REVISITING FEATURE ACQUISITION BIAS FOR FEW-SHOT FINE-GRAINED IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on metric-learning based few-shot fine-grained image classification (FSFGIC) has achieved promising success in classification accuracy, where various convolutional neural networks with different similarity measures are utilized to learn a common feature representation for each category for FSFGIC. In this paper, we identify and analyze for the first time a fundamental problem of existing metric-learning based FSFGIC methods which fail to effectively address the bias in feature information obtained from each input image that causes misclassification. To solve this problem, we present a robust feature acquisition network (RFANet) that has the ability to effectively address the bias in the feature information obtained from each input image and guide convolution-based embedding models to significantly increase the accuracy. Our proposed architecture can be easily embedded into any episodic training mechanisms for end-to-end training from scratch. Extensive experiments on FSFGIC tasks demonstrate the superiority of the proposed method over the state-of-the-arts.

1 INTRODUCTION

Learning from very limited training data has attracted increasing attention due to the high cost of data collection and annotation. In recent years, a variety of few-shot fine-grained image classification (FSFGIC) methods (Vinyals et al., 2016; He et al., 2016) have been presented for classifying similar images with very limited data. A basic and straightforward technique is to only leverage feature information from only few data for image classification. Take the k-nearest neighbors algorithm (Fix & Hodges, 1989) as an example, the label of a query image can be predicted based on the similarity measure between the query image and a few examples without any learning mechanism. However, it is nearly impossible for the k-nearest neighbors algorithm to learn a real concept with diverse and complex feature representations from just very limited examples (Li et al., 2019b). Another widely used technique (Koch et al., 2015; Snell et al., 2017) is to utilize different transfer learning mechanisms to develop transferable feature representations based on additional auxiliary datasets, which can be roughly classified into two groups: meta-learning based methods (Xu et al., 2021; Tsutsui et al., 2022) and metric-learning based methods (Huang et al., 2020; Liao et al., 2022; Huang et al., 2022). Meta-learning based methods aim to train a meta learner with a meta-learning paradigm or a learning-to-learning paradigm for FSFGIC. Metric-learning based methods aim to learn a transferable model based on similarity metrics between different categories.

In this work, we address a fundamental problem in the metric-learning. Generally, the metric-learning based FSFGIC methods include two main components: a feature embedding network and a similarity metric learning network. One of the main issues of the metric-learning methods is how to employ a convolution neural network (CNN) (Vinyals et al., 2016; He et al., 2016) with different similarity measures (e.g., nearest neighbor metric (Snell et al., 2017), hyperbolic distance (Khrulkov et al., 2020), cosine metric (Li et al., 2019a), or learned parametric options (Sung et al., 2018)) to learn an effective feature representation for each category. Although the existing metric-learning based FSFGIC methods have achieved some degree of success, the following questions remain open: (1) Have we properly considered how to robustly acquire feature information from each input image? (2) Have we properly considered how to effectively address the *bias in the feature information*[1] obtained from each input image due to image *quantization noise*[2] caused by data augmentation (DA) operations on images (e.g., image affine transformations, random cropping, or color jittering operations)? and (3) Have we considered how to properly guide the CNN to acquire feature information from each input image? Take the deep nearest neighbor neural network (DN4) (Li et al., 2019a) as an example as shown in Fig. 1, the five support images belong to five different categories, the support image $\boldsymbol{s}_5$ and query image $\boldsymbol{q}$ belong to the same category, and DA operations with random cropping, random horizontal flipping, and color jittering are randomly performed on the six images. DAs are carried out once on the five support images and DAs are carried out three times on the query image for obtaining three new query images (i.e., $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, and $\boldsymbol{q}_3$). Under the condition that all parameter settings remain unchanged, the five support images and the three query images after DA are sent into the trained DN4 model, and their corresponding similarities are shown in Fig. 1. Although the objects of the same category have the same feature properties, it can be seen from the cosine similarities in Fig. 1 that only query image $\boldsymbol{q}_2$ and support image $\boldsymbol{s}_5$ can be correctly classified into one category in these three classification instances. It can also be seen from Fig. 1 that the similarities between a query image and different support images from different categories are very close. The reasons are: (1) There is a bias in obtaining image feature information from a given CNN; (2) DA operation increases the quantization noise of images to a certain degree; and (3) Fine-grained images have small inter-class and large intra-class variations.

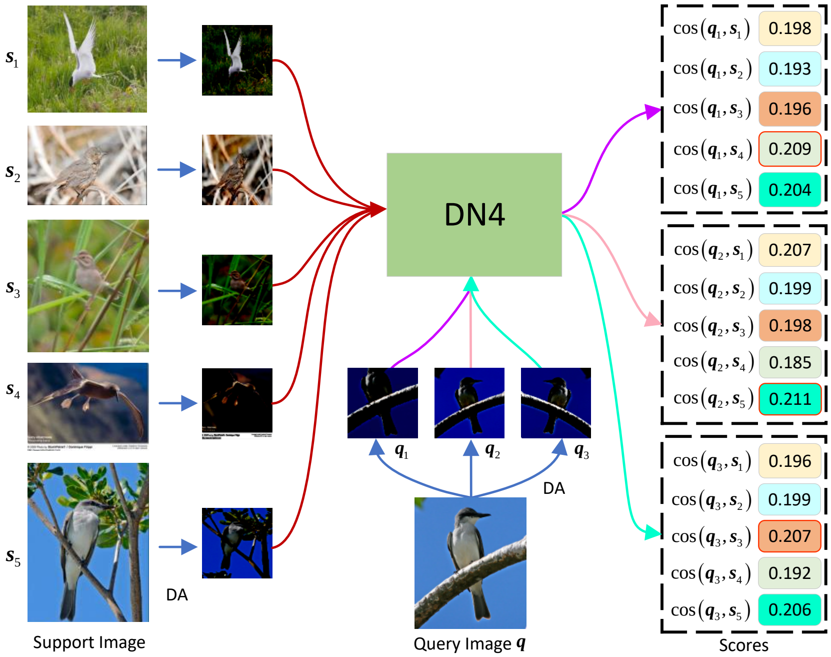

Figure 1: The effect of bias on acquired feature information on FSFGIC.

To address the aforementioned problems, this paper designs a novel and effective architecture for robust feature acquisition. The contributions of the paper can be summarized as follows: (1) We identify and analyze for the first time that existing metric-learning based FSFGIC algorithms are prone to the bias in the feature information obtained from each input image due to image quantization noise; (2) A novel and robust feature acquisition network (RFANet) is proposed to address this feature representation bias; and (3) The proposed RFANet has the capability to guide any convolution-based embedding models to obtain robust image feature representation for FSFGIC. It is worth to note that our designed architecture can be easily embedded into any episodic training mechanisms for end-to-end training from scratch. Experiments on five fine-gained image classification benchmark datasets, i.e., Stanford Dogs (Khosla et al., 2011), Stanford Cars (Krause et al., 2013), CUB (Wah et al., 2011), NABirds (Van Horn et al., 2015), and Plant Disease (Sharma, 2018), demonstrate that the proposed RFANet significantly outperforms the baseline methods on both 5-way 1-shot and 5-way 5-shot FSFGIC tasks. Furthermore, extensive ablation studies on the proposed method have

[1] "Bias in the feature information" denotes the deviation of features from difference instances (such as augmented versions) of each input image.

[2] Image mapping processes, such as applying DA operations introduces errors because a digital image is quantized to image grids, which is named quantization noise in this paper.

also been conducted. The results validate the effectiveness of our proposed RFANet architecture with very encouraging large performance boostings.

2 Related works

Existing FSFGIC methods can be roughly classified into two main streams: meta-learning based methods and metric-learning based methods. In the following, representative FSFGIC methods and some representative few-shot learning methods which have been used for performing the fine-grained image classification tasks are briefly reviewed.

2.1 Meta-learning based FSFGIC methods

The purpose of meta-learning based methods is to learn a good parameter initialization for FSFGIC. Wei et al. (2019) proposed the first FSFGIC model by utilizing two components. The first component is to obtain subtle image feature information and the second component is to learn the decision boundaries of input datasets. Zhu et al. (2020) presented a multi-attention meta-learning (MattML) algorithm which employed multi-attention on the base learner and the task learner for localizing the discriminate parts in images. Tang et al. (2020) explored the gap between machine and humans for fine-grained image classification and indicated that humans usually utilize stable feature representations which are robust to spatial deformations for FSFGIC. Then a pose-normalization feature representation network is designed for FSFGIC. Xu et al. (2021) argued that an isotropic Gaussian distribution can be used for modelling the intra-class variability. Then, feature information is repeatedly sampled from the learned intra-class variability distribution and added to the class-discriminative feature information for obtaining augmented feature information and performing FSFGIC. Wang et al. (2021) introduced a foreground object transformation (FOT) technique for removing image background. Then a posture transformation generator is presented to generate additional samples for increasing the diversity of data samples.

2.2 Metric-learning based FSFGIC Methods

The purpose of metric-learning based methods is to learn a good task independent embedding for FSFGIC. Currently, many discriminative feature learning architectures (Huang et al., 2019; 2020; Li et al., 2019a) have been presented for FSFGIC. Huang et al. (2020) presented a low-rank pairwise alignment bilinear network (LRPABN) to obtain feature relationships between different fine-grained categories for FSFGIC. To learn more discriminative features, Li et al. (2019a) proposed DN4 which designs an image-to-class similarity measure for image classification with very limited samples. To reduce similarity bias, Li et al. (2020) proposed a bi-similarity network (BSNet) which applies the sum of the two different similarity measures (i.e., the Euclidean distance and the cosine distance) for FSFGIC. Huang et al. (2022) presented a target-oriented alignment network (TOAN) which uses utilizes feature alignment for effectively reducing the intra-class variation in fine-grained images. Dong et al. (2020) presented an adaptive task-aware local representations network (ATL-Net) which aims to localize distinguished local patches in the entire task. Sun et al. (2020) utilized salient patch localization (Selvaraju et al., 2017) and high-order integration for obtaining discriminative feature representations for FSFGIC. Liao et al. (2022) presented an automatic salient region selection architecture for FSFGIC.

Meanwhile, feature distributions with different categories have been explored for FSFGIC. Li et al. (2019b) introduced a covariance metric networks (CovaMNet) which employs the feature distribution consistency for performing FSFGIC. Yang et al. (2021) argued that feature representations from each category follow a Gaussian distribution. Then a distribution calibration (DC) strategy is designed to make the distributions of feature representations more Gaussian-like. It is worth to note that the DC strategy (Yang et al., 2021) can also be implemented in meta-learning based FSFGIC. Furthermore, feature reconstruction techniques (Zhang et al., 2020; Wertheimer et al., 2021; Li & Bian, 2022) have been employed for FSFGIC. Zhang et al. (2020) used the earth mover's distance (EMD) technique (Rubner et al., 2000) and proposed DeepEMD which formulates feature reconstruction as an optimal transportation problem for finding optimal matching between a query image and support images. Wertheimer et al. (2021) introduced a feature map reconstruction network (FRN) in which a query feature is reconstructed directly from support features via ridge regression in a closed form. Li & Bian (2022) argued that it is inaccurate in FRN (Wertheimer et al., 2021) that each descriptor contributes equally to the reconstruction error and proposed a foreground-aware FRN (FAFRN) which uses the weighted squared Euclidean distance as the reconstruction error for FSFGIC.

Within the scope of our investigation, all the aforementioned meta-learning and metric-learning based FSFGIC methods utilize DA techniques (e.g., random cropping, random horizontal flipping, and color jittering) for enhancing the diversity of data samples, reducing overfitting, and improving classification accuracy. It is worth to note that DA operations on image inevitably lead to the generation of quantization noise. The current FSFGIC methods do not address the bias in feature information obtained from each input image which may cause biased feature representations and misclassification. How to effectively acquire feature information and suppress the interference of quantization noise from images has constrained the development of FSFGIC.

3 Methodology

3.1 Problem Statement

A typical few-shot image classification setting contains a support set $\boldsymbol{S}$ and a query set $\boldsymbol{Q}$. Support set $\boldsymbol{S}$ contains $\mathcal{C}$ different image classes and each class in $\mathcal{C}$ is composed of $\mathcal{K}$ labeled samples. Query set $\boldsymbol{Q}$ is composed of unlabeled samples. Set $\boldsymbol{S}$ and set $\boldsymbol{Q}$ have the same data-label space. The goal of FSFGIC is to train a model which has the ability to successfully classify each query sample $\boldsymbol{q}$ ($\boldsymbol{q} \in \boldsymbol{Q}$) into its corresponding class in $\mathcal{C}$. Thus, the FSFGIC task is called a $\mathcal{C}$-way $\mathcal{K}$-shot task.

3.2 Motivation of the Proposed Method

This work is largely inspired by the experimental observation shown in Fig. 1. Although as described in (Krizhevsky et al., 2012) that DA is one of "the two primary ways in which we combat overfitting", DA operations on each image inevitably lead to the generation of quantization noise, which results in bias in feature information obtained from each input image. The bias of feature representation has a great impact on the performance of FSFGIC. The main reasons are as follows: (1) Fine-grained images have small inter-class and large intra-class variations. (2) The training samples of FSFGIC are very limited. They present us with a dilemma. On the one hand, it is necessary to perform DA operations on each image for ensuring the diversity of data images and reducing overfitting. On the other hand, it is necessary to effectively address bias in the feature information obtained from each input image which is caused by the data augmentation (DA) operations on images.

In general, the attributes of the features obtained by a convolutional embedding module from different DA operations on the same image have better consistency than the features obtained from different images under different DA operations. The above analysis motivates us to propose a robust feature acquisition (RFA) mechanism for FSFGIC as follows: DA operations (i.e., random cropping, random horizontal flipping, and color jittering) are randomly performed on a support image and a query image. DA is performed once on the support image and DA is performed n times on the query image for obtaining n augmented query images. The augmented support and query images are sent into a given convolutional embedding module for obtaining feature descriptors. Then the regularized feature descriptors of the augmented support image and the regularized feature descriptors of the n augmented query images are employed for similarity calculation. Finally, the nth root of the product of these n similarities is used as the similarity measure between the query image and the support image. Under the condition that the DA operation is utilized to achieve diversity on training samples, the proposed RFA mechanism addresses the intrinsic bias problem caused by the image quantization noise.

3.3 The Proposed RFANet

The overview of the proposed RFANet framework for FSFGIC is illustrated in Fig. 2. A novel robust feature acquisition strategy is designed for effectively addressing the interference of quantization noise on feature descriptors under the condition of achieving the diversity of training samples. Predominantly, the proposed RFANet architecture can be trained in an end-to-end manner from scratch. In the following, the proposed RFANet framework will be presented in detail.

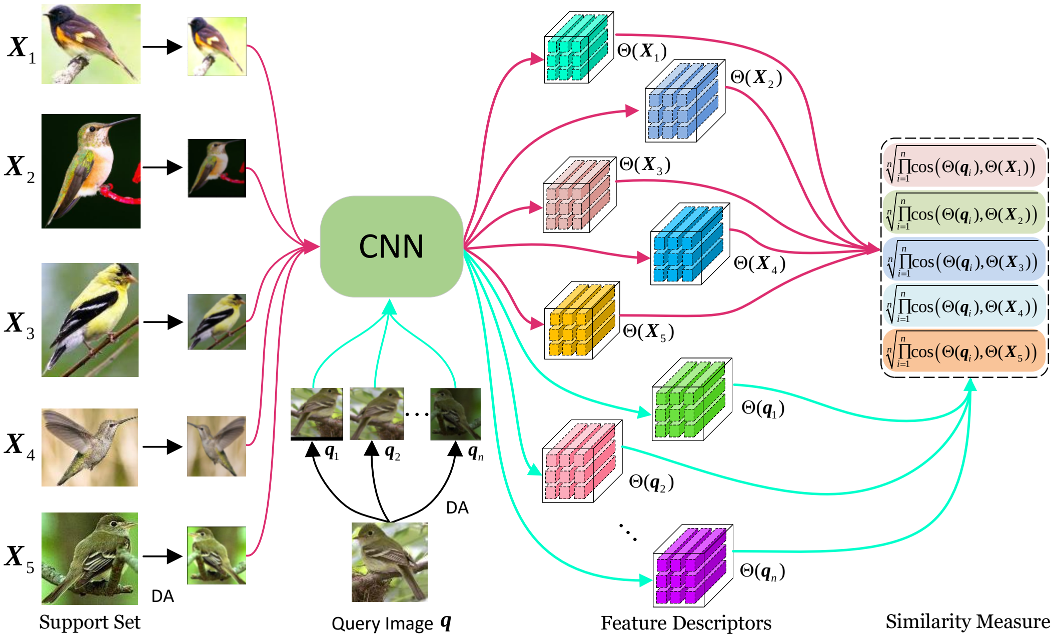
X_1
X_2
X_3
X_4
X_5
DA
Support Set
CNN
q_1
q_2
q_n
DA
Query Image q
$\Theta(X_1)$
$\Theta(X_2)$
$\Theta(X_3)$
$\Theta(X_4)$
$\Theta(X_5)$
$\Theta(q_1)$
$\Theta(q_2)$
$\Theta(q_n)$
Feature Descriptors
$\sqrt[n]{\prod_{i=1}^{n}\cos(\Theta(q_i),\Theta(X_1))}$
$\sqrt[n]{\prod_{i=1}^{n}\cos(\Theta(q_i),\Theta(X_2))}$
$\sqrt[n]{\prod_{i=1}^{n}\cos(\Theta(q_i),\Theta(X_3))}$
$\sqrt[n]{\prod_{i=1}^{n}\cos(\Theta(q_i),\Theta(X_4))}$
$\sqrt[n]{\prod_{i=1}^{n}\cos(\Theta(q_i),\Theta(X_5))}$
Similarity Measure

Figure 2: An overview of the proposed RFANet for a 5-way 1-shot FSFGIC task.

3.3.1 Robust Feature Acquisition Strategy

In this work, the most commonly used networks (e.g., Conv64F (Vinyals et al., 2016) and ResNet256 (He et al., 2016)) in the current FSFGIC methods (Li et al., 2019b;a; Dong et al., 2020; Huang et al., 2022) are selected as the base neural networks. For each query sample $\boldsymbol{q}$ and support sample $\boldsymbol{s}$, DA operations with random cropping, random horizontal flipping, and color jittering are performed once on the support sample and n times on the query sample for obtaining augmented query images (i.e., $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, $\cdots$, $\boldsymbol{q}_n$). The augmented support and query images are sent into a given convolutional embedding module for obtaining feature descriptors which can be represented as $\Theta(\boldsymbol{q}_1)$=$[\widetilde{\boldsymbol{q}}_{1,1}, ..., \widetilde{\boldsymbol{q}}_{1,j}, ..., \widetilde{\boldsymbol{q}}_{1,m}]\in R^{d\times(h\times w)}$, $\Theta(\boldsymbol{q}_2)$=$[\widetilde{\boldsymbol{q}}_{2,1}, ..., \widetilde{\boldsymbol{q}}_{2,j}, ..., \widetilde{\boldsymbol{q}}_{2,m}]\in R^{d\times(h\times w)}$, $\cdots$, $\Theta(\boldsymbol{q}_n)$=$[\widetilde{\boldsymbol{q}}_{n,1}, ..., \widetilde{\boldsymbol{q}}_{n,j}, ..., \widetilde{\boldsymbol{q}}_{n,m}]\in R^{d\times(h\times w)}$, and $\Theta(\boldsymbol{s})$=$[\widetilde{\boldsymbol{s}}_1, ..., \widetilde{\boldsymbol{s}}_j, ..., \widetilde{\boldsymbol{s}}_m]\in R^{d\times(h\times w)}$, where $\widetilde{\boldsymbol{s}}_j$ is the j-th feature descriptor with a length of d, h and w are the height and the width of the feature tensor map, d is the number of filters, R denotes real space, and m (m=$h\times w$) is the total number of feature descriptors for training samples. In this way, the j-th feature descriptor $\widehat{\boldsymbol{q}}_j$ (j=1, ..., m) of query sample $\boldsymbol{q}$ is represented by $\widetilde{\boldsymbol{q}}_{1,j}$, $\widetilde{\boldsymbol{q}}_{2,j}$, $\cdots$, and $\widetilde{\boldsymbol{q}}_{n,j}$. Then standardization operation is performed on feature descriptors $\widetilde{\boldsymbol{q}}_{1,j}$, $\widetilde{\boldsymbol{q}}_{2,j}$, $\cdots$, $\widetilde{\boldsymbol{q}}_{n,j}$, and $\widetilde{\boldsymbol{s}}_j$ (j=1, ..., m) as follows

$$\begin{aligned} \ddot{\boldsymbol{q}}_{1,j} &= \frac{\widetilde{\boldsymbol{q}}_{1,j}-\mho(\widetilde{\boldsymbol{q}}_{1,j})}{\sqrt{\wp(\widetilde{\boldsymbol{q}}_{1,j})}}, \ddot{\boldsymbol{q}}_{2,j} = \frac{\widetilde{\boldsymbol{q}}_{2,j}-\mho(\widetilde{\boldsymbol{q}}_{2,j})}{\sqrt{\wp(\widetilde{\boldsymbol{q}}_{2,j})}}, \cdots, \ddot{\boldsymbol{q}}_{n,j} = \frac{\widetilde{\boldsymbol{q}}_{n,j}-\mho(\widetilde{\boldsymbol{q}}_{n,j})}{\sqrt{\wp(\widetilde{\boldsymbol{q}}_{n,j})}}, \\ \ddot{\boldsymbol{s}}_j &= \frac{\widetilde{\boldsymbol{s}}_j-\mho(\widetilde{\boldsymbol{s}}_j)}{\sqrt{\wp(\widetilde{\boldsymbol{s}}_j)}}, j=1,...,m, \end{aligned} \tag{1}$$

where $\mho(\cdot)$ and $\wp(\cdot)$ denote the mean operation and the variance operation respectively. For each feature descriptor $\widehat{\boldsymbol{q}}_j$ and $\widetilde{\boldsymbol{s}}_j$ (j=1, ..., m), their corresponding similarity measure is

$$\lambda(\widehat{\boldsymbol{q}}_j, \widetilde{\boldsymbol{s}}_j) = \sqrt[n]{\prod_{i=1}^{n}\cos(\ddot{\boldsymbol{q}}_{i,j}, \ddot{\boldsymbol{s}}_j)}, \ \cos(\ddot{\boldsymbol{q}}_{i,j}, \ddot{\boldsymbol{s}}_j) = \frac{\ddot{\boldsymbol{q}}_{i,j}^{\top}\ddot{\boldsymbol{s}}_j}{||\ddot{\boldsymbol{q}}_{i,j}||\cdot||\ddot{\boldsymbol{s}}_j||}, \tag{2}$$

where $\cos(\cdot)$ denotes the cosine similarity. In this way, k-nearest neighbors $\breve{\boldsymbol{s}}_j^t|_{t=1}^k$ for feature descriptor $\widehat{\boldsymbol{q}}_j$ can be found in each class $\mathcal{C}$ based on the k-nearest-neighbor technique (Boiman et al., 2008), where t denotes the t-th k-nearest neighbor. Then the image-to-class similarity measure (Li et al., 2019a) is calculated for FSFGIC as follows

$$\Im(\boldsymbol{q}, \mathcal{C}) = \sum_{j=1}^{m} \sum_{t=1}^{k} \lambda(\widehat{\boldsymbol{q}}_j, \check{\boldsymbol{s}}_j^t). \tag{3}$$

Finally, the Adam optimization (Kingma & Ba, 2015) with a cross-entropy loss is utilized to train the whole network for learning the parameters and performing FSFGIC tasks. In this work, k and n are set to 1 and 3 respectively (See Table 4 and Table 5). We provide the ablation studies of choosing different values of k and n to explore the sensitivity of the proposed method to the hyper-parameters.

4 Experiments

4.1 Datasets

Five fine-gained datasets (i.e., Stanford Dogs (Khosla et al., 2011), Stanford Cars (Krause et al., 2013), CUB (Wah et al., 2011), NABirds (Van Horn et al., 2015), and Plant Disease (Sharma, 2018)) are used for evaluating the performance of the proposed RFANet. The Stanford Dogs dataset contains 120 dog classes with 20,580 samples. The Stanford Cars dataset consists of 196 car classes with 16,185 samples. The CUB dataset contains 200 bird classes with 11,788 samples. The NABirds dataset consists of 555 bird classes with 48,562 samples. The Plant Disease dataset contains 38 plant disease classes with 54,306 samples. Following the latest data split introduced in (Huang et al., 2022), the same data splits are used in our work which are summarized in Table 1.

Table 1: The class split of five fine-grained datasets. N_{train}, N_{val}, and N_{test} are the numbers of classes in the training set, validation set, and test set respectively.

Dataset	N_{train}	N_{val}	N_{test}
Stanford Dogs	70	20	30
Stanford Cars	130	17	49
CUB	120	30	50
NABirds	350	66	139
Plant Disease	20	10	8

4.2 Experimental Setup

We conduct 5-way 1-shot and 5-way 5-shot experiments, which are standard FSFGIC settings, on the aforementioned five datasets. Each input image is resized to a fixed size of 100×100 and randomly cropped into 84×84. Random cropping, random horizontal flipping, and color jittering are utilized for data augmentation. In terms of the episodic training paradigm (Vinyals et al., 2016), 300,000 episodes are randomly selected and constructed for training models. In each episode, 15 and 10 query images are randomly selected from each class for the 1-shot and 5-shot settings respectively. Meanwhile, the Adam optimization technique (Kingma & Ba, 2015) with an initial learning rate of 0.001 is used for optimizing the designed models. The learning rate is reduced by half for every 100,000 episodes. During the testing process, 600 episodes are randomly constructed from the testing set for obtaining the top-1 results on mean classification accuracy. Meanwhile, the 95% confidence intervals are obtained and reported. It is worth to note that the proposed RFANet architecture is trained in an end-to-end manner from scratch and without any fine-tuning during testing.

4.3 Performance Comparison

The proposed RFANet architecture has been compared with fourteen state-of-the-art methods (MatchingNet (Vinyals et al., 2016), Prototypical Net (ProtoNet) (Snell et al., 2017), RelationNet (Sung et al., 2018), CovaMNet (Li et al., 2019b), PCM (Wei et al., 2019), DN4 (Li et al., 2019a), MattML (Zhu et al., 2020), PABN+$_{\text{cpt}}$ (Huang et al., 2020), LRPABN$_{\text{cpt}}$ (Huang et al.,

Table 2: Comparison results on the Stanford Dogs, Stanford Cars, and CUB datasets ($\flat$ denotes that the results are from (Huang et al., 2022)).

Model	Backbone	5-way Accuracy (%)					
		Stanford Dogs		Stanford Cars		CUB	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet (Vinyals et al., 2016)	Conv64F	45.05±0.66	60.60±0.62	48.03±0.60	64.22±0.59	57.59±0.74	70.57±0.62
ProtoNet (Snell et al., 2017)	ConvNet256	42.58±0.63	59.45±0.65	45.27±0.61	64.24±0.61	53.88±0.72	70.85±0.63
RelationNet (Sung et al., 2018)	Conv64F	44.75±0.70	58.36±0.66	56.02±0.74	66.93±0.63	59.82±0.77	71.83±0.61
CovaMNet (Li et al., 2019b)	Conv64F	49.10±0.76	63.04±0.65	56.65±0.86	71.33±0.62	58.51±0.94	71.15±0.80
PCM$^\flat$ (Wei et al., 2019)	AlexNet	28.78±2.33	46.92±2.00	29.63±2.38	52.28±1.46	42.10±1.96	62.48±1.21
DN4 (Li et al., 2019a)	Conv64F	45.41±0.76	63.51±0.62	59.84±0.80	88.65±0.44	55.60±0.89	77.64±0.68
DN4$^\flat$ (Li et al., 2019a)	ResNet256	51.83±0.80	69.83±0.66	76.62±0.70	89.57±0.40	67.17±0.81	82.09±0.56
MattML (Zhu et al., 2020)	Conv64F	54.84±0.53	71.34±0.38	66.11±0.54	82.80±0.28	66.29±0.56	80.34±0.30
PABN$^\flat$+$_{\text{cpt}}$ (Huang et al., 2020)	Conv64F	45.65±0.71	61.24±0.62	54.44±0.71	67.36±0.61	63.36±0.80	74.71±0.60
LRPABN$^\flat{}_{\text{cpt}}$ (Huang et al., 2020)	Conv64F	45.72±0.75	60.94±0.66	60.28±0.76	73.29±0.58	63.63±0.77	76.06±0.58
SoSN (Zhang & Koniusz, 2019)	Conv64F	48.01±0.76	64.95±0.64	62.84±0.68	75.75±0.52	63.95±0.72	78.79±0.60
BSNet(R&C) (Li et al., 2020)	Conv64F	51.24±0.96	71.65±0.76	61.71±0.85	85.65±0.78	66.91±0.89	84.52±0.67
ATL-Net (Dong et al., 2020)	Conv64F	54.49±0.92	73.20±0.69	67.95±0.84	89.16±0.48	63.42±0.86	86.21±0.69
FOT (Wang et al., 2021)	Conv64F	49.32±0.74	68.18±0.69	54.55±0.73	73.69±0.65	67.46±0.68	83.19±0.43
TOAN (Huang et al., 2022)	Conv64F	49.30±0.77	67.16±0.49	65.90±0.72	84.24±0.48	65.34±0.75	80.43±0.68
TOAN (Huang et al., 2022)	ResNet256	51.83±0.80	69.83±0.66	76.62±0.70	89.57±0.40	67.17±0.81	82.09±0.56
Ours	Conv64F	**55.72±0.66**	**75.96±0.51**	**69.76±0.66**	**91.52±0.31**	**71.18±0.63**	**88.38±0.41**
	ResNet256	**65.67±0.79**	**82.17±0.40**	**83.78±0.63**	**93.46±0.42**	**78.48±0.71**	**92.54±0.44**

2020), SoSN (Zhang & Koniusz, 2019), BSNet (Li et al., 2020), ATL-Net (Dong et al., 2020), FOT (Wang et al., 2021), and TOAN (Huang et al., 2022)). Table 2 and Table 3 show the comparison results on the Stanford Dogs, Stanford Cars, CUB, NABirds, and Plant Disease datasets. It is observed from Table 2 and Table 3 that our proposed method achieves the best performance on both 5-way 1-shot and 5-way 5-shot FSFGIC tasks. For the 5-way 1-shot and 5-way 5-shot FSFGIC tasks on the CUB dataset, our proposed RFANet with the backbone of Conv64F achieves 13.59%, 17.30%, 11.36%, 12.67%, 29.08%, 13.58%, 4.89%, 7.82%, 7.55%, 7.23%, 4.27%, 7.76%, 3.72%, and 5.84% improvements and 17.81%, 17.53%, 16.55%, 17.23%, 25.90%, 10.74%, 8.04%, 13.67%, 12.32%, 9.59%, 3.86%, 2.17%, 5.19%, and 7.95% improvements over MatchingNet, ProtoNet, RelationNet, CovaMNet, PCM, DN4, MattML, PABN+$_{\text{cpt}}$, LRPABN$_{\text{cpt}}$, SoSN, BSNet, ATL-Net, FOT, and TOAN respectively.

It is worth to note that the proposed RFANet architecture only considers how to effectively address the bias in feature information obtained from each input image and does not perform further processing on the extracted features such as attention mechanism feature reconstruction. The algorithm that can perform a fair performance comparison with the proposed RFANet is DN4. The differences between the proposed RFANet and the DN4 (Li et al., 2019a) are as follows: DAs are performed three times on the query image in this work for obtaining three augmented query images and the regularized feature descriptors of the three augmented query images and the regularized feature descriptors of the augmented support image are used for similarity calculation. Then the cubic root of the product of these three similarities is used as the similarity measure between the query image and the support image. It can be seen from Table 2 and Table 3 that the proposed RFANet achieves far better performance than the DN4. The reason is as described in (Huang et al., 2022): Different fine-grained categories have very similar shapes and visual appearances; the similarities between any two samples are always high which means that the k-nearest features sorted by DN4 in different support classes are also similar which leads to the performance degradation of DN4 in dealing with FSFGIC. By addressing the bias problem, the proposed RFANet has the capability to guide any convolution-based embedding model for robustly obtaining feature representations from images and achieves much better performance than DN4 and the other thirteen state-of-the-art methods.

Table 3: Comparison results on the NABirds and Plant Disease datasets ($\flat$ denotes that the results are from (Huang et al., 2022) and $-$ indicates that the results are not reported).

Model	Backbone	5-way Accuracy (%)			
		NABirds		Plant Disease	
		1-shot	5-shot	1-shot	5-shot
MatchingNet (Vinyals et al., 2016)	Conv64F	60.70±0.78	76.23±0.62	64.22±0.88	80.79±0.99
ProtoNet (Snell et al., 2017)	ConvNet256	55.85±0.78	75.34±0.63	64.97±0.85	82.73±0.91
RelationNet (Sung et al., 2018)	Conv64F	64.34±0.81	77.52±0.60	65.16±0.81	84.62±0.85
CovaMNet (Li et al., 2019b)	Conv64F	60.03±0.98	75.63±0.79	70.72±0.89	88.92±0.81
DN4 (Li et al., 2019a)	Conv64F	51.81±0.91	83.38±0.60	72.47±0.76	90.68±0.44
PABN$^{\flat}$+$_{\text{cpt}}$ (Huang et al., 2020)	Conv64F	66.94±0.82	79.66±0.62	-	-
LRPABN$^{\flat}$$_{\text{cpt}}$ (Huang et al., 2020)	Conv64F	67.73±0.81	81.62±0.58	-	-
SoSN (Zhang & Koniusz, 2019)	Conv64F	69.53±0.77	83.87±0.51	71.02±0.89	89.31±0.67
ATL-Net (Dong et al., 2020)	Conv64F	69.29±0.94	85.18±0.65	72.18±0.92	90.11±0.65
TOAN (Huang et al., 2022)	Conv64F	70.02±0.80	85.52±0.50	-	-
TOAN (Huang et al., 2022)	ResNet256	76.14±0.75	90.21±0.40	-	-
Ours	Conv64F	**70.62±0.67**	**87.11±0.62**	**76.98±0.74**	**91.32±0.53**
	ResNet256	**78.53±0.68**	**91.12±0.64**	**77.10±0.72**	**91.91±0.61**

Table 4: The impact of the k-nearest neighbors in the proposed RFANet obtained on the Stanford Dogs and Stanford Cars datasets.

Model	5-way Accuracy (%)			
	Stanford Dogs		Stanford Cars	
	1-shot	5-shot	1-shot	5-shot
RFANet_{$k = 5$}	54.36±0.62	74.58±0.54	68.28±0.69	89.82±0.34
RFANet_{$k = 3$}	55.11±0.61	75.36±0.58	69.01±0.62	91.11±0.33
RFANet_{$k = 1$}	**55.72±0.66**	**75.96±0.51**	**69.76±0.66**	**91.52±0.31**

4.4 Ablation Studies

Influence of the k-nearest neighbors for FSFGIC. The performance of our designed RFANet architecture on FSFGIC tasks will be affected by the selection of parameter k. Additional experiments are conducted for investigating the effect of the parameter k in Equation (3) on performance. Based on the backbone of Conv64F, three different k ($k \in \{1, 3, 5\}$) are selected for performing 5-way 1-shot and 5-way 5-shot FSFGIC tasks on the Stanford Dogs and Standford Cars datasets with the fixed parameter n (n=3). The results are shown in Table 4. It is observed from Table 4 that when k equals 1, the proposed RFANet achieves the best classification performance. Therefore, k-nearest neighbors with k=1 is recommended for our designed architecture.

Influence of n DA operations on the query image for FSFGIC. The performance of our designed RFANet architecture on FSFGIC tasks will be affected by the selection of parameter n. Additional experiments on FSFGIC are conducted for investigating the effect of the parameter n in Equation (2) on the performance. Based on the backbone of Conv64F, four different n ($n \in \{2, 3, 4, 5\}$) are selected for performing 5-way 1-shot and 5-way 5-shot FSFGIC tasks on the Stanford Dogs and Standford Cars datasets with a fixed parameter k (k=1). The results on the 5-way 1-shot and 5-way 5-shot tasks are shown in Table 5. It is observed from Table 5 that when n equals 3, the proposed RFANet achieves the allover best classification performance. Therefore, DA operations on the query image with n=3 is recommended for our designed architecture.

The loss curves and the accuracy curves of the DN4 and the proposed RFANet on the Stanford Dogs dataset. Take the 5-way 1-shot FSFGIC task on the Stanford Dogs dataset as an example, the loss curves of training and validation of the DN4 and the proposed RFANet are shown in Fig. 3

Table 5: The impact of the n DA operations of the proposed RFANet obtained on the Stanford Dogs and Stanford Cars.

Model	5-way Accuracy (%)			
	Stanford Dogs		Stanford Cars	
	1-shot	5-shot	1-shot	5-shot
RFANet_$\{n = 5\}$	55.18±0.69	75.25±0.59	69.11±0.61	91.28±0.36
RFANet_$\{n = 4\}$	55.31±0.67	75.62±0.56	**69.79±0.69**	91.47±0.39
RFANet_$\{n = 3\}$	**55.72±0.66**	**75.96±0.51**	69.76±0.66	**91.52±0.31**
RFANet_$\{n = 2\}$	51.08±0.64	67.89±0.68	65.81±0.59	89.81±0.43

and the accuracy curves of training and validation of the DN4 and the proposed RFANet are shown in Fig. 4. It can be seen from Fig. 3 that the proposed RFANet has better convergence with faster convergence speed and less loss than DN4. Meanwhile, it can be seen from Fig. 4 that the proposed RFANet has a better accuracy than DN4.

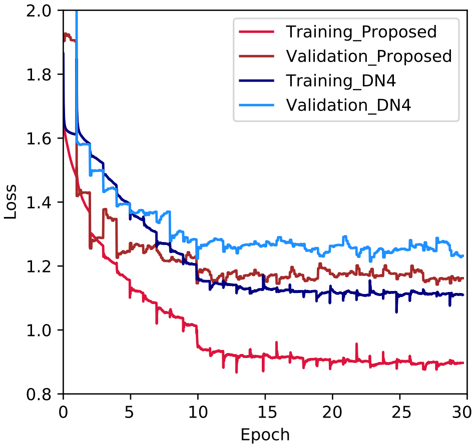

Figure 3: The loss curves of the DN4 and the proposed RFANet for a 5-way 1-shot FSFGIC task on the Stanford Dogs dataset.

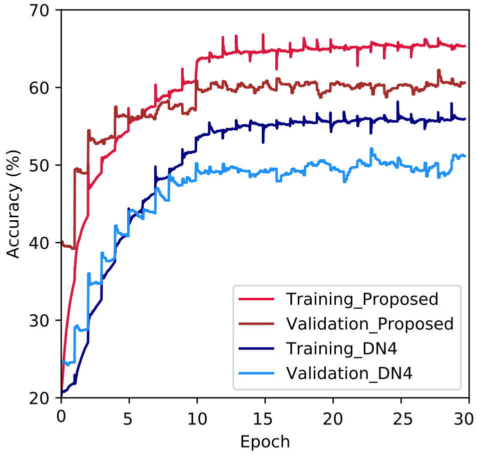

Figure 4: The accuracy curves of the DN4 and the proposed RFANet for a 5-way 1-shot FSFGIC task on the Stanford Dogs dataset.

Runtime. In our implementation, during training for a 5-way 1-shot or 5-way 5-shot FSFGIC task, the average running time for one episode is 0.07 and 0.17 second respectively on a single NVIDIA V100 GPU and a single Intel Xeon Gold 6150 CPU. During testing, the average running time for one episode is 0.03 second. Our approach is efficient in computation for practical applications and can be improved with further parallel implementations.

5 Conclusion

In this paper, we identify and analyze for the first time a fundamental problem in existing metric-learning based FSFGIC methods which fail to effectively address the bias in feature information obtained from each input image that causes misclassification. Then a simple and effective robust feature acquisition network (RFANet) is proposed for FSFGIC. The proposed RFANet architecture address the image quantization noise problem and has the capability to guide any convolution-based embedding models for robustly obtaining feature information from images. Our proposed architecture can be easily embedded into any episodic training mechanisms for end-to-end training from scratch. We have validated the effectiveness of the proposed RFANet on five benchmark datasets which demonstrate the very encouraging great superiority in performance of the proposed method over state-of-the-arts. Furthermore, the proposed RFA mechanism also has a great potential to be applied in object detection, image segmentation, and many other fields.

REFERENCES

Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Chuanqi Dong, Wenbin Li, Jing Huo, Zheng Gu, and Yang Gao. Learning task-aware local representations for few-shot learning. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 716–722, 2020.

Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57 (3):238–247, 1989.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

H. Huang, J. Zhang, J. Zhang, Q. Wu, and J. Xu. Compare more nuanced: Pairwise alignment bilinear network for few-shot fine-grained learning. In *IEEE International Conference on Multimedia and Expo*, pp. 91–96, 2019.

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.

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, 2022.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford Dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop on Fine-Grained Visual Categorization*, volume 2, 2011.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 6418–6428, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *Proceedings of the International Conference on Machine Learning Deep Learning Workshop*, volume 2, 2015.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10657–10665, 2019.

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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7270–7268, 2019a.

Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. Distribution consistency based covariance metric networks for few-shot learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, volume 33, pp. 8642–8649, 2019b.

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.

Yangfan Li and Chunjiang Bian. Few-shot fine-grained ship classification with a foreground-aware feature map reconstruction network. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12, 2022.

Yi Liao, Weichuan Zhang, Yongsheng Gao, Changming Sun, and Xiaohan Yu. ASRSNet: Automatic salient region selection network for few-shot fine-grained image classification. In *Pattern Recognition and Artificial Intelligence*, pp. 627–638, 2022.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

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 international Conference on Computer Vision*, pp. 618–626, 2017.

Saroj Raj Sharma. Plant disease. https://www.kaggle.com/saroz014/plant-disease, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Conference on Neural Information Processing Systems*, pp. 4077–4087, 2017.

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.

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 Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Luming Tang, Davis Wertheimer, and Bharath Hariharan. Revisiting pose-normalization for fine-grained few-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14352–14361, 2020.

Satoshi Tsutsui, Yanwei Fu, and David Crandall. Reinforcing generated images via meta-learning for one-shot fine-grained visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Conference on Neural Information Processing Systems*, 29:3630–3638, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The CalTech-UCSD Birds-200-2011 dataset. *California Institute of Technology*, 2011.

Chaofei Wang, Shiji Song, Qisen Yang, Xiang Li, and Gao Huang. Fine-grained few shot learning with foreground object transformation. *Neurocomputing*, 466:16–26, 2021.

Xiu-Shen Wei, Peng Wang, Lingqiao Liu, Chunhua Shen, and Jianxin Wu. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE Transactions on Image Processing*, 28(12):6116–6125, 2019.

Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8012–8021, 2021.

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 International Conference on Computer Vision*, pp. 8812–8821, 2021.

Shuo Yang, Songhua Wu, Tongliang Liu, and Min Xu. Bridging the gap between few-shot and many-shot learning via distribution calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

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 Conference on Computer Vision and Pattern Recognition*, pp. 12203–12213, 2020.

Hongguang Zhang and Piotr Koniusz. Power normalizing second-order similarity network for few-shot learning. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1185–1193, 2019.

Yaohui Zhu, Chenlong Liu, and Shuqiang Jiang. Multi-attention meta learning for few-shot fine-grained image recognition. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pp. 1090–1096, 2020.

APPENDIX

6 MORE EXPERIMENTAL RESULTS

Additional experiments are conducted for investigating the effect of different DA operation. Three different operations (i.e., random cropping, random horizontal flipping, and color jittering) are selected individually for performing 5-way 1-shot and 5-way 5-shot FSFGIC tasks respectively on the Stanford Cars dataset. Their results are shown in Table 6. It is observed that the proposed RFANet is more effective when applied on random cropping than on flipping and jittering.

Table 6: Comparison results on the Stanford Cars dataset based on Conv64F.

Model	Backbone	20-way Accuracy (%)	
		1-shot	5-shot
RFANet$_{\text{random cropping}}$	Conv64F	**68.19±0.67**	**89.38±0.36**
RFANet$_{\text{random horizontal flipping}}$	Conv64F	67.63±0.67	86.77±0.35
RFANet$_{\text{color jittering}}$	Conv64F	64.60±0.66	87.59±0.39

The proposed RFANet has also been compared with DN4 on the Stanford Cars dataset for 20-way 1-shot and 20-way 5-shot FSFGIC tasks. 127 classes, 20 classes, and 49 classes of Stanford Cars dataset are randomly selected for training, validation, and testing respectively. Their classification accuracies are summarized in Table 7, consistently demonstrating the superior performance of the proposed method when class number increases.

Table 7: Comparison results on the Stanford Cars dataset based on Conv64F.

Model	Backbone	20-way Accuracy (%)	
		1-shot	5-shot
DN4 (Li et al., 2019a)	Conv64F	37.67±0.59	69.58±0.34
RFANet	Conv64F	**48.55±0.46**	**77.65±0.31**

Based on ResNet12 (He et al., 2016) as the backbone, the proposed RFANet is compared with nine methods (i.e., Baseline (Chen et al., 2019), Baseline++ (Chen et al., 2019), MatchingNet (Vinyals et al., 2016), ProtoNet (Snell et al., 2017), RelationNet (Sung et al., 2018), MetaOptNet (Lee et al., 2019), BSNet (Li et al., 2020), DeepEMD (Zhang et al., 2020), and FRN (Wertheimer et al., 2021)) on the CUB dataset. Following the same data split as used in (Zhang et al., 2020; Wertheimer et al., 2021), 100 classes, 50 classes, and 50 classes of CUB dataset are randomly selected for training, validation, and testing respectively.

In this way, ResNet12 is employed with the same implementation as in (Zhang et al., 2020; Wertheimer et al., 2021). The classification accuracy on 5-way 1-shot and 5-way 5-shot FSFGIC tasks for each method is summarized in Table 8. It is observed that the proposed RFANet achieves the best performance on 5-way 1-shot and 5-way 5-shot FSFGIC tasks.

Table 8: Comparison results on the CUB dataset based on ResNet12.

Model	Backbone	5-way Accuracy (%)	
		1-shot	5-shot
Baseline (Chen et al., 2019)	ResNet12	63.90±0.88	82.54±0.54
Baseline++ (Chen et al., 2019)	ResNet12	68.46±0.85	81.02±0.46
MatchingNet (Vinyals et al., 2016)	ResNet12	72.62±0.90	84.14±0.50
ProtoNet (Snell et al., 2017)	ResNet12	71.57±0.89	86.37±0.49
RelationNet (Sung et al., 2018)	ResNet12	70.20±0.84	84.28±0.46
MetaOptNet (Lee et al., 2019)	ResNet12	75.15±0.46	87.09±0.30
BSNet (Li et al., 2020)	ResNet12	69.73±0.97	82.85±0.61
DeepEMD (Zhang et al., 2020)	ResNet12	75.65±0.83	88.69±0.50
FRN (Wertheimer et al., 2021)	ResNet12	78.07±0.21	89.33±0.12
RFANet	ResNet12	**78.33±0.67**	**91.61±0.34**