Learning Attention-Guided Pyramidal Features for Few-shot Fine-grained Recognition

Chengcheng Yuan* , **Hao Tang*** , **Dong Zhang** , **Xinguang Xiang** , **Zechao Li†**

School of Computer Science and Engineering, Nanjing University of Science and Technology

{yuancheng, tanghao0918, dongzhang, xgxiang, zechao.li}@njust.edu.cn,

Abstract

Few-shot fine-grained recognition aims to distinguish several highly similar objects from different sub-categories with limited supervisions. However, the current solutions mainly focus on capturing high-level global semantics but surprisingly ignore exploring low-level details, resulting in a new problem of inconspicuous but distinguishable information loss. Therefore, how to effectively tackle the fine-grained recognition task given limited data still remains challenging. In this paper, we propose an effective pyramidal architecture with the attention mechanism to capture both high-level semantic features and low-level detailed features for the fine-grained image recognition in the few-shot learning scenario. Specifically, features in different granular spaces are first gradually combined via a multi-scale feature pyramid and a multi-level attention pyramid on the backbone. Besides, we further present an attention-guided refinement strategy based on the multi-level attention pyramid. Based on this strategy, the raw input features can be refined by enhancing foreground and eliminating background noises. The proposed model is trained with the meta-learning framework in an end-to-end fashion without any extra supervisions. Extensive experimental results on four challenging and widely-used fine-grained datasets show that our model can achieve state-of-the-art performance, especially in the one-shot scenario.

1 Introduction

Fine-grained recognition [Lin *et al.*, 2015], also called sub-category recognition, aims to differentiate objects belonging to different sub-categories under the same super-category. Due to the higher intra-class and lower inter-class variations, fine-grained recognition needs to distinguish subtle visual differences, which is more challenging than generic object recognition. Recently, the rapid development of CNNs has made great progress on this research topic [Wei *et al.*, 2017; Wei *et al.*, 2018; Zhang *et al.*, 2021]. However, there are two major limitations that make many existing works less practical and scalable in real-world applications. Firstly, some works rely on extra annotations (*e.g.,* bounding box or part annotations) where pre-defining them usually requires professional knowledge. Secondly, almost all works heavily rely on large-scale well-labeled training data, which is labor-intensive to collect them. What's worse is that in many scenarios we cannot get enough training data. Hence, inspired by human visual recognition mechanism, researchers pay attention on how to solve the fine-grained recognition with limited labeled data [Wei *et al.*, 2019], especially in a weakly-supervised manner where only image-level class labels are available. In this paper, we study the fine-grained recognition problem in a more practical few-shot setting, where only a few or even one labeled sample is available.

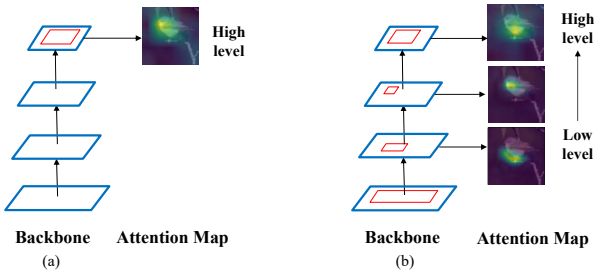

Figure 1: Our motivation (Best view in zooming-in). The interested regions learned from diverse layers: (a) high-level single layer and (b) pyramidal hierarchy layers. Although most previous methods only pay more attention to high-level features, we observe that the low-level attention maps can capture more subtle parts but such detailed information is not explored. In our work, we propose to extract low-level features as supplement information for FSFGR.

Few-shot fine-grained recognition (FSFGR), a novel task that almost few works have been explored before, is first proposed by PCM [Wei *et al.*, 2019] which introduces the bilinear features to learn a piecewise mapping classifier. Tang *et al.* [Tang *et al.*, 2020b] propose to leverage the part annotations for training a pose estimator to localize significant regions. Instead, other methods use a weakly-supervised scheme to locate the distinguishable parts or learn discriminative features. For example, LRPABN [Huang *et al.*, 2020] employs feature alignment to learn a appropriate metric space by introducing pairwise bilinear pooling operator. Recently, MattML [Zhu *et al.*, 2020] introduces a task embedding net-

*These two authors contribute equally to this work.

†Corresponding author: Zechao Li.

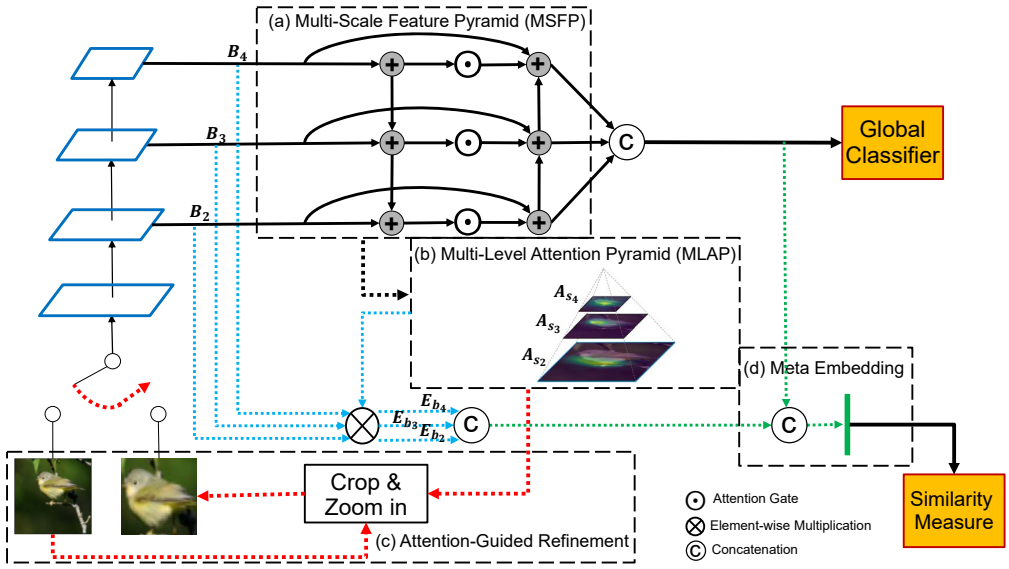

Figure 2: Overview architecture of the proposed framework, which consists of (a) Multi-Scale Feature Pyramid, (b) Multi-Level Attention Pyramid, (c)Attention-Guided Refinement, and (d) Meta Embedding. The process of attention-based reweighing and attention-guided refinement is shown in the blue flow and red flow. The production of meta embedding for few-shot recognition task is indicated in green.

work to combine attention mechanisms with meta learning for FSFGR, which has achieved state-of-the-art performance. As we all know, fine-grained categories are distinguished by subtle visual differences, but almost all of the above methods, even includes most existing few-shot learning methods, have a commonality that high-level features contain global concepts are directly used to compute the metric distance where lacking of low-level information prevents further performance gains. Although attention mechanisms are utilized in some methods to strengthen dominative object, multi-scale representation is less explored for the contextual understanding. Figure 1 shows the differences of feature activations in different layers. It can be observed that high-level features contain the global semantic information, but the low-level features contain local detailed information. Consequently, we are able to distinguish several highly similar objects from various sub-categories with the combination of enhanced high-level and low-level information.

In this work, we formulate the proposed method as a two-stage meta-learning paradigm for FSFGR, which jointly integrates the multi-level feature learning and attention-guided foreground refinement to construct attention-guided pyramidal features. An overview architecture of the proposed framework is shown in Figure 2. The proposed method has two stages, namely coarse stage and refined stage, respectively. The coarse stage mainly obtains the overall characteristics as well as the informative foreground regions of the target. The refined stage, involving the cropping mechanism, takes the finer scale of the raw image as input to mainly study the fine-grained characteristics with more discriminability and less redundancy. All of the above operations are done through a multi-scale feature pyramid (MSFP) combined with a multi-level attention pyramid (MLAP) to simultaneously capture both high-level semantic features and low-level detailed features. To be more specific, we introduce MLAP to reweight the outputs of the backbone and integrate them with the output of the MSFP to explore pyramidal features with complementary information in metric space. Consider the case of mislocalization in the coarse stage which causes attention-guided refinement result is only part of the target, we concatenate output embeddings of both stages to improve the robustness of meta embeddings that used to similarity calculation. To demonstrate the superiority of our proposed model, experiments are carried out on for challenging and widely-used fine-grained datasets. Extensive experimental results declare that the proposed method can significantly improve the accuracy of FSFGR and outperforms the most existing methods by a a large margin, especially in the one-shot scenario.

2 Related Work

2.1 Fine-Grained Visual Recognition

Fine-grained visual recognition has been a hot research topic in recent years, which is more challenging than general object recognition due to the high intra-category variance and low inter-category variance in fine-grained sub-categories. In early studies, some traditional methods [Chai *et al.*, 2013; Xie *et al.*, 2013] are proposed to exploit hand-crafted features in training and inference stages, but the performance and generalization of this type of approach is unsatisfactory. Benefiting from the remarkable success of deep neural networks, the research direction of fine-grained recognition has turned to how to acquire rich information in a weakly supervised manner, where only image-level labels are available. Deep learning-based methods can be roughly divided into two categories: feature encoding-based methods [Lin *et al.*, 2015; Gao *et al.*, 2016] and part localization-based methods [Fu *et al.*, 2017; Yang *et al.*, 2018]. Recently, some studies [Zhang *et al.*, 2021; Ding *et al.*, 2021] are extensively explored to the integration of part localization and feature learning in an end-

to-end framework without bounding-boxes or part annotations. For example, MMAL [Zhang *et al.*, 2021] predicts the positions of the object and informative part regions through multi-branch networks with attention mechanisms. However, most methods ignores the fact that the multi-level information can also contribute to fine-grained recognition, especially in the few-shot scenarios. By comparison, our proposed method introduces the pyramidal hierarchy to take full advantage of attention-guided multi-level features for FSFGR.

2.2 Few-shot Learning

The purpose of the few-shot learning is to learn to classify novel instances into a set of unseen categories given limited labeled instances per category, which has attracted considerable research attention [Peng *et al.*, 2019; Tang *et al.*, 2020a]. Deep neural networks have made significant progress toward few-shot learning problem, recent efforts can be roughly cast into two categories, including meta-learning based methods [Finn *et al.*, 2017; Munkhdalai *et al.*, 2018; Jamal and Qi, 2019] and metric-learning based methods [Vinyals *et al.*, 2016; Snell *et al.*, 2017; Sung *et al.*, 2018]. The former aims to find an appropriate gradient-based optimization strategy and the latter aims to learn a well generalized embedding as well as an appropriate comparison metric. In this paper, we will mainly follow the metric-based methods due to the simplicity and effectiveness, which have achieved higher performance on few-shot learning tasks. However, not like most existing metric-learning based methods that directly compute the similarity of features based on the global concepts, our proposed two-stage method conducts attention-guided refinement at the end of the first stage and takes the finer scale of object image as the input of the second stage to enhance richer representation for the final decision of FSFGR.

3 Method

3.1 Preliminaries

For the few-shot learning problem, the dataset is typically divided into three part, including training set $\mathcal{D}_{base}$, support set $\mathcal{S}$ and query set $\mathcal{Q}$. Episode training mechanism [Vinyals *et al.*, 2016] is an effective approach for few-shot learning, where the training process simulate the settings in the testing. Few-shot recognition task aims to recognize the unlabeled samples in $\mathcal{Q}$ given $\mathcal{D}_{base}$ and $\mathcal{S}$, where $\mathcal{S}$ and $\mathcal{Q}$ share the same label space but disjoint with that of $\mathcal{D}_{base}$. If the support set $\mathcal{S}$ consists of N categories and K labeled samples per category, the target problem is also called "N-way K-shot" learning. In our work, we also adopt the meta-learning setting to conduct FSFGR.

3.2 Framework

An overview of our proposed framework is shown in Figure 2. As a two-stage method that contains a coarse stage and a refined stage, in which each stage establishes a multi-scale feature pyramid (MSFP) and a multi-level attention pyramid (MLAP) on the basic backbone to explore more fine-grained features for improving performance. Noted that the two stages share the same model architecture and parameters. In the coarse stage, MSFP generates pyramidal features, which contain both high-level semantic information and low-level texture information, as well as MLAP by attention gates. Once the MLAP is established, then we can obtain the region of interest with a bounding box according to the attention-guided refinement strategy. The refined stage distill the more discriminative and less redundant information from the region of interests clipped from the raw image. Significantly, both stages set individual classifiers for multi-scale pyramidal features, which only used in the training stage. We also construct a special *meta embedding* for instance similarity calculation in FSFGR, which consists of (a) the outputs of MSFP, and (b) the inputs of MSFP but re-weighted by the MLAP. The proposed method is trained in an end-to-end manner without bounding box or part annotations.

3.3 Multi-Scale Feature Pyramid

The motivation of the proposed method is to capture both low-level subtle information and high-level semantic information for fine-grained recognition in the few-shot scenario. As shown in Figure 2(b), the backbone generates feature maps with different scales, denoted as $B_1, B_2, \ldots, B_n$. Many previous works directly classify generic objects by the last global feature B_n that only contains high-level semantic information. But for fine-grained recognition in the few-shot scenario, the detailed texture information in low-level features is essential. The proposed method analyzes feature maps with different scales to simultaneously capture high-level semantic information and low-level texture information by establishing multi-scale feature pyramid. Inspired by FPT [Zhang *et al.*, 2020], we construct a top-down pathway to combine high-level semantic features with low-level texture features and a bottom-up pathway to deliver subtle information from low levels to higher levels. In particular, we further introduce an additional attention gate to better locate discriminative regions and a skip connection from inputs to maintain the backbone information in each pyramid level. In other words, MSFP learns multiple scale-specific representations in an explicit way, which contains more discriminative and rich information. In order to make the model fully and effectively learn the representations obtained by MSFP, we aggregate the above outputs to predict the global semantic label of the input image in $\mathcal{D}_{base}$ during the training stage. For each episode, we make semantic prediction on each sample and choose cross-entropy loss as the global classification loss as follow:

$$\mathcal{L}_1(\theta) = \mathop{\mathbb{E}}_{(\mathcal{S},\mathcal{Q})} - \sum_{i=1}^{Q} \log p_\theta(y_q^i \mid q_i) - \sum_{j=1}^{S} \log p_\theta(y_s^j \mid s_j), \tag{1}$$

where $(q_i, y_q^i) \in \mathcal{Q}$ and $(s_j, y_s^j) \in \mathcal{S}$ are query sample and support sample in each episode, and θ indicates the parameters of the model.

3.4 Multi-Level Attention Pyramid

As shown in Figure 2(c), we introduce an attention gate in the MSFP to construct the MLAP, where the purposed model consists of (a) reweighting the outputs of backbone from different scales, and (b) locating discriminative regions by attention-guided refinement.

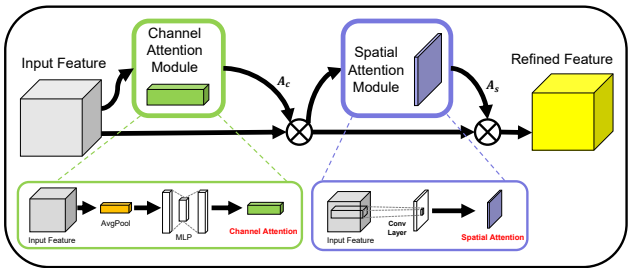

Figure 3: The overview of our proposed attention gate, which has two sequential sub-structures, channel attention module and spatial attention module. $\otimes$ represents element-wise multiplication.

Attention Gate. Inspired by CBAM [Woo *et al.*, 2018], we introduce a simple and effective attention gate to perform adaptive feature optimization along channel and spatial dimension. Figure 3 shows the flow diagram. Given an input feature map X, the channel attention map A_c and spatial attention map A_s can be represented as

$$\begin{aligned} A_c &= W_2 * ReLU\left(W_1 * GAP(X)\right), \\ A_s &= W_3 * \left(A_c \otimes X\right), \end{aligned} \tag{2}$$

where GAP and $ReLU$ denote the global average pooling and the ReLU function,respectively. The $\otimes$ represents the element-wise multiplication. Here W_1, W_2, W_3 refer to the parameters of the convolution kernel. As a result, we obtain multi-level attention pyramid $\{A_{s_2}, A_{s_3}, \cdots, A_{s_n}\}$ based on multi-scale feature maps.

Attention-based Reweighting. The attention map A_{s_n} contains spatial activation of corresponding feature B_n in each pyramid level, so we use the *absolute value* of each element in A_{s_n} as weight to construct the corresponding weight map $\widetilde{A}_{s_n}$, As shown in Figure 2 with the blue flow, the reweighted features with different scales E_{b_i} and final output E_b can be represented as:

$$\begin{aligned} E_{b_i} &= GAP(B_i \otimes \widetilde{A}_{s_i}), \\ E_b &= Concat(E_{b_2}, E_{b_3}, E_{b_4}). \end{aligned} \tag{3}$$

3.5 Attention-guided Refinement.

The intuition of the proposed attention-guided refinement is that the positions with higher activation values on the spatial attention maps are often the areas where the target object is located. As shown in Figure 2(d), we conduct attention-guided refinement on the pyramid bottom features to learn the minimum bounding box in the input image in a weakly-supervised manner, which further to improve the performance in the refinement stage. Then, we crop region of interested object according to the location coordinates and enlarge it to the same size as raw image to obtain the refined features. The main process of generating cropping coordinates by processing the spatial activation map as shown in Algorithm 1, and Figure 4 illustrates a visualized example.

3.6 Meta Embedding

As shown in Figure 2(e), in each stage, we conduct a special *meta embedding* for each input sample by concatenating the outputs of MSFP and the inputs of MSFP reweighted by the MLAP. Noticeably, we measure the similarity between

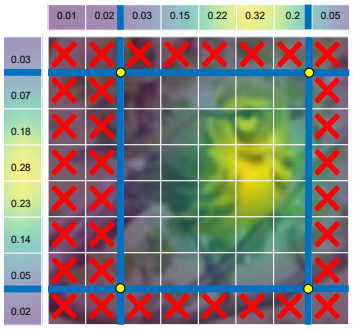

Figure 4: A specific example of attention-guided refinement with threshold $\theta = 0.05$. The yellower the region, the more discriminative it is.

Algorithm 1 Attention-guided refinement algorithm

Input: Raw Image $B \in R^{H \times W}$,
Spatial Activation Map $S \in R^{h \times w}$, Threshold θ
Cropping Coordinate $[x_{left}, y_{top}, x_{right}, y_{bottom}]$
Output: Refined Image $B_{refined}$

1: Zoom-in $S \in R^{h \times w} \longrightarrow S \in R^{H \times W}$;
2: $S_{norm} \longleftarrow S \,/\, \sum_{i=1}^{H} \sum_{j=1}^{W} S^{i,j}$;
3: Compute $x_{left} = \arg\max_{l} \sum_{i=1}^{H} \sum_{j=1}^{l-1} S_{norm}^{i,j} \leq \theta$;
4: Compute $x_{right} = \arg\min_{r} \sum_{i=1}^{H} \sum_{j=r+1}^{W} S_{norm}^{i,j} \leq \theta$;
5: Compute $y_{top} = \arg\max_{t} \sum_{i=1}^{t-1} \sum_{j=1}^{W} S_{norm}^{i,j} \leq \theta$;
6: Compute $y_{bottom} = \arg\min_{b} \sum_{i=b+1}^{H} \sum_{j=1}^{W} S_{norm}^{i,j} \leq \theta$;
7: Crop B by $[x_{left}, y_{top}, x_{right}, y_{bottom}] \longrightarrow B_{refined}$;
8: Zoom-in $B_{refined} \longrightarrow B_{refined} \in R^{H \times W}$;
9: **return** $B_{refined}$;

meta embeddings, which aggregate complementary information discovered from different scales and stages. In this work, our solution to the few-shot recognition task is in line with the Prototypical Network [Snell *et al.*, 2017], due to the simplicity and effectiveness. However, it is worth noting that our method can work with any other metric-based meta-learning methods. We first calculate the prototype for each category by averaging all meta embeddings belong to that category, and then the query samples will be considered to be the same category as the closest prototype. Specifically, given the prototype p_c of the category c, the probability of query sample q_t as category c is:

$$p_\theta(y = c|q_t) = \frac{exp(-d(f(q_t), p_c))}{\sum_k exp(-d(f(q_t), p_k))}, \tag{4}$$

where $d(\cdot)$ is the Euclidean distance. For an episode, the nearest neighbor recognition loss is then defined as:

$$\mathcal{L}_2(\theta) = \mathop{\mathbb{E}}_{\mathcal{Q}} - \sum_{i=1}^{Q} \log p_\theta \left(y = c_{q_i} \mid q_i\right), \tag{5}$$

where $(q_i, c_{q_i}) \in \mathcal{Q}$ and the parameters of the model are θ. Finally, incorporating Eq. (1) and Eq. (5), the overall objective function is to minimize the following equation:

$$\mathcal{L}(\theta) = \mathcal{L}_1(\theta) + \mathcal{L}_2(\theta). \tag{6}$$

Methods	Ref.	Backbone	FGVC-Aircraft		CUB-200-2011		Stanford Cars		Stanford Dogs	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MaNet	NeurIPS'16	Conv4	54.41 ± 0.47	70.04 ± 0.35	60.52 ± 0.88	75.29 ± 0.75	34.80 ± 0.98	44.70 ± 1.03	35.80 ± 0.99	47.50 ± 1.03
ProtoNet	NeurIPS'17	Conv4	57.62 ± 0.49	71.43 ± 0.38	50.46 ± 0.88	76.39 ± 0.64	40.90 ± 1.01	52.93 ± 1.03	37.59 ± 1.00	48.19 ± 1.03
ReNet	CVPR'18	Conv4	63.94 ± 0.51	76.22 ± 0.34	62.34 ± 0.94	77.84 ± 0.68	47.79 ± 0.49	60.60 ± 0.41	43.29 ± 0.46	55.15 ± 0.39
MAML	ICML'17	Conv4	60.24 ± 0.34	75.24 ± 0.24	54.73 ± 0.97	75.75 ± 0.76	47.25 ± 0.30	61.11 ± 0.29	44.84 ± 0.31	58.61 ± 0.30
AdaCNN	ICML'18	Conv4	58.60 ± 0.48	72.82 ± 0.38	-	-	41.88 ± 0.40	49.87 ± 0.37	42.16 ± 0.43	54.12 ± 0.39
CovaMNet	AAAI'19	Conv4	60.10 ± 0.93	70.24 ± 0.73	58.87 ± 1.00	70.46 ± 0.84	56.65 ± 0.86	71.33 ± 0.62	49.10 ± 0.76	63.04 ± 0.65
DN4	CVPR'19	Conv4	59.71 ± 0.91	85.05 ± 0.60	57.45 ± 0.89	84.41 ± 0.58	61.51 ± 0.85	**89.60 ± 0.44**	45.73 ± 0.76	66.33 ± 0.66
LRPABN	TMM'20	Conv4	-	-	63.63 ± 0.77	76.06 ± 0.58	60.28 ± 0.76	73.29 ± 0.58	45.72 ± 0.75	60.94 ± 0.66
MattML	IJCAI'20	Conv4	75.69 ± 0.54	86.23 ± 0.31	66.29 ± 0.56	80.34 ± 0.30	66.11 ± 0.54	82.80 ± 0.28	54.84 ± 0.53	71.34 ± 0.38
Ours		Conv4	**79.16 ± 0.86**	**87.35 ± 0.58**	**74.03 ± 0.90**	**86.54 ± 0.50**	**78.14 ± 0.84**	87.42 ± 0.57	**60.89 ± 0.98**	**78.14 ± 0.62**

Table 1: The few-shot classification accuracy (%) with 95% confidence intervals on four datasets, when using Conv4 as the backbone.

4 Experiment

4.1 Dataset

In this work, we conduct few-shot fine-grained recognition task on the four challenging and widely-used fine-grained datasets, including CUB-200-2011 [Wah *et al.*, 2011], Stanford-Cars [Krause *et al.*, 2013], Stanford-Dogs [Khosla *et al.*, 2011], and FGVC-Aircraft [Maji *et al.*, 2013]. Significantly, the proposed method does not require additional box annotations or part annotations except image-level annotation. All images from the four datasets are resized to 84×84. For the category split of above datasets, we followed the separation of DN4 [Li *et al.*, 2019] and MattML [Zhu *et al.*, 2020]. Note that the validation set is only used to select the optimal experiment setting.

4.2 Experimental Setting

For a fair comparison, we follow the same experimental settings as in [Chen *et al.*, 2019; Zhu *et al.*, 2020] and use widely-adopted Conv4 and ResNet12 networks as embedding modules to perform two types of tasks called 5-way 1-shot and 5-way 5-shot image recognition. In the meta-training stage, we conduct $60,000$ episodes during the training for all experiments. For N-way K-shot recognition task, each episode contains $N \times K$ labeled samples as support set and $16 \times N$ unlabeled samples as query set. In the meta-testing stage, we report the average accuracy and the corresponding 95% confidence interval over the 600 episodes as the final results. Adam is used as the optimizer to optimize our model from scratch with initial learning rate 0.001. We adopt random crop, horizontal flip, color jitter, and random erasing as data augmentation.

4.3 Experimental Results

Table 1 shows the comparison of classification accuracy with state-of-the-are methods on different fine-grained datasets using Conv4 as the backbone. We can see that the proposed method shows superiority compared with other methods on almost all datastes. Especially for 5-way 1-shot task, our method gains **3.47**%, **7.74**%, **12.03**% and **6.05**% improvements over the state-of-the-art methods on the 4 datasets mentioned in Section 4.1, respectively. When using deeper ResNet12 as the backbone, we use the CUB-200-2011 dataset to conduct main experiments, and their results are shown in the Table 2. By comparison, our method is **3.58**%, **0.37**% higher than state-of-the-art methods for 1-shot and 5-shot task, respectively. Notably, our method based on ResNet12 outperforms two methods based on ResNet18, which demonstrates the superiority of attention-guided pyramid refinement, and also reflects that multi-scale representations guided by multi-level attention activations can deal with few-shot fine-grained recognition well than a single global descriptor. In summary, the proposed method can be well extended to all four fine-grained datasets.

Methods	Ref.	Backbone	CUB-200-2011	
			5-way 1-shot	5-way 5-shot
MaNet	NeurIPS'16	ResNet12	71.87 ± 0.85	85.08 ± 0.57
ProtoNet	NeurIPS'17	ResNet12	66.09 ± 0.92	82.50 ± 0.58
ReNet	CVPR'18	ResNet12	70.20 ± 0.84	84.28 ± 0.46
Baseline++	ICLR'19	ResNet12	67.30 ± 0.86	84.75 ± 0.60
MetaOptNet	CVPR'19	ResNet12	75.15 ± 0.46	87.09 ± 0.30
Neg-Cosine	ECCV'20	ResNet18	72.66 ± 0.85	89.40 ± 0.43
Centroid-A	ECCV'20	ResNet18	74.22 ± 1.09	88.65 ± 0.55
Ours		ResNet12	**78.73 ± 0.84**	**89.77 ± 0.47**

Table 2: Average accuracy (%) comparisons to state-of-the-arts by using ResNet12 as the backbone network on CUB-200-2011 dataset.

Methods	Stanford-Cars	
	5-way 1-shot	5-way 5-shot
Baseline	67.28 ± 0.90	79.35 ± 0.70
Baseline+MSFP	74.40 ± 0.82	85.12 ± 0.60
Baseline+MSFP+MLAP (1 stage)	75.30 ± 0.84	85.81 ± 0.60
Baseline+MSFP+MLAP (2 stage)	**78.14 ± 0.84**	**87.42 ± 0.57**

Table 3: Ablation study analysis on Stanford Cars dataset by using Conv4 as the backbone network.

4.4 Ablation Study

Table 3 shows ablation studies of the 5-way 1-shot and 5-way 5-shot tasks on Stanford Cars. The first model called Baseline, only uses single-scale features directly obtained from the top layer of backbone with an attention gate for recognition. The second model called Baseline+MSFP, uses multi-scale features for recognition, which directly aggregate the outputs of multi-scale feature pyramid. The third model called Baseline+MSFP+MLAP(1stage), uses singe-stage features for recognition, which aggregates the outputs of multi-scale feature pyramid and multi-level attention pyramid at coarse stage. The fourth model called

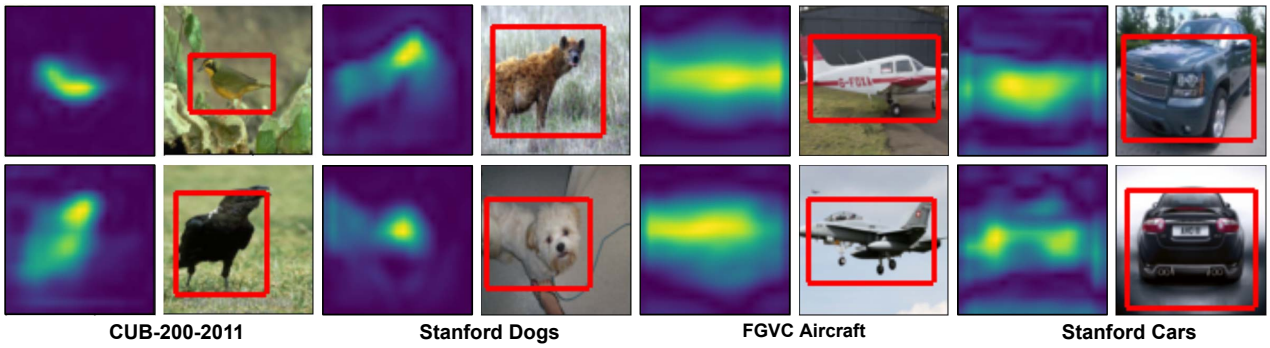

Figure 5: Visualizations of attention-guided pyramid refinement which can locate more vital regions to improve recognition.

Methods	MSFP	Backbone	CUB-200-2011	
			5-way 1-shot	5-way 5-shot
MatchingNet	✗	Conv4	60.52 ± 0.88	75.29 ± 0.75
MatchingNet	✓	Conv4	71.10 ± 0.91 $_{\uparrow 10.58}$	84.94 ± 0.57 $_{\uparrow 9.65}$
RelationNet	✗	Conv4	62.34 ± 0.94	77.84 ± 0.68
RelationNet	✓	Conv4	66.32 ± 1.01 $_{\uparrow 3.98}$	80.37 ± 0.62 $_{\uparrow 2.53}$
MatchingNet	✗	ResNet12	71.87 ± 0.85	85.08 ± 0.57
MatchingNet	✓	ResNet12	74.00 ± 0.83 $_{\uparrow 2.13}$	87.77 ± 0.48 $_{\uparrow 2.69}$
RelationNet	✗	ResNet12	70.20 ± 0.84	84.28 ± 0.46
RelationNet	✓	ResNet12	72.98 ± 0.96 $_{\uparrow 2.78}$	85.89 ± 0.54 $_{\uparrow 1.61}$

Table 4: Effectiveness of multi-scale feature pyramid which can extract the feature representations with contextual information.

Baseline+MSFP+MLAP(2stage), uses two-stage features for recognition, which aggregates the outputs of multi-scale feature pyramid and multi-level attention pyramid at both of coarse stage and refined stage. We can observe that: i) Only employing pyramidal architecture, we gains 7.12% and 5.77% improvements, which demonstrates that the multi-scale feature pyramid with multi-level information is essential for fine-grained recognition. ii) Compare with Baseline+MSFP, the method equipped with MLAP obtains further improvements, which indicates that the pyramidal outputs of backbone reweighted by pyramidal attention contain complementary discriminative information. iii) On the basis of the coarse stage, our method achieves about 2.84% and 1.61% improvements going through the refined stage, which explains that the attention-guided refinement can learn fine-grained features with less redundancy in different scales other than the structural features.

4.5 Additional Analysis

Effectiveness of MSFP

To further illustrate the effectiveness of our proposed multi-scale feature pyramid (MSFP), we plug this module into another two metric-based algorithms (*i.e.*, MatchingNet [Vinyals *et al.*, 2016] and RelationNet [Sung *et al.*, 2018]) to conduct extensive experiments on CUB-200-2011 [Wah *et al.*, 2011]. As shown in Table 4, we can observe that our proposed MSFP achieves consistent performance gains on the different benchmark algorithms and backbones. Although the proposed module is simple, these results also prove that gradually integrating features of different granularity from low-level to high-level is more beneficial to fine-grained recognition than only using features of the most discriminative regions, especially in scenarios with limited labeled samples.

Visualization of Refinement

To understand our proposed attention-guided refinement more intuitively, we visualize spatial activation maps in multi-level attention pyramid (MLAP) and corresponding results of attention-guided pyramid refinement in Figure 5, in which highlighted regions are relevant to the more discriminative local parts playing a key role in fine-grained recognition. We observe that the proposed attention-guided refinement based on MLAP not only can locate more discriminative regions but also can eliminate the background noises, which contributes to the ability of grasping the structure information and enhancing the rich representation.

5 Conclusion

In this paper, we proposed an effective pyramidal architecture with the attention-guided pyramidal features for few-shot fine-grained image recognition without additional supervisions. The proposed model made full use of the high-level semantic information and the low-level detailed information via building a multi-scale feature pyramid and a multi-level attention pyramid upon the backbone network. Specifically, we introduced an attention-guided refinement to achieve excellent performance by enhancing the discriminative regions and eliminating the background noises. Our model has great flexibility, which can be trained in an end-to-end manner and worked together with other metric-based meta-learning frameworks. Experiments conducted on four challenging and widely-used fine-grained datasets demonstrated the superiority of our method. In the future, we will explore how to apply our proposed model to other computer vision tasks, such as object detection, semantic segmentation, and person re-identification.

Acknowledgement

The authors would like to thank all the anonymous reviewers for their constructive comments and suggestions. This work was partially supported by the National Key Research and Development Program of China under Grant 2017YFC0820601, and the National Natural Science Foundation of China (Grant No. U20B2064 and 61720106004).

References

[Chai *et al.*, 2013] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *ICCV*, 2013.

[Chen *et al.*, 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019.

[Ding *et al.*, 2021] Yifeng Ding, Zhanyu Ma, Shaoguo Wen, Jiyang Xie, Dongliang Chang, Zhongwei Si, Ming Wu, and Haibin Ling. Ap-cnn: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, 2021.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[Fu *et al.*, 2017] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017.

[Gao *et al.*, 2016] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, 2016.

[Huang *et al.*, 2020] 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*, 2020.

[Jamal and Qi, 2019] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *CVPR*, 2019.

[Khosla *et al.*, 2011] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop*, 2011.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013.

[Li *et al.*, 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. In *CVPR*, 2019.

[Lin *et al.*, 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015.

[Maji *et al.*, 2013] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. In *arXiv*, 2013.

[Munkhdalai *et al.*, 2018] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, 2018.

[Peng *et al.*, 2019] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *ICCV*, 2019.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[Sung *et al.*, 2018] 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 *CVPR*, 2018.

[Tang *et al.*, 2020a] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *ACM MM*, 2020.

[Tang *et al.*, 2020b] Luming Tang, Davis Wertheimer, and Bharath Hariharan. Revisiting pose-normalization for fine-grained few-shot recognition. In *CVPR*, 2020.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, 2016.

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

[Wei *et al.*, 2017] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 2017.

[Wei *et al.*, 2018] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 2018.

[Wei *et al.*, 2019] 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*, 2019.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[Xie *et al.*, 2013] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *ICCV*, 2013.

[Yang *et al.*, 2018] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, 2018.

[Zhang *et al.*, 2020] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *ECCV*, 2020.

[Zhang *et al.*, 2021] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MMM*, 2021.

[Zhu *et al.*, 2020] Yaohui Zhu, Chenlong Liu, and Shuqiang Jiang. Multi-attention meta learning for few-shot fine-grained image recognition. In *IJCAI*, 2020.