \mathbb{R}^{(K\times H\times W)\times C}$ while the the query features are $\mathbf{Q} \in \mathbb{R}^{(H\times W)\times C}$. $\mathbf{Q}$ is reconstructed by each $\mathbf{S}_n$ via ridge regression:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\text{argmin}}\, ||\mathbf{Q} - \mathbf{W}\mathbf{S}_n||_2^2 + \lambda||\mathbf{W}||_2^2, \tag{7}$$

where $\mathbf{W} \in \mathbb{R}^{(H\times W)\times(K\times H\times W)}$ is the weight matrix and λ is a non-negative value that controls the contribution of the ridge penalty. The reconstructed query image by the nth class is calculated as

$$\hat{\mathbf{Q}}_n = \hat{\mathbf{W}}\mathbf{S}_n. \tag{8}$$

Then, the task-wise weight vector $\boldsymbol{\beta}_n$ is applied to the original and the reconstructed query feature maps to re-weight the channels:

$$\begin{aligned}\mathbf{Q}_n^r &= (\mathbf{1}_{H\times W}\boldsymbol{\beta}_n^T) \odot \mathbf{Q},\\ \hat{\mathbf{Q}}_n^r &= (\mathbf{1}_{H\times W}\boldsymbol{\beta}_n^T) \odot \hat{\mathbf{Q}}_n,\end{aligned} \tag{9}$$

where $\mathbf{1}_{H\times W}$ is a vector of $H \times W$ 1s and $\odot$ is the element-wised Hadamard product.

Lastly, to assign the class membership of the jth query image, we calculate its probability of belonging to the nth class as

$$P(\hat{y}_j = n|\mathbf{x}_j) = \frac{e^{-\gamma d(\mathbf{Q}^r_{n,j}, \hat{\mathbf{Q}}^r_{n,j})}}{\sum_{n' \in [1,N]} e^{-\gamma d(\mathbf{Q}^r_{n',j}, \hat{\mathbf{Q}}^r_{n',j})}}, \tag{10}$$

where $d(\mathbf{Q}^r_n, \hat{\mathbf{Q}}^r_n) = \frac{1}{H\times W}||\mathbf{Q}^r_n - \hat{\mathbf{Q}}^r_n||^2_2$ and γ is a non-negative parameter.

The training process of FRN+TDM is guided by the cross-entropy loss and the auxiliary loss in FRN:

$$\begin{aligned} L_{FRN} &= L_{CE} + L_{AUX} \\ &= -\sum_{j=1}^{Nq} \log(P(\hat{y}_j = y_j|\mathbf{x}_j)) \\ &+ \sum_{n\in[1,N]} \sum_{n'\in[1,N], n'\neq n} ||\hat{\mathbf{S}}_n(\hat{\mathbf{S}}_{n'})^T||^2, \end{aligned} \tag{11}$$

where $\hat{\mathbf{S}}_n$ is the row-normalised $\mathbf{S}_n$.

3.3. Adaptive feature map reconstruction network (AFRN)

Although FRN+TDM has achieved a state-of-the-art performance in few-shot fine-grained image classification, due to the small inter-class discrepancy omnipresent in fine-grained images and the small training set in the setting of few-shot learning, the training of FRN+TDM can still result in excessively flexible boundaries between subcategories and hence overfitting to the seen subclasses in the training set. To mitigate this issue, we propose a simple scheme to amplify inter-class discrepancy and thus improve FRN+TDM. Our scheme only involves two simple amendments to FRN+TDM: relaxing

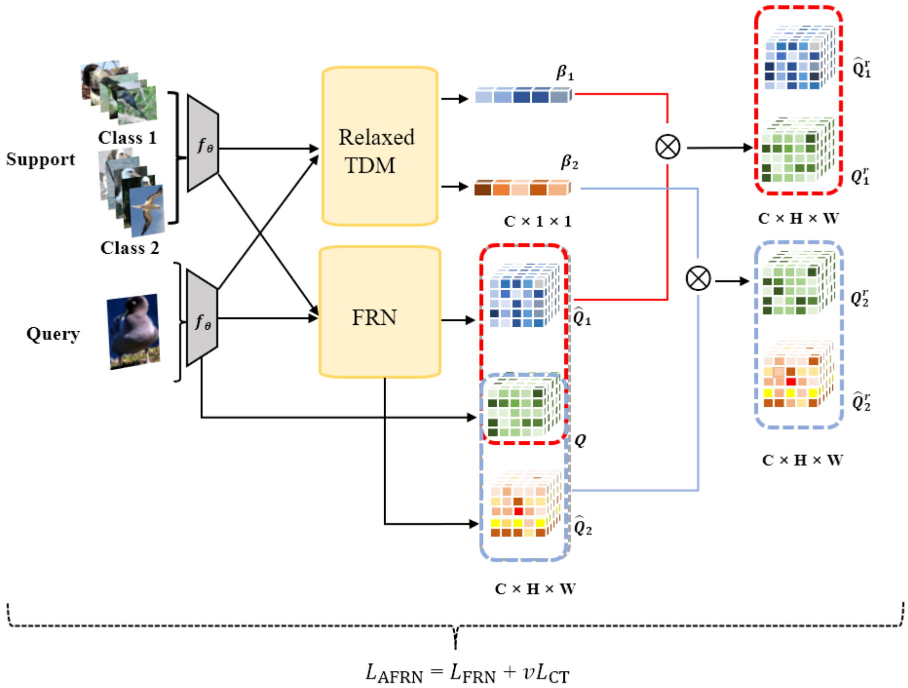

Figure 3: The structure of AFRN with an example of 2-way 5-shot classification. The embedded features of the support set and the query set are input to the FRN and the relaxed TDM modules. The FRN module reconstructs the query feature map by the pooled support features of each class and output the reconstructed query feature maps $\hat{\mathbf{Q}}_1$ and $\hat{\mathbf{Q}}_2$. The relaxed TDM module produce the task-wise channel weights $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Then, the original query feature map $\mathbf{Q}$ and the reconstructed $\hat{\mathbf{Q}}_1$ are re-weighted by $\boldsymbol{\beta}_1$ to obtain $\mathbf{Q}_1^r$ and $\hat{\mathbf{Q}}_1^r$. Similarly, $\mathbf{Q}$ and $\hat{\mathbf{Q}}_2$ are re-weighted by $\boldsymbol{\beta}_2$ to obtain $\mathbf{Q}_2^r$ and $\hat{\mathbf{Q}}_2^r$. Lastly, the two pairs of re-weighted query features are used to obtain probabilities in equation(10) to assign the membership of the query image.

the inter-class score in TDM, and adding a centre loss to FRN. We call the network incorporating our scheme to FRN+TDM the adaptive feature map reconstruction network (AFRN). The structure of AFRN is illustrated in Figure 3.

3.3.1. Relaxing inter-class score in TDM

In equation(3), $r_{n,c}^{\mathrm{inter}}$ measures the minimum distance between each channel and its closest mean spatial features of a different class. Therefore, the classes that are mostly difficult to distinguish are specifically considered. However, this may lead to extremely flexible classification boundaries in the setting of fine-grained image classification, which is even severer in the few-shot setting where the classes in the base set and the novel set are not the same, due to the overfitting to the seen subclasses in the base set. To mitigate this problem, we propose the relaxed TDM by revising the calculation of $r_{n,c}^{\mathrm{inter}}$ in equation (3) as

$$r_{n,c}^{\mathrm{inter}} = \frac{1}{H \times W} \max_{1 \leq l \leq N, n \neq l} ||\boldsymbol{\mathcal{P}}_{n,c} - \mathbf{M}_l||_2^2. \tag{12}$$

In this way, $r_{n,c}^{\mathrm{inter}}$ measures the differences between classes that are less difficult to distinguish, which makes the classification boundaries less flexible and thus mitigates the overfitting to a large extent.

3.3.2. Adding centre loss to FRN

The centre loss L_{CT} measures the intra-class variation of each class, which is calculated as

$$L_{CT} = \sum_{j=1}^{Nq} ||\mathbf{Q}_j - \mathbf{C}_{y_j}||_2^2, \tag{13}$$

where $\mathbf{C}_{y_j}$ denotes the centre of the y_jth class, and $\mathbf{Q}_j$ represents the feature of the jth query. To effectively update the centre, we compute the centre as the average of the query samples in one task.

Hence, the total loss function of AFRN is a simple amendment to that of FRN in equation (11):

$$L_{AFRN} = L_{FRN} + \nu L_{CT}. \tag{14}$$

4. Experiments

In this section, we empirically demonstrate the superior classification performance of AFRN on five fine-grained image datasets, by comparing it with eight state-of-the-art methods: MatchingNet [4], ProtoNet [2], CTX [20], DeepEMD [5], RENet [25], MixFSL [26], FRN [11] and FRN+TDM [12].

4.1. Datasets

We choose five publicly-available benchmark datasets for few-shot image classification, namely CUB-200-2011 [27], aircraft [28], Oxford flowers [29], Stanford cars [30] and Stanford dogs [31]. We name these datasets CUB, aircraft, flowers, cars and dogs for short hereafter.

The CUB dataset contains 200 species of birds, with a total of 11,788 images. We randomly divide the 200 categories into the training, validation and test sets, each consisting of 100, 50 and 50 categories, respectively.

The aircraft dataset has 100 classes of aircrafts, with a total of 10,000 images. We randomly divide the dataset into the training set with 50 classes, the validation set with 25 classes and the test set with 25 classes.

The flowers dataset consists of 102 categories of flowers with 8,189 images. Each type of flower consists of 40 to 258 images, mainly featuring common British flowers. We randomly select 51 classes as the training set, 26 classes as the validation set, and 25 classes as the test set.

The cars dataset includes 196 classes of cars, with a total of 16,185 images. We randomly divide the dataset into the training set with 130 classes, the validation set with 17 classes and the test set with 49 classes.

The dogs dataset contains 120 breeds of dogs, with a total of 20,580 images. We randomly divide the 120 categories into the training set with 60 categories, the validation set with 30 categories and the testing set with 30 categories.

4.2. Implementation details

We adopt ResNet-12 as the backbone with the same implementation details as in [28, 32, 33]. The ResNet-12 backbone consists of 4 residual blocks, and each residual block has 3 convolutional layers. We adopt the leaky ReLU with $\alpha = 0.1$ and 2×2 max pooling. We also adopt the deep block from the original implementation [32, 28, 33], so the output size of each residual block is 64, 160, 320 and 640. Therefore, the shape of the output feature map of an input image of size 84×84 is $640 \times 5 \times 5$. During the training process, we implement the standard data augmentation step, including random cropping, horizontal flipping and color jittering, as in [28, 5, 34, 35].

Following [14, 33], we train ResNet-12 for 1,200 epochs and reduce the learning rate proportionally at the 400th and 800th epochs. We use the validation set to select the best performing model during the training process and validate every 20 epochs. We train the models with the 10-way 5-shot setting and test the models with the 5-way 1-shot and 5-way 5-shot setting.

For AFRN, we follow TDM [12] to set $\alpha = \eta = 0.5$, and set $\nu = 0.05$. In section 4.5, we will show the robustness of ν.

AFRN and FRN+TDM have the same amount of parameters and they have the same FLOPs. For the 5-way 1-shot task with 16 query images, their FLOPs is 299.6G per task while for the 5-way 5-shot setting with 16 query images, their FLOPs is 370G per task.

Table 1: 5-way few-shot classification accuracies on the CUB, aircraft, flowers, cars and dogs datasets with the ResNet-12 backbone. Methods labeled by † denote our implementations. The best classification accuracies are labelled in bold fonts.

Method	CUB		Aircraft		Flowers		Cars		Dogs	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet[4] †	71.87±0.24	85.08±0.24	56.74±0.87	73.75±0.69	71.89±0.90	85.46±0.59	45.29±0.82	64.00±0.74	66.48±0.88	79.57±0.63
ProtoNet[2] †	81.02±0.20	91.93±0.11	46.68±0.81	71.27±0.27	75.41±0.22	89.46±0.14	82.29±0.20	93.11±0.10	73.81±0.21	87.39±0.12
CTX[20] †	80.39±0.20	91.01±0.11	65.60±0.25	80.20±0.25	-	-	85.03±0.19	92.63±0.11	73.22±0.22	85.90±0.13
DeepEMD[5]†	75.59±0.30	88.23±0.18	-	-	70.00±0.35	83.63±0.26	73.30±0.29	88.37±0.17	70.38±0.30	85.24±0.18
RENet[25] †	77.45±0.45	90.50±0.26	59.16±0.47	76.48±0.37	79.91±0.42	92.33±0.22	79.66±0.44	91.95±0.22	71.69±0.47	85.60±0.30
MixFSL[26] †	64.53±0.92	80.67±0.64	60.55±0.86	77.57±0.69	72.60±0.91	86.52±0.65	58.15±0.87	80.54±0.63	67.26±0.90	82.05±0.56
FRN[11] †	82.33±0.19	92.02±0.11	70.26±0.22	83.58±0.14	81.68±0.20	92.61±0.11	86.59±0.18	95.01±0.08	76.49±0.21	88.22±0.12
FRN+TDM[12] †	83.31±0.19	92.70±0.10	70.61±0.21	84.53±0.13	82.95±0.19	93.61±0.10	**89.38±0.16**	**96.98±0.06**	76.67±0.21	88.53±0.12
Ours	**83.95±0.18**	**93.17±0.10**	**72.19±0.21**	**85.59±0.13**	**83.59±0.19**	**94.05±0.09**	89.27±0.16	96.89±0.06	**77.01±0.21**	**88.60±0.12**

Table 2: The results of the one-sided paired t-test of comparing the classification accuracies of our method with those of the state-of-the-art methods in Table 1. The null hypothesis H_0 is $\mu_{\mathrm{AFRN}} < \mu_m$, where μ is the mean classification accuracy and $m \in$ {MatchingNet, ProtoNet, CTX, DeepEMD, RENet, MixFSL, FRN, FRN+TDM}.

Ours vs.	MatchingNet	ProtoNet	CTX	DeepEMD	RENet	MixFSL	FRN	FRN+TDM
p value	1×10^{-3}	7×10^{-3}	3.9×10^{-5}	2.8×10^{-4}	2.8×10^{-4}	1.4×10^{-4}	3.3×10^{-5}	7×10^{-3}
Reject at 1% level	✓	✓	✓	✓	✓	✓	✓	✓

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

We report the classification accuracies of AFRN and the eight state-of-the-art methods on five fine-grained image datasets in Table 1. Obviously,

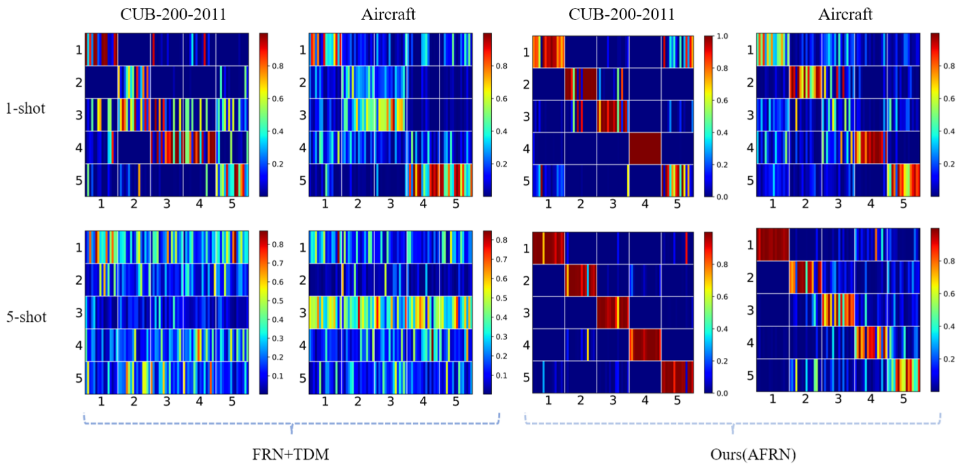

Figure 4: The visualisations of the confusion matrices of AFRN and FRN+TDM on the CUB and aircraft datasets under the 5-way 1-shot and 5-way 5-shot settings.

Table 3: The ablation study on the relaxed TDM module and the centre loss.

	Relaxed TDM	Centre loss	CUB		Aircraft	
			1-shot	5-shot	1-shot	5-shot
(a)	-	-	83.31±0.19	92.70±0.10	70.61±0.21	84.53±0.13
(b)	✓	-	83.73±0.12	92.86±0.10	71.59±0.22	85.06±0.13
(c)	-	✓	83.77±0.18	93.09±0.10	71.05±0.21	84.58±0.13
(d)	✓	✓	**83.95±0.18**	**93.17±0.10**	**72.19±0.21**	**85.59±0.13**

our method can beat all state-of-the-art methods on the CUB, aircraft, flowers and dogs dataset, while providing competitive classification results with FRN+TDM on the cars dataset. This demonstrates the effectiveness of involving the centre loss and the relaxed TDM module. To have a deep insight to the results, we compare the visualisations of the confusion matrices of AFRN and FRN+TDM in Figure 4 on the CUB and aircraft datasets. It is clear that AFRN is better than FRN+TDM on the two datasets with more deep red stripes or higher values on the diagonals. To confirm that AFRN is significantly better than the state-of-the-art methods, we perform one-sided paired t-test to compare the classification accuracies of AFRN and those of other methods in Table 1, with a null hypothesis H_0 of $\mu_{\text{AFRN}} < \mu_m$, where μ is the mean classification accuracy and $m \in \{$MatchingNet, ProtoNet, CTX, DeepEMD, RENet, MixFSL, FRN, FRN+TDM$\}$. H_0 can be rejected at 1% level for all methods compared, suggesting that the classification accuracy of AFRN is significantly better than those of other methods.

4.4. Ablation studies

Here we explore the impacts of the relaxed TDM module and the centre loss on the classification performance and report the results on the CUB and aircraft datasets in Table 3. For the relaxed TDM column, '-' represents adopting the original TDM module while '✓' is for the proposed relaxed TDM module. For the centre loss column, '-' is to train the model by the original FRN loss in (11) while '✓' represents training the model by the AFRN loss in (14). Thus, scenario-(a) corresponds to FRN+TDM while scenario-(d) represents AFRN. Clearly, the classification accuracy of TDM can be raised by only modifying the inter-class score via the relaxed TDM in scenario-(b). It is worth noting that, for the 1-shot classification of the aircraft dataset, the accuracy is improved greatly by almost 1%, suggesting that the subcatergories of aircraft are highly similar and the relaxed score is required to reduce potential overfitting. In scenario-(c), when we only involve the additional centre loss, the improvement is more substantial for the CUB dataset, suggesting that the variation within each subcategory of the CUB dataset is relatively large and thus making intra-class variation smaller via centre loss is beneficial. Finally, utilising the relaxed TDM module as well as the centre loss can provide the best classification accuracies.

Table 4: The effect of ν in (14) of the AFRN loss.

ν	CUB		Flowers	
	1-shot	5-shot	1-shot	5-shot
0.5	83.22±0.19	92.75±0.10	82.75±0.19	93.46±0.10
0.05	**83.95±0.18**	**93.17±0.10**	**83.59±0.19**	**94.05±0.09**
0.005	83.69±0.18	93.07±0.10	82.35±0.20	93.22±0.10

4.5. The effect of ν in (14)

In this section, we present the effect of ν in (14), i.e. the parameter controlling the contribution of the centre loss, on the classification performance. The classification accuracies of the CUB and flowers datasets for three values of ν, 0.5, 0.05 and 0.005, are summarised in Table 4. It shows that 0.05 is a proper choice. In addition, the accuracies of using the three values of ν are all higher than or competitive with FRN+TDM.

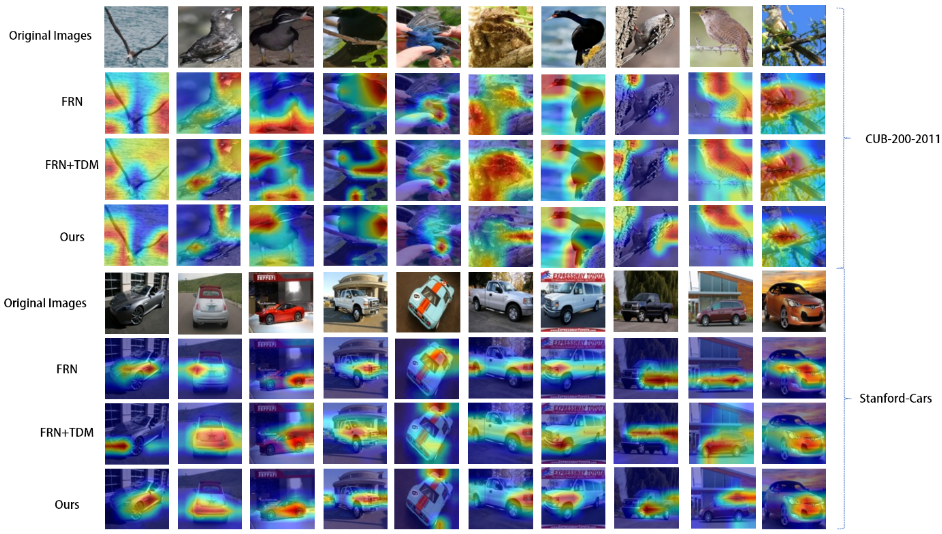

Figure 5: The visualisation of the discriminative features extracted by FRN, FRN+TDM and AFRN ('Ours') on the CUB and cars datasets. AFRN focuses on the most discriminative regions compared with FRN and FRN+TDM.

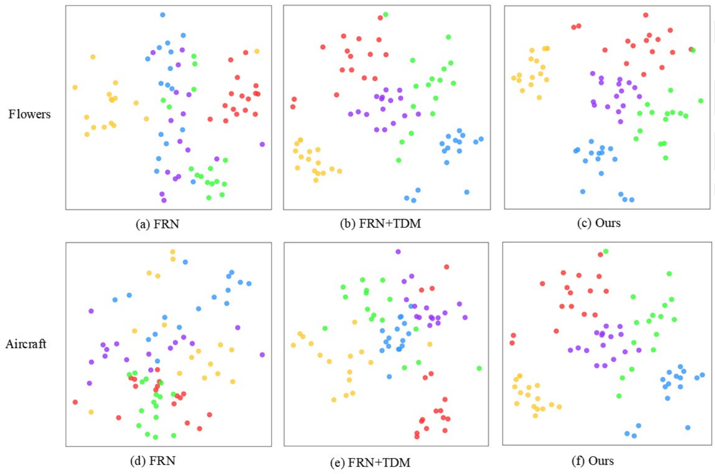

Figure 6: The visualisation of the feature embeddings of FRN, FRN+TDM and AFRN ('Ours') on the flowers and aircraft datasets. AFRN can provide the best separation of different classes.

4.6. The visual comparisons of FRN, FRN+TDM and AFRN

4.6.1. Visualisation of discriminative features

To demonstrate that AFRN can focus on the most discriminative regions for classification, we visually compare the discriminative regions identified by FRN, FRN+TDM and AFRN, following the Grad-CAM technology [36] in Figure 5. For the CUB and cars datasets, we randomly select 10 images for visualisation. We can observe that FRN tends to focus on both the objects and irrelevant backgrounds. FRN+TDM can improve this by identifying smaller discriminative regions, while AFRN can usually make the areas even smaller by focusing on the highly discriminative ones.

4.6.2. Visualisation of feature embeddings

To further show that AFRN can amplify the inter-class discrepancy, we visualise the feature embeddings learnt by FRN, FRN+TDM and AFRN via t-distributed stochastic neighbour embedding (t-SNE) [37] in Figure 6. The results of the flowers and aircraft datasets are presented in the first and second rows in Figure 6, respectively. For each dataset, we randomly select five classes with 16 test samples each and label them by different colours. The five classes are severely mixed in FRN while better separated in FRN+TDM. Obviously, the best separation of the classes is achieved by FRN: the inter-class discrepancy is amplified, which also supports our motivation in Figure 1.

4.7. Discussion

Table 5: The classification accuracies of FRN, FRN+TDM and AFRN ('Ours') on two coarse-grained datasets, mini-ImageNet and FC100, with the ResNet-12 backbone. The best classification accuracies are labelled in bold fonts.

	mini-ImageNet		FC100	
	1-shot	5-shot	1-shot	5-shot
FRN	**63.26±0.21**	77.68±0.15	**40.31±0.17**	**55.34±0.17**
FRN+TDM	62.18±0.20	78.41±0.15	39.84±0.17	54.16±0.17
Ours	62.78±0.20	**78.60±0.15**	40.09±0.18	54.38±0.18

In this section, we further test the ability of AFRN to classify coarse-grained data, where larger categories or super-categories with large intra-class variations are considered. We adopt two benchmark coarse-grained datasets, the mini-ImageNet dataset [4] and the FC100 dataset [38]. The mini-ImageNet dataset contains 60,000 images distributed evenly over 100 classes. We randomly divide the dataset to a training set with 64 classes, a validation set with 16 classes and a test set with 20 classes. The FC100 dataset has 100 object categories which are merged to 20 super-categories. We randomly divide it to a training set with 12 super-categories containing 60 object categories, a validation set with 4 super-categories containing 20 object categories and a test set with 4 super-categories containing 20 object categories.

The classification accruacies of FRN, FRN+TDM and AFRN on coarse-grained datasets are reported in Table 5. Clearly, the original FRN dominates FRN+TDM and AFRN in most scenarios, except for the classification of 5-shot mini-ImageNet. However, we note that AFRN performs slightly better than FRN+TDM in all cases, which demonstrate that the two amendments also work on coarse-grained data, but not effective enough to beat the original FRN. One explanation to this result is that TDM or relaxed TDM put too much attention on few channels while ignore information from other channels that may be valuable for coarse-grained data. Thus, they perform worse than the original FRN when all channels are considered equally.

5. Conclusions

In this paper, we propose AFRN, a simple scheme to amplify the inter-class discrepancy and thus improve the classification performance of FRN+TDM on few-shot fine-grained images. To mitigate the potential overfitting to the seen subclasses, we propose to relax the inter-class score in TDM. To enlarge the subtle differences between the subclasses of fine-grained images, we propose to incorporate the centre loss to FRN. Extensive experiments on five fine-grained datasets showcase that our scheme can produce the state-of-the-art performance, verified by statistical tests. Results in ablation study also reveal the effectiveness of each amendment. Moreover, we note one limitation of our method on classifying coarse-grained data, which we identify as our future work.

Acknowledgement

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62176110, the Key Research and Development Program of Gansu Province under Grant 22YF7GA130, S&T Program of Hebei under grant SZX2020034, Hong-liu Distinguished Young Talents Foundation of Lanzhou University of Technology and the Royal Society under International Exchanges Award IEC\NSFC\201071.

References

[1] X. Li, X. Yang, Z. Ma, J.-H. Xue, Deep metric learning for few-shot image classification: A review of recent developments, Pattern Recognition (2023) 109381.

[2] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[3] X. Huang, S. H. Choi, SAPENet: Self-attention based prototype enhancement network for few-shot learning, Pattern Recognition 135 (2023) 109170.

[4] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, D. Wierstra, Matching networks for one shot learning, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.

[5] C. Zhang, Y. Cai, G. Lin, C. Shen, DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, 2020, pp. 12200–12210.

[6] V. N. Nguyen, S. Løkse, K. Wickstrøm, M. Kampffmeyer, D. Roverso, R. Jenssen, SEN: A novel feature normalization dissimilarity measure for prototypical few-shot learning networks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, Springer, 2020, pp. 118–134.

[7] C. Chen, K. Li, W. Wei, J. T. Zhou, Z. Zeng, Hierarchical graph neural networks for few-shot learning, IEEE Transactions on Circuits and Systems for Video Technology 32 (1) (2021) 240–252.

[8] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, J.-H. Xue, BSNet: Bi-similarity network for few-shot fine-grained image classification, IEEE Transactions on Image Processing 30 (2020) 1318–1331.

[9] H. Huang, J. Zhang, J. Zhang, J. Xu, Q. Wu, Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification, IEEE Transactions on Multimedia 23 (2020) 1666–1680.

[10] H. Huang, J. Zhang, L. Yu, J. Zhang, Q. Wu, C. 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) (2021) 853–866.

[11] D. Wertheimer, L. Tang, B. Hariharan, Few-shot classification with feature map reconstruction networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, 2021, pp. 8012–8021.

[12] S. B. Lee, W. Moon, J. Heo, Task discrepancy maximization for fine-grained few-shot classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, 2022, pp. 5321–5330.

[13] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, 2016, pp. 499–515.

[14] S. Gidaris, N. Komodakis, Dynamic few-shot visual learning without forgetting, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 4367–4375.

[15] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. V. Oseledets, V. S. Lempitsky, Hyperbolic image embeddings, in: 2020 IEEE/CVF Conference

on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, 2020, pp. 6417–6427.

[16] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, T. M. Hospedales, Learning to compare: Relation network for few-shot learning, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 1199–1208.

[17] Z. Li, Z. Hu, W. Luo, X. Hu, SaberNet: Self-attention based effective relation network for few-shot learning, Pattern Recognition 133 (2023) 109024.

[18] V. G. Satorras, J. B. Estrach, Few-shot learning with graph neural networks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, 2018.

[19] F. Hao, F. He, J. Cheng, D. Tao, Global-local interplay in semantic alignment for few-shot learning, IEEE Transactions on Circuits and Systems for Video Technology 32 (7) (2021) 4351–4363.

[20] C. Doersch, A. Gupta, A. Zisserman, CrossTransformers: Spatially-aware few-shot transfer, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 21981–21993.

[21] X. Wang, X. Wang, B. Jiang, B. 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) (2023) 7789–7802.

[22] C. Xu, Y. Fu, C. Liu, C. Wang, J. Li, F. Huang, L. Zhang, X. Xue, Learning dynamic alignment via meta-filter for few-shot learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, 2021, pp. 5182–5191.

[23] J. Sun, X. Shen, Q. Sun, Efficient feature reconstruction via l2,1-norm regularization for few-shot classification, IEEE Transactions on Circuits and Systems for Video Technology 33 (12) (2023) 7452–7465.

[24] X. Li, Z. Li, J. Xie, X. Yang, J.-H. Xue, Z. Ma, Self-reconstruction network for fine-grained few-shot classification, Pattern Recognition 153 (2024) 110485.

[25] D. Kang, H. Kwon, J. Min, M. Cho, Relational embedding for few-shot classification, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021, pp. 8802–8813.

[26] A. Afrasiyabi, J. Lalonde, C. Gagné, Mixture-based feature space learning for few-shot image classification, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, 2021, pp. 9021–9031.

[27] C. Wang, H. K. Galoogahi, C. Lin, S. Lucey, Deep-LK for efficient adaptive object tracking, in: 2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018, 2018, pp. 627–634.

[28] H. Ye, H. Hu, D. Zhan, F. Sha, Few-shot learning via embedding adaptation with set-to-set functions, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, 2020, pp. 8805–8814.

[29] M. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008, 2008, pp. 722–729.

[30] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013, 2013, pp. 554–561.

[31] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization: Stanford dogs, in: Proc. CVPR workshop on fine-grained visual categorization (FGVC), Vol. 2, Citeseer, 2011.

[32] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, P. Isola, Rethinking few-shot image classification: a good embedding is all you need?, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, 2020, pp. 266–282.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[34] Y. Wang, W.-L. Chao, K. Q. Weinberger, L. Van Der Maaten, SimpleShot: Revisiting nearest-neighbor classification for few-shot learning, arXiv preprint arXiv:1911.04623.

[35] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, arXiv preprint arXiv:1703.03130.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 618–626.

[37] G. Hinton, L. Van Der Maaten, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.

[38] B. Oreshkin, P. Rodríguez López, A. Lacoste, TADAM: Task dependent adaptive metric for improved few-shot learning, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.

- We propose AFRN, a simple scheme to amplify inter-class discrepancy.
- We relax the inter-class score in TDM to mitigate the negative impact of overfitting.
- We incorporate the guidance of the centre loss to FRN to enhance the discriminative power of learnt features.
- The experimental results demonstrate that our scheme is simple yet effective to improve the few-shot classification performance.

Xiaoxu Li received the Ph.D. degree from Beijing University of Posts and Telecommunications 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.

Zijie Guo received the B.E. degree in Management from Hankou University in 2021. He is a postgraduate student in Lanzhou University of Technology. His research interests include computer vision and few-shot learning.

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. She is an Associate Editor of Neurocomputing.

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

Jun Guo received B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and network management.

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 in the Department of Statistical Science at University College London. His research interests include statistical pattern recognition, machine learning and computer vision. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Cybernetics, and the IEEE Transactions on Neural Networks and Learning Systems.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: