Cross-Layer and Cross-Sample Feature Optimization Network for Few-Shot Fine-Grained Image Classification

Zhen-Xiang Ma, Zhen-Duo Chen*, Li-Jun Zhao, Zi-Chao Zhang, Xin Luo, Xin-Shun Xu

School of Software, Shandong University, Jinan, China
{mazhenxiang0923, zhangzichao1008, lj_zhao1028}@163.com, {chenzd.sdu, luoxin.lxin}@gmail.com, xuxinshun@sdu.edu.cn

Abstract

Recently, a number of Few-Shot Fine-Grained Image Classification (FS-FGIC) methods have been proposed, but they primarily focus on better fine-grained feature extraction while overlooking two important issues. The first one is how to extract discriminative features for Fine-Grained Image Classification tasks while reducing trivial and non-generalizable sample-level noise introduced in this procedure, to overcome the over-fitting problem under the setting of Few-Shot Learning. The second one is how to achieve satisfying feature matching between limited support and query samples with variable spatial positions and angles. To address these issues, we propose a novel Cross-layer and Cross-sample feature optimization Network for FS-FGIC, C2-Net for short. The proposed method consists of two main modules: Cross-Layer Feature Refinement (CLFR) module and Cross-Sample Feature Adjustment (CSFA) module. The CLFR module further refines the extracted features while integrating outputs from multiple layers to suppress sample-level feature noise interference. Additionally, the CSFA module addresses the feature mismatch between query and support samples through both channel activation and position matching operations. Extensive experiments have been conducted on five fine-grained benchmark datasets, and the results show that the C2-Net outperforms other state-of-the-art methods by a significant margin in most cases. Our code is available at: https://github.com/zenith0923/C2-Net.

Introduction

With the ongoing advancement of deep learning (He et al. 2016; Du et al. 2023; Yang et al. 2022b), Fine-Grained Image Classification (FGIC) (Fu, Zheng, and Mei 2017) has made significant progress under the condition of having sufficient training samples. However, the cost of collecting large-scale label-rich fine-grained images is expensive, so many researchers have shifted their focus to Few-Shot Learning (FSL) (Munkhdalai et al. 2018; Vinyals et al. 2016) to avoid this issue. Built based on meta-learning, metric learning, and some other strategies, FSL methods aim to learn task-transferable knowledge according to related base classes, to reduce over-fitting and achieve satisfying classification results on testing (query) samples belonging to novel classes with extremely limited training (support) samples. By combining the idea of FSL with the FGIC task, several effective methods have been proposed recently for Few-Shot Fine-Grained Image Classification (FS-FGIC), but there are still several issues to be further addressed in this new task.

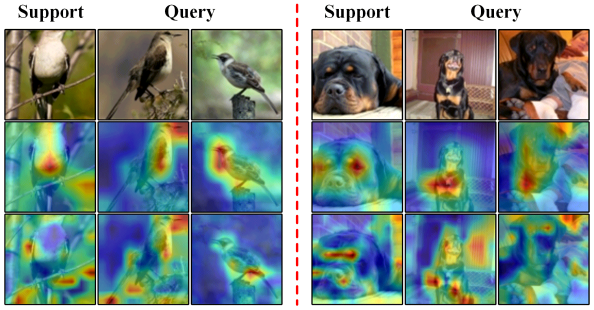

Figure 1: Visualization results of features extracted by the backbone on the CUB and Stanford-Dogs datasets. The first row shows the raw images, the second row displays the visualization of features extracted from the last layer of the backbone, and the third row shows the visualization of features extracted from the penultimate layer.

Generally speaking, feature learning is the most critical point to fine-grained related tasks. However, there exists a potential contradiction in the design of the feature learning module for the FS-FGIC task. Specifically, on the one hand, the task of FGIC requires that the model can obtain sufficient detailed information to achieve adequate fine-grained discriminating power. On the other hand, the FSL task requires the model to effectively learn semantic concepts at the category level and identify inter-class differences from extremely limited training data. Although exploring and incorporating more detailed information can improve the fine-grained discrimination capability of models, it inevitably leads to the expansion of the final feature space and introduces more trivial and non-generalizable sample-level noise, which will exacerbate the over-fitting problem under the FSL setting. For example, as shown in Fig. 1, although combining feature maps from multiple layers is a well-tested strategy for FGIC methods (Du et al. 2020; Yang et al. 2022a) because the features extracted from different layers of the backbone are complementary, it also introduces

*Corresponding Author

more sample-level and background noise. Inheriting the traditional idea of the FGIC methods, current FS-FGIC methods (Huang et al. 2021; Xu et al. 2022; Wang, Fu, and Ma 2022) generally tend to focus on exploring more local information to enhance the discrimination ability of the model, while neglecting this potential contradiction between FSL and FGIC task, which may limit their applicability to addressing the FS-FGIC problem. In this paper, we argue that one of the keys to designing a promising FS-FGIC model lies in striking a balance between the requirements of both FSL and FGIC tasks, rather than simply mining and stacking fine-grained features. In other words, it means the model should not only effectively learn and integrate sufficient fine-grained features, but also be able to further refine the extracted features by discarding or reducing trivial and non-generalizable information.

After the feature learning procedure, because of the extremely limited number of training instances under the FSL setting, there is also a feature mismatch issue to be considered for fine-grained samples collected from real-world environments. As illustrated in Fig. 1, the key discriminative regions in support and query images can be mismatched in both semantic information ($e.g.$, the presence of the body of a dog or the head of a bird in the query, but not in the support sample) and spatial position ($e.g.$, the head of a dog, the wings of a bird). Under the setting of FSL, there are usually no sufficient training instances for a specific category to ensure the model's generalization ability on the target task. As a result, the negative impact caused by this problem will be much more severe in the FSL task than that in conventional classification tasks, especially for the FS-FGIC task that deals with highly similar fine-grained categories. Therefore, to reduce performance degradation resulting from the information loss during support-query instance matching, an additional mechanism is necessary to adjust the extracted features and ensure alignment between input samples from the perspective of information and position.

Based on the above discussion, we propose a novel Cross-layer and Cross-sample feature optimization Network (C2-Net) for FS-FGIC, to achieve effective feature learning and accurate classification. In response to the aforementioned two issues, the C2-Net consists of two main modules, $i.e.$, the Cross-Layer Feature Refinement (CLFR) module as well as the Cross-Sample Feature Adjustment (CSFA) module. The CLFR module aims to overcome the contradictions between FSL and FGIC tasks with an extracting-and-refining procedure. Specifically, to meet the requirements of the FGIC task, the CLFR module draws on a typical fine-grained feature learning strategy, which integrates information learned from multiple backbone network layers to combine both mid-level features and high-level semantic concepts for a comprehensive feature representation. However, in contrast to traditional methods that simply aggregate extracted features for final classification, the CLFR module goes a further step to refine extracted features with a reconstructing procedure based on cross-layer feature correlation matrices. While maintaining fine-grained discriminative power, this procedure can significantly suppress noise and non-generalizable sample-level characteristics, leading to the reduction of over-fitting indirectly under the setting of FSL. Subsequently, to address the issue of feature mismatch between support and query samples caused by diverse factors, such as position and angle variations, the CSFA module incorporates a channel activation operation that adaptively re-calibrates channel-wise features based on trainable channel weights to enhance shared information and suppress unshared query information, as well as a position matching operation to adjust the position of targets by learning a dynamic adjustment matrix,

Contributions of this paper can be summarized as follows:

- By analyzing the FSL and FGIC task, we raise two critical issues that are needed to be addressed in the design of the FS-FGIC methods. The first one is how to extract fine-grained features while reducing trivial and non-generalizable sample-level noise. The second one is addressing the feature mismatch between query and support samples caused by positions and angles. Then we propose a novel C2-Net to address these issues.
- To overcome the feature learning contradiction between FSL and FGIC tasks, we design the CLFR module to integrate feature maps from multiple network layers, and further refine extracted features with a cross-layer feature correlation-based reconstructing procedure.
- To address the issue of feature mismatch between support and query samples, we propose the CSFA module to improve sample match results by adjusting the query features from both channel and position perspectives.
- We conduct extensive experiments and analyses on five widely-used fine-grained benchmarks, and experimental results demonstrate the superiority of the C2-Net.

Related Work

Few-Shot Image Classification

Recent Few-Shot Image Classification methods can be roughly categorized into three groups: meta-learning based methods (Finn, Abbeel, and Levine 2017; Lee et al. 2019), metric-learning based methods (Snell, Swersky, and Zemel 2017; Sung et al. 2018; Vinyals et al. 2016), and data-augmentation based methods (Hariharan and Girshick 2017; Chen et al. 2019b; Tang et al. 2020). Meta-learning based methods try to find a suitable gradient-based optimization strategy that can quickly adapt to new tasks with few gradient updates. For example, MAML (Finn, Abbeel, and Levine 2017) aims to learn a parameter initialization that can be easily adapted to a new task with few gradient updates. Metric-learning based methods try to learn a generalizable embedding and a suitable metric function that can measure the similarity of two samples. The most representative one is ProtoNet (Snell, Swersky, and Zemel 2017), which computes prototypes as the mean feature of each class in the support and measures the distance from a query to each class using them. FRN (Wertheimer, Tang, and Hariharan 2021) reconstructs the query features using the support feature pools to calculate their similarity, which achieves more performant efficient. Data-augmentation based methods try to learn a generator from base classes and use it to generate novel samples or features for data augmentation.

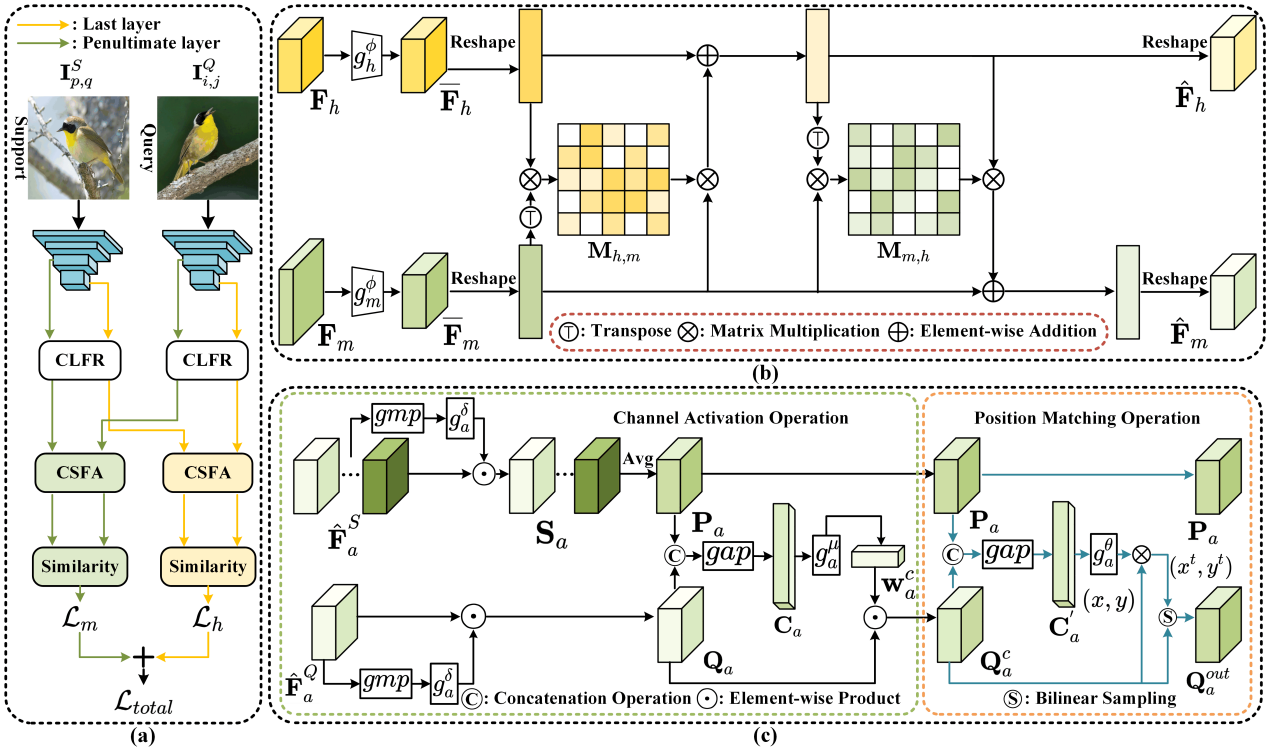

Figure 2: The proposed C2-Net. (a) is the overall of C2-Net, consisting of the backbone network, the Cross-Layer Feature Refinement (CLFR) module, the Cross-Sample Feature Adjustment (CSFA) module, and the Similarity module; (b) is the schemetic illustration of the CLFR module; (c) is the schemetic illustration of the CSFA module.

Few-Shot Fine-grained Image Classification

FS-FGIC (Li et al. 2021; Wu et al. 2023) is a challenging problem that requires more effective feature learning power than traditional FSL task. (Wei et al. 2019) first define the few-shot fine-grained recognition task and use bilinear features to learn a piecewise mapping classifier. Thereafter, PoseNorm (Tang, Wertheimer, and Hariharan 2020) explores the effect of part annotations and shows that learning part features can significantly improve the performance of few-shot learning methods on the fine-grained dataset. MattML (Zhu, Liu, and Jiang 2020) uses a task embedding network to automatically learn a task-specific initialization with attention mechanisms. (Xu et al. 2022) uses spatial attention to capture the fine-grained details of the object and channel attention to capture the global context of the image. HelixFormer (Zhang et al. 2022) is a transformer-based double-helix model that solves the FS-FGIC task by learning cross-image object semantic relations in local regions of images. BiFRN (Wu et al. 2023) utilizes a bidirectional reconstruction process to increase inter-class variations and decrease intra-class variations.

In addition, AGPF (Tang et al. 2022) tries to take advantage of multiple network layer fusion, but it ignores the negative impact resulting from such operation under the setting of FSL. In contrast, we further refine the extracted features while integrating multi-layer outputs, thereby suppressing the interference of instance-level feature noise and achieving better performance. OLSA (Wu et al. 2021) aims to align the spatial features of the object and learns the long-range semantic correspondence across different tasks. However, this alignment method only involves spatial adjustment and can damage the integrity of extracted features, which may exacerbate the over-fitting problem caused by sample-level features. The C2-Net adjusts extracted and refined features as a complete object, achieving optimized feature matching between query and support samples through both channel activation and position matching operations.

Method

Problem Definition and Overall Framework

The objective of FSL is to acquire transferable knowledge from the base classes D_{base} so that the model can perform well in the novel classes D_{novel} with the help of limited labeled samples, where $D_{base} \cap D_{novel} = \emptyset$. In accordance with the standard setting of FSL, we adopt the episodic training strategy. During the meta-training phase, each episode can be referred to as an "N-way K-shot" classification problem, where N classes are randomly chosen from D_{base}. Each class in the support set S contains K labeled samples, and each class in the query set Q contains U unlabeled samples. During the meta-testing phase, the model is evaluated on the novel classes D_{novel} using the knowledge acquired from the meta-training phase. This evaluation helps measure the model's generalization capability to the novel

classes. Without loss of generality, we assume the following discussion is within one episode.

As illustrated in Fig. 2(a), given support image $\mathbf{I}^S_{p,q}$, where $p \in \{1, ..., N\}$ and $q \in \{1, ..., K\}$, and query image $\mathbf{I}^Q_{i,j}$, where $i \in \{1, ..., N\}$ and $j \in \{1, ..., U\}$, the backbone network $B(\cdot)$ first takes them as input, and feature maps ($\mathbf{F}^S_{p,q,m}$, $\mathbf{F}^S_{p,q,h}$, $\mathbf{F}^Q_{p,q,m}$, and $\mathbf{F}^Q_{p,q,h}$) from last L layers will be obtained, where L=2 in this paper in consideration of performance and model complexity. Then the CLFR module integrates and refines the two feature maps extracted from each image with a feature reconstruction procedure. Thereafter, the CSFA module adjusts the refined feature maps in both channel and position perspectives for better support-query matching. Finally, the Similarity module calculates the distances between query and support samples, the mean of two distances is used for query sample classification.

Cross-Layer Feature Refinement

Given the insufficient discriminative power of the final layer output of the network in describing fine-grained objects, the strategy of combining features extracted from multiple network layers has been adopted by several existing FGIC methods (Du et al. 2020). This strategy can integrate both high-level semantic information and mid-level detailed characteristics to comprehensively describe and identify fine-grained objects, and it is also easier to implement and more flexible than spacial localization-based fine-grained feature learning. However, mid-level features usually introduce more trivial and non-generalizable noise as illustrated in Fig. 1, and integrating more information also results in the expansion of feature space. As mentioned in the Introduction, this is unfavorable for the FSL task. To solve this problem, we design the Cross-Layer Feature Refinement (CLFR) module, whose detail is illustrated in Fig. 2(b), to reconstruct feature maps based on cross-layer feature correlation.

Specifically, given an image $\mathbf{I}$ (superscript and subscript are omitted for clarity), inputs to the CLFR module can be obtained by

$$\mathbf{F}_h, \mathbf{F}_m = B(\mathbf{I}), \tag{1}$$

where $\mathbf{F}_h \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_m \in \mathbb{R}^{C'\times H'\times W'}$ are feature maps obtained from the last and penultimate layer. Because the size of $\mathbf{F}_h$ and $\mathbf{F}_m$ are different, two sub-networks g^ϕ_h and g^ϕ_m, both of which consist of two convolutional layers, are applied to transfer these feature maps into the same size, which can be formulated as

$$\overline{\mathbf{F}}_h = g^\phi_h(\mathbf{F}_h) \in \mathbb{R}^{C\times H\times W}, \tag{2}$$

$$\overline{\mathbf{F}}_m = g^\phi_m(\mathbf{F}_m) \in \mathbb{R}^{C\times H\times W}. \tag{3}$$

Thereafter, the feature map $\overline{\mathbf{F}}_h$ and $\overline{\mathbf{F}}_m$ are reshaped into 2-D matrices with the size of $C \times HW$, and then a cross-layer feature correlation matrix can be obtained as follows,

$$\mathbf{M}_{h,m} = tanh\Big(\Big(\frac{\overline{\mathbf{F}}_m}{||\overline{\mathbf{F}}_m||_2}\Big)^T \frac{\overline{\mathbf{F}}_h}{||\overline{\mathbf{F}}_h||_2}\Big) \in \mathbb{R}^{HW\times HW}, \tag{4}$$

where $||\cdot||_2$ represents the L2-norm. Based on $\mathbf{M}_{h,m}$, the feature map obtained from the last layer could be further reconstructed and refined, which can be formulated as

$$\hat{\mathbf{F}}_h = (\overline{\mathbf{F}}_m \mathbf{M}_{h,m}) + \overline{\mathbf{F}}_h. \tag{5}$$

Afterward, a similar process can be used to further refine the feature map obtained from the penultimate layer as follows,

$$\mathbf{M}_{m,h} = tanh\Big(\Big(\frac{\hat{\mathbf{F}}_h}{||\hat{\mathbf{F}}_h||_2}\Big)^T \frac{\overline{\mathbf{F}}_m}{||\overline{\mathbf{F}}_m||_2}\Big), \tag{6}$$

$$\hat{\mathbf{F}}_m = (\hat{\mathbf{F}}_h \mathbf{M}_{m,h}) + \overline{\mathbf{F}}_m. \tag{7}$$

Following the above procedure, given support samples $\mathbf{I}^S_{p,q}$ and query samples $\mathbf{I}^Q_{i,j}$, corresponding refined feature maps $\hat{\mathbf{F}}^S_{p,q,h}$, $\hat{\mathbf{F}}^S_{p,q,m}$, $\hat{\mathbf{F}}^Q_{i,j,h}$, and $\hat{\mathbf{F}}^Q_{i,j,m}$ can be obtained. They are reshaped back to the size of $C \times H \times W$ for the next module.

Through the above procedure, the features shared by two input feature maps can be highlighted, and the influence of unshared features can be reduced, resulting in two main advantages. Firstly, noise caused by background or sample-level trivial characteristics typically exists primarily in mid-level feature maps. Therefore, this reconstruction operation can significantly suppress such noise. Secondly, if features are shared by both high-level and mid-level feature maps, it also means that they are capable of describing both the overall semantics and the detailed characteristics of the target object. The above feature refining procedure enhances such features while suppressing the rest, which can achieve the compression of the feature space while maximizing the retention of the fine-grained discriminative capability.

Cross-Sample Feature Adjustment

After suppressing or removing unimportant features with the CLFR module, the Cross-Sample Feature Adjustment (CSFA) module is designed to reduce critical information loss caused by spatial diversity and enhance the consistency between samples belonging to the same category during support-query matching. For the sake of clarity, in this subsection, we simplify the symbols of refined feature maps ($\hat{\mathbf{F}}^S_{p,q,h}$, $\hat{\mathbf{F}}^S_{p,q,m}$, $\hat{\mathbf{F}}^Q_{i,j,h}$, and $\hat{\mathbf{F}}^Q_{i,j,m}$) to $\hat{\mathbf{F}}^S_a$ and $\hat{\mathbf{F}}^Q_a$ unless necessary, where $a = \{h, m\}$, because feature maps related to different layers and each query-support pair are processed in the same way in this module. The overall architecture of the CSFA module is illustrated in Fig. 2(c), the CSFA module incorporates two sequential operations for channel weights adjustment and feature positions adjustment.

Channel Activation Operation In the beginning, given refined feature maps generated by the CLFR module, they are first performed intra-sample channel weighting, which serves as a complement to the previous module, in order to better activate discriminative features while suppressing potential background noise that might still be present. Specifically, two sub-networks consisting of two fully connected layers, *i.e.*, g^δ_a and $a = \{h, m\}$, are applied to generated channel-wise attention and the input feature maps will be re-weighted accordingly. This can be formulated as

$$\begin{aligned} \mathbf{S}_a &= \hat{\mathbf{F}}^S_a \odot g^\delta_a(gmp(\hat{\mathbf{F}}^S_a)), \\ \mathbf{Q}_a &= \hat{\mathbf{F}}^Q_a \odot g^\delta_a(gmp(\hat{\mathbf{F}}^Q_a)), \end{aligned} \tag{8}$$

Methods	Backbone	Stanford-Dogs		Stanford-Cars	
		1-shot	5-shot	1-shot	5-shot
DN4 (CVPR-19)	Conv-4	45.41 ± 0.76	63.51 ± 0.62	59.84 ± 0.80	88.65 ± 0.44
CovaMNet (AAAI-19)	Conv-4	49.10 ± 0.76	63.04 ± 0.65	56.65 ± 0.86	71.33 ± 0.62
MattML (IJCAI-20)	Conv-4	54.84 ± 0.53	71.34 ± 0.38	66.11 ± 0.54	82.80 ± 0.28
ATL-Net (IJCAI-20)	Conv-4	54.49 ± 0.92	73.20 ± 0.69	67.95 ± 0.84	89.16 ± 0.48
BSNet (TIP-21)	Conv-4	43.13 ± 0.85	62.61 ± 0.73	44.56 ± 0.83	63.72 ± 0.78
LRPABN (TMM-21)	Conv-4	45.72 ± 0.75	60.94 ± 0.66	60.28 ± 0.76	73.29 ± 0.58
TOAN (TCSVT-21)	Conv-4	49.30 ± 0.77	67.16 ± 0.49	65.90 ± 0.72	84.24 ± 0.48
OLSA (MM-21)	Conv-4	55.53 ± 0.45	71.68 ± 0.36	70.13 ± 0.48	84.29 ± 0.31
DAN (AAAI-22)	Conv-4	59.81 ± 0.50	77.19 ± 0.35	70.21 ± 0.50	85.55 ± 0.31
AGPF (PR-22)	Conv-4	60.89 ± 0.98	78.14 ± 0.62	78.14 ± 0.84	87.42 ± 0.57
PaCL (MM-22)	Conv-4	59.76 ± 0.70	77.50 ± 0.48	72.21 ± 0.68	88.02 ± 0.36
HelixFormer (MM-22)	Conv-4	59.81 ± 0.50	73.40 ± 0.36	75.46 ± 0.37	89.68 ± 0.25
BiFRN (AAAI-23)	Conv-4	61.39 ± 0.23	78.86 ± 0.15	76.22 ± 0.20	90.66 ± 0.11
Ours	Conv-4	**66.42 ± 0.50**	**81.23 ± 0.34**	**81.29 ± 0.45**	**91.08 ± 0.26**
BSNet (TIP-21)	ResNet-18	-	-	60.36 ± 0.98	85.28 ± 0.64
OLSA (MM-21)	ResNet-12	64.15 ± 0.49	78.28 ± 0.32	77.03 ± 0.46	88.85 ± 0.46
HelixFormer (MM-22)	ResNet-12	65.92 ± 0.49	80.65 ± 0.36	79.40 ± 0.43	92.26 ± 0.15
BiFRN (AAAI-23)	ResNet-12	72.54 ± 0.22	85.86 ± 0.13	88.43 ± 0.17	**96.34 ± 0.07**
Ours	ResNet-12	**75.50 ± 0.49**	**87.65 ± 0.28**	**88.96 ± 0.37**	95.16 ± 0.20

Table 1: 5-way classification accuracy (%) on the Stanford Dogs and Stanford Cars datasets. The highest results are highlighted, while the second-highest results are underlined.

where $gmp(\cdot)$ is global max pooling and $\odot$ denotes element-wise product operation with the broadcasting mechanism. In addition, to improve efficiency, we follow ProtoNet to generate prototypes for each class to replace original support samples for final sample matching (classification). Specifically, the prototype of p-th class can be obtained by

$$\mathbf{P}_{p,a} = \frac{1}{K}\sum_{q=1}^{K}\mathbf{S}_{p,q,a}. \tag{9}$$

Similarly, the above output can be also simplified to $\mathbf{P}_a$.

To reduce the influence of semantic information mismatch resulting from the angle or some other factors, we further address this issue from a cross-sample perspective. Firstly, a query sample $\mathbf{Q}_a$ and a support prototype $\mathbf{P}_a$ are concatenated into a cross-sample representation as follows,

$$\mathbf{C}_a = gap(\mathbf{P}_a||\mathbf{Q}_a) \in \mathbb{R}^{2C\times 1\times 1}, \tag{10}$$

where $gap(\cdot)$ is global average pooling and $||$ represents the concatenation operation.

Thereafter, a group of trainable channel weights are used to adaptively modify the feature maps of the query samples, activating shared features between query samples and corresponding prototypes while suppressing query-specific features. Specifically, we use two fully connected layers g_a^{μ} to generate channel weights $\mathbf{w}_a^c \in \mathbb{R}^{C\times 1\times 1}$, then, the calibrated feature $\mathbf{Q}_a^c \in \mathbb{R}^{C\times H\times W}$ can be computed as follows,

$$\mathbf{w}_a^c = 1 + tanh(g_a^{\mu}(\mathbf{C}_a)), \tag{11}$$

$$\mathbf{Q}_a^c = \mathbf{Q}_a \odot \mathbf{w}_a^c, \tag{12}$$

where $\odot$ denotes element-wise product operation with the broadcasting mechanism. To some extent, this operation can also contribute to reducing the over-fitting problem by feature space compressing.

Position Matching Operation As the name indicates, this operation aims to achieve feature alignment between the crucial feature regions in the query $\mathbf{Q}_a^c$ and the support prototype $\mathbf{P}_a$ by adjusting the query feature map. Specifically, a feature matching network g_a^{θ}, which includes two fully connected layers, takes the cross-sample representation $\mathbf{C}_a^{'} = gap(\mathbf{P}_a||\mathbf{Q}_a^c)$ as input and generates position matching matrices $\mathbf{M}_a^{\theta}$ and $\mathbf{M}_a^{\Delta}$ as follows,

$$\mathbf{M}_a^{\theta},\ \mathbf{M}_a^{\Delta} = g_a^{\theta}(\mathbf{C}_a^{'}) = \{\theta_1,\ \theta_2\},\ \{\Delta x,\ \Delta y\}, \tag{13}$$

Afterward, generate a regular grid with the same size as $\mathbf{Q}_a^c$ within a range of $[-1, 1]^{2\times H\times W}$, where the number 2 represents the x and y coordinates respectively. Next, the position matching matrices $\mathbf{M}_a^{\theta}$ and $\mathbf{M}_a^{\Delta}$ are applied to the original coordinates (x, y) to obtain the adjusted coordinates (x^t, y^t) as follows,

$$(x^t, y^t)^T = \mathbf{M}_a^{\theta}(x, y)^T + \mathbf{M}_a^{\Delta}. \tag{14}$$

Finally, use bilinear sampling to map the original features $\mathbf{Q}_a^c$ to the adjusted coordinates (x^t, y^t), resulting in the position matching feature $\mathbf{Q}_a^{out} \in \mathbb{R}^{C\times H\times W}$. The output $\mathbf{Q}_a^{out}$ is obtained by adjusting the features in both channel and position perspectives to enhance consistency among sample features, thereby alleviating the issues of both semantic information and position features mismatch between support and query samples.

To sum up, given the feature maps ($\hat{\mathbf{F}}_{p,q,h}^S$, $\hat{\mathbf{F}}_{p,q,m}^S$, $\hat{\mathbf{F}}_{i,j,h}^Q$, and $\hat{\mathbf{F}}_{i,j,m}^Q$), the corresponding support prototypes ($\mathbf{P}_{p,h}$ and $\mathbf{P}_{p,m}$) and adjusted query feature maps ($\mathbf{Q}_{i,j,h}^{out}$ and $\mathbf{Q}_{i,j,m}^{out}$) can be obtained through the CSFA module.

Methods	1-shot	5-shot
Closer (ICLR-19)	60.53	79.34
SAML (ICCV-19)	69.35	81.37
DN4 (CVPR-19)	64.02 ± 0.92	82.97 ± 0.66
CovaMNet (AAAI-19)	58.87 ± 1.00	70.46 ± 0.84
MattML (IJCAI-20)	66.29 ± 0.56	80.34 ± 0.30
PoseNorm (CVPR-20)	64.17	81.96
FEAT (CVPR-20)	68.87	82.90
BSNet (TIP-21)	55.81 ± 0.97	76.34 ± 0.65
FRN (CVPR-21)	69.45 ± 0.22	85.16 ± 0.14
OLSA (MM-21)	73.07 ± 0.46	86.24 ± 0.29
DAN (AAAI-22)	72.89 ± 0.50	86.60 ± 0.31
FRN+TDM (CVPR-22)	71.37 ± 0.22	86.45 ± 0.14
AGPF (PR-22)	74.03 ± 0.90	86.54 ± 0.50
PaCL (MM-22)	74.04 ± 0.70	88.75 ± 0.38
BiFRN (AAAI-23)	74.36 ± 0.20	88.64 ± 0.10
Ours	**78.66 ± 0.46**	**89.43 ± 0.28**

Table 2: 5-way classification accuracy (%) using Conv-4 backbone on the CUB (using raw images) dataset. The highest results are highlighted, while the second-highest results are underlined.

Overall Objectives

The meta-training process: In a "*N*-way *K*-shot" training episodic, the loss function for features from each layer is formulated as follows,

$$\mathcal{L}_a = -\frac{1}{N}\frac{1}{U}\sum_{i=1}^{N}\sum_{j=1}^{U} log\frac{\exp\left(-\tau_a d\big(\mathbf{Q}_{i,j,a}^{out}, \mathbf{P}_{i,a}\big)\right)}{\sum_{p=1}^{N}\exp\left(-\tau_a d\big(\mathbf{Q}_{i,j,a}^{out}, \mathbf{P}_{p,a}\big)\right)}, \tag{15}$$

where $a = \{h, m\}$, τ_a is learnable scaling parameter and $d(\cdot)$ represents the similarity scores between queries and support prototypes measured by Euclidean distance. The overall objective function to train the C2-Net can be expressed as follows,

$$\mathcal{L}_{total} = \alpha\mathcal{L}_h + (1-\alpha)\mathcal{L}_m, \tag{16}$$

where α is a balancing hyper-parameter within a range of $[0, 1]$.

The meta-validation/testing process: After the meta-training process, given an unlabeled query sample $\mathbf{I}_q$ and support samples in a "*N*-way *K*-shot" testing episodic, this query sample can be classified as the i-th class as follows,

$$arg\ min_i \frac{1}{2}\left(d\big(\mathbf{Q}_m^{out}, \mathbf{P}_{i,m}\big) + d\big(\mathbf{Q}_h^{out}, \mathbf{P}_{i,h}\big)\right). \tag{17}$$

where $\mathbf{Q}_m^{out}$ and $\mathbf{Q}_h^{out}$ are the obtained features given $\mathbf{I}_q$, and $\mathbf{P}_{i,m}$ and $\mathbf{P}_{i,h}$ are feature maps corresponding to prototypes of support samples.

Experiments

Dataset

Stanford Dogs (Khosla et al. 2011) comprises 20,580 images and 120 classes of dogs. Following (Zhu, Liu, and Jiang 2020), we use 70, 20, and 30 classes for meta-training, meta-validation, and meta-testing, respectively.

Methods	meta-iNat		tiered meta-iNat	
	1-shot	5-shot	1-shot	5-shot
ProtoNet (NIPS-17)	53.78	73.80	35.47	54.85
Covar. Pool (CVPR-19)	57.15	77.20	36.06	57.48
DN4 (CVPR-19)	62.32	79.76	43.82	64.17
DSN (CVPR-20)	58.08	77.38	36.82	60.11
CTX (NeurIPS-20)	60.03	78.80	36.83	60.84
DeepEMD (CVPR-20)	54.48	68.36	36.05	48.55
FRN (CVPR-21)	61.98	80.04	43.95	63.45
FRN+TDM (CVPR-22)	63.97	81.60	44.05	62.91
MCL (CVPR-22)	64.66	81.31	44.08	64.61
Ours	**71.47**	**85.47**	**49.04**	**67.25**

Table 3: 5-way classification accuracy (%) using Conv-4 backbone on the meta-iNat and tiered meta-iNat datasets. The highest results are highlighted, while the second-highest results are underlined.

Stanford Cars (Krause et al. 2013) has 16,185 images from 196 classes of cars. We follow the dataset split introduced in (Zhu, Liu, and Jiang 2020), which employs 130, 17, and 49 classes for meta-training, meta-validation, and meta-testing.

CUB-200-2011 (Wah et al. 2011) is a dataset of 11,788 images of birds from 200 classes. We use the same class split as (Wertheimer, Tang, and Hariharan 2021), which uses 100, 50, and 50 classes for meta-training, meta-validation, and meta-testing. It should be emphasized that we use raw images without annotated bounding boxes.

meta-iNat (Horn et al. 2018; Wertheimer and Hariharan 2019) which has 13 super categories and 1,135 species. The dataset is split into 908 training classes and 227 testing classes according to (Wertheimer and Hariharan 2019).

tiered meta-iNat (Wertheimer and Hariharan 2019) is a more challenging variant of meta-iNat. The dataset comprises 781 classes for training and 354 classes for testing.

Implementation Details

We adopt widely used two backbone networks in few-shot classification tasks: Conv-4 and ResNet-12, which are consistent with the common protocols. Following convention settings, the input images are resized to 84×84. We utilize the same standard data augmentation techniques, including random crop, horizontal flip, and color jitter as existing methods. During the meta-training stage, we use SGD with a Nesterov momentum of 0.9. The initial learning rate is set to 0.1 and weight decay is set to 5e-4. The value of α is set to 0.5. The entire process lasts for 150 epochs, and the learning rate decays to 0.01 and 0.001 at the 70 and 110 epochs, respectively. During the meta-testing stage, we apply the standard 5-way 1-shot and 5-way 5-shot settings, using 15 query images per class in both settings. We test on the best-performing model on the validation set and report the 95% confidence interval results for 2,000 test episodes.

Experimental Results

In this section, we compare our proposed C2-Net with a number of state-of-the-art methods, experimental results are summarized in Table 1-3.

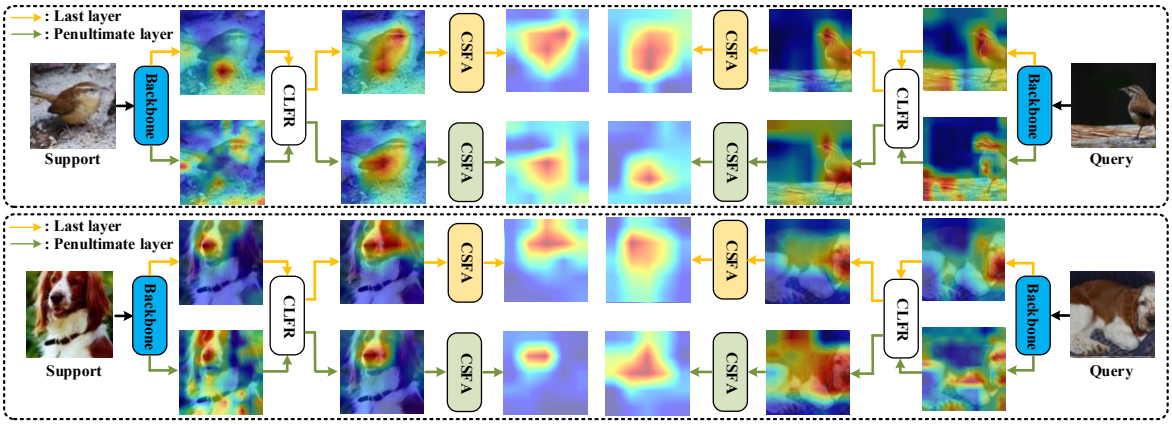

Figure 3: Visualization results of features extracted by Backbone, CLFR, and CSFA on the CUB and Stanford-Dogs datasets.

Baseline	CLFR	CSFA	CUB		Stanford-Dogs	
			1-shot	5-shot	1-shot	5-shot
✓			54.87	79.09	44.65	67.74
✓	✓		69.81	86.05	58.54	78.02
✓		✓	76.51	88.15	63.32	78.77
✓	✓	✓	**78.66**	**89.43**	**66.42**	**81.23**

Table 4: Module-wise ablation study using Conv-4 backbone on the CUB (using raw images) and Stanford Dogs datasets.

Baseline	CSFA		CUB		Stanford-Dogs	
	CA	PM	1-shot	5-shot	1-shot	5-shot
✓			54.87	79.09	44.65	67.74
✓	✓		75.93	87.97	62.60	78.17
✓		✓	75.46	87.56	61.33	77.85
✓	✓	✓	**76.51**	**88.15**	**63.32**	**78.77**

Table 5: Ablation study of CSFA using Conv-4 backbone on the CUB (using raw images) and Stanford Dogs datasets.

According to the results, it can be observed that: Firstly, our proposed method is tested on five fine-grained datasets, including three widely-used datasets, including Stanford-Dogs, Stanford-Cars, and CUB and the more challenging meta-iNat and tiered meta-iNat datasets, and consistently achieved the state-of-the-art results, demonstrating the effectiveness of the proposed C2-Net. Secondly, compared to FS-FGIC methods, including alignment-based OLSA and multi-layer feature fusion-based AGPF, our method also shows significant performance improvement, demonstrating the rationality of the two issues we addressed regarding the FS-FGIC problem, as well as the effectiveness of the modules proposed to tackle these two issues. Finally, our method shows an even more prominent performance advantage on the 1-shot setting, indicating that the C2-Net can better adapt to the FSL task with extremely limited training samples, indirectly demonstrating the effectiveness of the two proposed modules in preventing over-fitting.

Analyses

Since the overall C2-Net has been proven to be effective in the former section, we give more analyses and experimental results to further demonstrate the effectiveness of C2-Net.

Ablation Study of the Overall Model In order to verify the validity of the two modules proposed, we conduct an ablation study of submodules in this section, and the results are shown in Table 4. It can be seen that both module CLFR and module CSFA consistently improve the performance of the baseline, and the complete C2-Net consisting of both modules achieves the best performance.

Ablation Study of the CSFA Module In consideration that there are two operations in the CSFA module, Table 5 shows the impact of the Channel Activation (CA) operation and Position Matching (PM) operation separately. CA and PM both bring stable performance improvement to the baseline, further validating the effectiveness of adjusting query features from two perspectives with this module.

Visualization

To gain a deeper understanding of the role of each modules, we visualize the query and support feature maps processed by the backbone, the CLFR, and the CSFA module, respectively. As shown in Fig. 3, CLFR can extract both category semantic and detail features while suppressing background noise and non-generalizable features, and the CSFA module can further adjust the feature maps of the query samples according to input support-query pairs for better feature match.

Conclusion

In this paper, we propose the C2-Net to address the issues in the design of the FS-FGIC method. The C2-Net consists of the CLFR module and the CSFA module. The CLFR module resolves the contradiction between FSL and FGIC tasks through feature extraction and refinement. The CSFA module addresses the feature mismatch between support and query samples from the perspective of both semantic information and spatial position. Our proposed method achieves comprehensive superiority in results on five benchmark datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202272, 62172256, 62202278, in part by Natural Science Foundation of Shandong Province under Grant ZR2019ZD06, ZR2020QF036, ZR2021ZD15, in part by the Young Scholars Program of Shandong University, and in part by the Major Program of the National Natural Science Foundation of China under Grant 61991411.

References

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019a. A Closer Look at Few-shot Classification. In *ICLR*.

Chen, Z.; Fu, Y.; Zhang, Y.; Jiang, Y.; Xue, X.; and Sigal, L. 2019b. Multi-Level Semantic Feature Augmentation for One-Shot Learning. *IEEE Transactions on Image Processing*, 28(9): 4594–4605.

Doersch, C.; Gupta, A.; and Zisserman, A. 2020. CrossTransformers: spatially-aware few-shot transfer. In *NeurIPS*.

Dong, C.; Li, W.; Huo, J.; Gu, Z.; and Gao, Y. 2020. Learning Task-aware Local Representations for Few-shot Learning. In *IJCAI*, 716–722.

Du, F.; Yang, P.; Jia, Q.; Nan, F.; Chen, X.; and Yang, Y. 2023. Global and Local Mixture Consistency Cumulative Learning for Long-tailed Visual Recognitions. In *CVPR*, 15814–15823.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.; and Guo, J. 2020. Fine-Grained Visual Classification via Progressive Multi-granularity Training of Jigsaw Patches. In *ECCV*, 153–168.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*, 1126–1135.

Fu, J.; Zheng, H.; and Mei, T. 2017. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *CVPR*, 4476–4484.

Hao, F.; He, F.; Cheng, J.; Wang, L.; Cao, J.; and Tao, D. 2019. Collect and Select: Semantic Alignment Metric Learning for Few-Shot Learning. In *ICCV*, 8459–8468.

Hariharan, B.; and Girshick, R. B. 2017. Low-Shot Visual Recognition by Shrinking and Hallucinating Features. In *ICCV*, 3037–3046.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Horn, G. V.; Aodha, O. M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. J. 2018. The INaturalist Species Classification and Detection Dataset. In *CVPR*, 8769–8778.

Huang, H.; Zhang, J.; Yu, L.; Zhang, J.; Wu, Q.; and Xu, C. 2022. 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.

Huang, H.; Zhang, J.; Zhang, J.; Xu, J.; and Wu, Q. 2021. Low-Rank Pairwise Alignment Bilinear Network For Few-Shot Fine-Grained Image Classification. *IEEE Transactions on Multimedia*, 23: 1666–1680.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel Dataset for Fine-Grained Image Categorization. In *CVPR Workshop*, 806–813.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCV Workshops*, 554–561.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-Learning With Differentiable Convex Optimization. In *CVPR*, 10657–10665.

Lee, S. B.; Moon, W.; and Heo, J. 2022. Task Discrepancy Maximization for Fine-grained Few-Shot Classification. In *CVPR*, 5321–5330.

Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; and Luo, J. 2019a. Revisiting Local Descriptor Based Image-To-Class Measure for Few-Shot Learning. In *CVPR*, 7260–7268.

Li, W.; Xu, J.; Huo, J.; Wang, L.; Gao, Y.; and Luo, J. 2019b. Distribution Consistency Based Covariance Metric Networks for Few-Shot Learning. In *AAAI*, 8642–8649.

Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; and Xue, J. 2021. BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification. *IEEE Transactions on Image Processing*, 30: 1318–1331.

Liu, Y.; Zhang, W.; Xiang, C.; Zheng, T.; Cai, D.; and He, X. 2022. Learning to Affiliate: Mutual Centralized Learning for Few-shot Classification. In *CVPR*, 14391–14400.

Munkhdalai, T.; Yuan, X.; Mehri, S.; and Trischler, A. 2018. Rapid Adaptation with Conditionally Shifted Neurons. In *ICML*, 3661–3670.

Shi, X.; Xu, L.; Wang, P.; Gao, Y.; Jian, H.; and Liu, W. 2020. Beyond the Attention: Distinguish the Discriminative and Confusable Features For Fine-grained Image Classification. In *ACM MM*, 601–609.

Simon, C.; Koniusz, P.; Nock, R.; and Harandi, M. 2020. Adaptive Subspaces for Few-Shot Learning. In *CVPR*, 4135–4144.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *NeurIPS*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *CVPR*, 1199–1208.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. BlockMix: Meta Regularization and Self-Calibrated Inference for Metric-Based Meta-Learning. In *ACM MM*, 610–618.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130: 108792.

Tang, L.; Wertheimer, D.; and Hariharan, B. 2020. Revisiting Pose-Normalization for Fine-Grained Few-Shot Recognition. In *CVPR*, 14340–14349.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *NeurIPS*, 3630–3638.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset.

Wang, C.; Fu, H.; and Ma, H. 2022. PaCL: Part-level Contrastive Learning for Fine-grained Few-shot Image Classification. In *ACM MM*, 6416–6424.

Wei, X.; Wang, P.; Liu, L.; Shen, C.; and Wu, J. 2019. Piecewise Classifier Mappings: Learning Fine-Grained Learners for Novel Categories With Few Examples. *IEEE Transactions on Image Processing*, 28(12): 6116–6125.

Wertheimer, D.; and Hariharan, B. 2019. Few-Shot Learning With Localization in Realistic Settings. In *CVPR*, 6558–6567.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *CVPR*, 8012–8021.

Wu, J.; Chang, D.; Sain, A.; Li, X.; Ma, Z.; Cao, J.; Guo, J.; and Song, Y. 2023. Bi-directional Feature Reconstruction Network for Fine-Grained Few-Shot Image Classification. In *AAAI*, 2821–2829.

Wu, Y.; Zhang, B.; Yu, G.; Zhang, W.; Wang, B.; Chen, T.; and Fan, J. 2021. Object-aware Long-short-range Spatial Alignment for Few-Shot Fine-Grained Image Classification. In *ACM MM*, 107–115.

Xu, S.; Zhang, F.; Wei, X.; and Wang, J. 2022. Dual Attention Networks for Few-Shot Fine-Grained Recognition. In *AAAI*, 2911–2919.

Yang, X.; Wang, Y.; Chen, K.; Xu, Y.; and Tian, Y. 2022a. Fine-Grained Object Classification via Self-Supervised Pose Alignment. In *CVPR*, 7389–7398.

Yang, Y.; Hu, Y.; Zhang, X.; and Wang, S. 2022b. Two-Stage Selective Ensemble of CNN via Deep Tree Training for Medical Image Classification. *IEEE Transactions on Cybernetics*, 52(9): 9194–9207.

Ye, H.; Hu, H.; Zhan, D.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *CVPR*, 8805–8814.

Zhang, B.; Yuan, J.; Li, B.; Chen, T.; Fan, J.; and Shi, B. 2022. Learning Cross-Image Object Semantic Relation in Transformer for Few-Shot Fine-Grained Image Classification. In *ACM MM*, 2135–2144.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *CVPR*, 12200–12210.

Zhu, Y.; Liu, C.; and Jiang, S. 2020. Multi-attention Meta Learning for Few-shot Fine-grained Image Recognition. In *IJCAI*, 1090–1096.