Multi-Scale Adaptive Task Attention Network for Few-Shot Learning

Haoxing Chen, Huaxiong Li, Yaohui Li, Chunlin Chen

Nanjing University
{haoxingchen, yaohuili}@smail.nju.edu.cn, {huaxiongli, clchen}@nju.edu.cn

Abstract. Few-shot learning has aroused considerable interests in recent years, which aims to recognize unseen categories by using few labeled samples. In various few-shot methods, low-level information metric-learning based methods have achieved promising performance. However, most of these methods deal with each category in the support set independently, which may be insufficient to measure the relations among features, especially in a specific task. Besides, the coexistence of dominant objects at different scales may degrade the performance of these methods. To address these issues, a novel Multi-Scale Adaptive Task Attention Network, MATANet for short, is proposed for few-shot learning. In MATANet, a multi-scale feature generator is first constructed to extract the image features at different scales. Then, an adaptive task attention module is built to select the most important LRs among the entire task. Finally, a similarity-to-class module is adopted to measure the similarities between query and support set. Extensive experiments on popular benchmarks show the effectiveness of the proposed MATANet compared with state-of-the-art methods.

Keywords: Few-shot learning · Metric-learning · Task attention · Image classification.

1 Introduction

Deeplearning based computer vision method has achieved great success in many practical problems. Most of these methods require a lot of labeled data for training. However, collecting a large amount of labeled data is time-consuming and laborious. Besides, some fine-grained data [10, 13, 27] require expert knowledge to be accurately labeled, i.e., ordinary people can only distinguish a few bird species, while bird experts can accurately label various birds. How to tackle the image classification under the few-shot learning setting accurately remains an open problem.

Humans can easily learn new concepts and objects with only one or a few samples. In order to imitate this ability of humans, many few-shot learning methods [1,5,6,15,16,20,23–26] have been proposed. Meta-learning and metric-based learning are two kinds of mainstream few-shot learning methods. Meta-learning based methods [5,20] focus on how to find a good parameter initialization or optimizers. Metric-learning based methods [2,5,6,15,16,24–26] aims to find a more

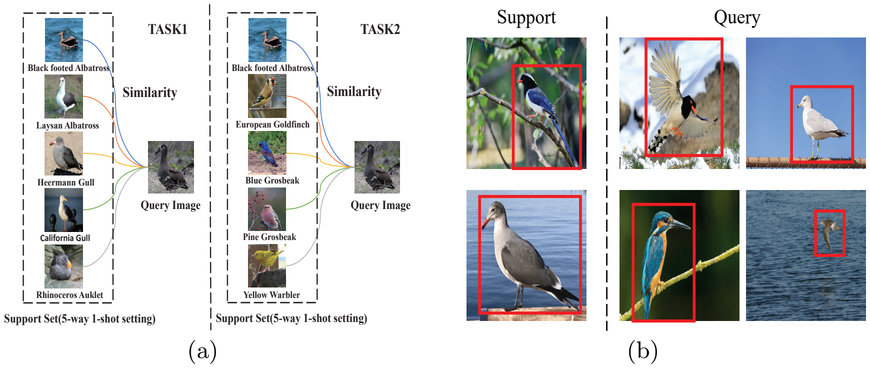

Fig. 1. The two main limitations of the previous local representation based methods. (a) In different tasks, the most discriminative features are different. In task 1, the beak is the key distinguishing feature, while the most critical feature is the wing in task 2. (b) The scales of the dominant objects vary from image to image.

discriminative distance measure to distinguish different categories of samples. However, most of these metric learning methods [2, 5, 6, 24–26] adopt image-level features for classification. Due to the scarcity of samples in few-shot image recognition tasks, classifying at this level may not be effective enough. Instead, many methods [6, 16] based on low-level information were proposed, i.e., Local Representations (LRs) of feature embeddings. These methods use low-level information to measure the distance between query images and support images, and they can achieve better recognition results [15]. However, these methods do not measure the similarity between query images and support images in the context of the whole task, which cannot make full use of the representation ability of local feature descriptors. When humans classify an image into one of several unseen classes, it is natural to focus on semantic features shared only between certain classes and the query image. In other words, humans do not pay much attention to the features shared between classes when recognizing a category they have not seen before. For example, consider two 5-way 1-shot tasks in Fig. 1(a). In task1, we need to recognize a 'Black-footed Albatross' among 'Laysan Albatross', 'Heermann Gull', 'California Gull', and 'Rhinoceros Auklet'. While in task 2, we need to recognize a 'Black-footed Albatross' among 'European Goldfinch', 'Blue Grosbeak', 'Pine Grosbeak', and 'Yellow Warbler'. For task 1, the beak is a very distinguishing feature, but for task 2, it is not the most critical feature. While the wing is more important for task 2 than task 1. In summary, the importance of each LR varies from task to task.

Although many existing methods [6, 15, 16] can extract the relation between the query image and each support set independently, they do not consider the importance of each LR under the whole task, and all the LRs are weighted equally, *and we argue that the task-relevant LRs should enjoy the higher weights.* Moreover, these methods can only calculate their similarity at a single scale. As shown in Fig. 1(b), the scale of dominant objects in different images are dissimilar, *and we argue that it is more reasonable to calculate the similarity between query image and support set at multiple scales simultaneously.* In addition, the features representation used in [6,15,16] are not discriminative, since CNN treats all features equally.

To this end, we propose a novel *Multi-Scale Adaptive Task Attention Network* for metric-learning based few-shot learning, which can be trained in an end-to-end manner. First, we use a self-attention module to enhance the discriminant ability of feature representations for few-shot learning and represent all images as a collection of LRs at different scales by a multi-scale feature generator, rather than a global feature representation at the image level. Second, we measure the semantic similarity between the query image and the support set by calculating the semantic relation matrix. Afterward, we employ an adaptive task attention module to select the most distinguishing feature in the current task. Third, to further make full use of LRs, we employ a similarity-to-class mechanism to determine which support class the query image belongs to at each scale. Finally, we adaptively fuse the similarities calculated from the features of different scales together.

To sum up, the main contributions are summarized as follows: First, we combine self-attention module and feature extractor to enhance the discriminant ability of local feature representations. Second, to generate different scale features, we propose a multi-scale feature generator in few-shot learning tasks, which can provide multi-scale information for more comprehensive measurements. Third, we further propose a novel adaptive task attention mechanism by finding and weighing the most discriminative local representations in the entire task, aiming to learn task-relevant feature representations for few-shot learning. Finally, we conduct sufficient experiments on four benchmark datasets to verify the advancement of our model, and the performance of our model achieves the state-of-the-art.

2 Related Work

In this section, we review the recent metric-learning based few-shot learning literature. The metric-learning based methods aim to learn an informative distance metric, as presented in [2,6,15,16,22–26]. Koch *et al.* [12] migrated the Siamese Network to the one-shot learning task. Snell *et al.* [24] proposed Prototypical Networks, which assumes that each type can be represented by a prototype, and the prototype can be obtained by calculating the mean value of the embedding representation of each class, then using a distance function for classification. Normally, we do not know which distance function is the best. Therefore, Sung *et al.* [25] proposed a Relation Network to obtain the most suitable distance metric function through learning. The above methods are based on the feature representation at the image level. Due to the scarcity of the number of samples, we cannot well represent the distribution of each category at the image-level features. In contrast, some recent work, such as SAML [6], DN4 [15] and Co-

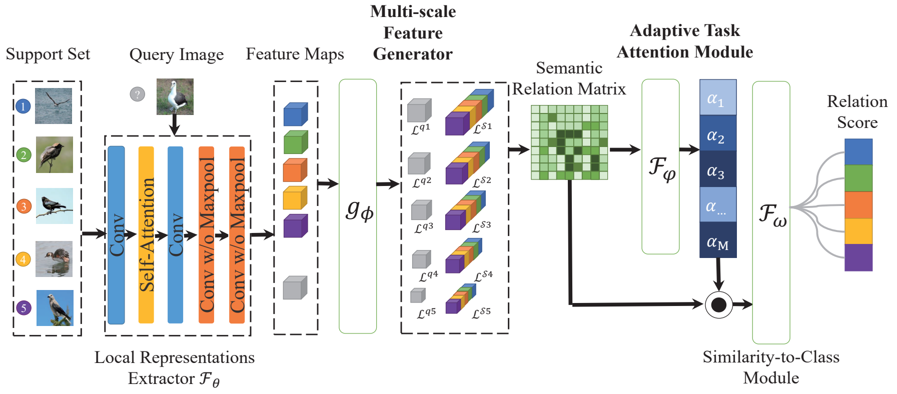
Support Set
Query Image
Feature Maps
Multi-scale Feature Generator
Semantic Relation Matrix
Adaptive Task Attention Module
Relation Score
Conv
Self-Attention
Conv
Conv w/o Maxpool
Conv w/o Maxpool
Local Representations Extractor $\mathcal{F}_\theta$
g_ϕ
$\mathcal{F}_\varphi$
$\mathcal{F}_\omega$
α_1
α_2
α_3
$\alpha_{\dots}$
α_M
$\mathcal{L}^{q1}$ $\mathcal{L}^{s1}$
$\mathcal{L}^{q2}$ $\mathcal{L}^{s2}$
$\mathcal{L}^{q3}$ $\mathcal{L}^{s3}$
$\mathcal{L}^{q4}$ $\mathcal{L}^{s4}$
$\mathcal{L}^{q5}$ $\mathcal{L}^{s5}$
Similarity-to-Class Module

Fig. 2. The architecture of the proposed MATANet. (Best view in color.)

vaMNet [16] shows that the rich low-level features (i.e., LRs) have better representation capabilities. Hao *et al.* [6] proposed the Semantic Alignment Metric Learning method to finding semantic relevance between query images and support set. Li *et al.* [15] proposed a Deep Nearest Neighbor Network to find the similarity between local feature descriptors. Li *et al.* [16] proposed a CovaMNet to measure the distance between query images and support set by a covariance measurement.

However, most previous few-shot learning methods mentioned above measure the similarity between the query image and each support class independently, without considering the entire task together. In order to solve this issue, Li *et al.* [14] proposed a Category Traversal Module (CTM), which can select task-relevant features. Although their method combines task information for few-shot learning, they find task-relevant features at the image-level, which may be not effective [5, 18]. Moreover, most few-shot learning methods only measure the similarity between query image and support set at a single scale, which may lead to a lower classification accuracy in the case that the scales of dominant objects are different. Jiang *et al.* [9] proposed a Multi-Scale Metric Learning (MSML) method to extract features and calculate similarities at multiple scales. MSML introduces complex network structures and classifies on image-level features, which may not be effective [1].

Unlike the above methods, our MATANet calculates the similarity between the query image and the support set at multiple scales. We can obtain the final result through integrate multiple similarities from different scales. In addition, our MATANet can adaptively select task-relevant local features with discriminative semantics, as the process of human recognition.

3 The proposed Method

3.1 Description on Few-Shot Learning

In few-shot learning, there are usually three sets of data: a query set $\mathcal{Q}$, a support set $\mathcal{S}$, and an auxiliary set $\mathcal{A}$. Note that $\mathcal{Q}$ and $\mathcal{S}$ share the same label space, while they have no intersection with the label space of $\mathcal{A}$.

In this paper, we follow the definition of the common few-shot learning task. Given a N-way K-shot task (e.g., 5-way 1-shot or 5-way 5-shot), we have N previously unseen classes with K samples, for every query image, we need to classify which class the query image belongs to. To achieve this goal, we use an auxiliary set to train a model to learn transferable knowledge. The model is trained by the episodic training mechanism [26]. In each episode, a new task is randomly constructed in $\mathcal{A}$, and each task consists of two subsets: auxiliary support set $\mathcal{A}_{\mathcal{S}}$ and auxiliary query set $\mathcal{A}_{\mathcal{Q}}$. Generally, in the training stage, hundreds of tasks are adopted to train the model.

Fig. 2 shows the architecture of our proposed MATANet, which consists of four components: a local representations extractor $\mathcal{F}_\theta$, a multi-scale feature generator g_ϕ, an adaptive task attention module $\mathcal{F}_\varphi$, and a similarity-to-class module $\mathcal{F}_\omega$. All image samples are first fed into the local representations extractor $\mathcal{F}_\theta$ to get rich LRs. In practice, we choose 4-layer CNN as our feature extractor [5,15,18] and we embed the self-attention module after the first block, to enhance the discriminant ability of feature representations for few-shot learning [7,28]. Then the multi-scale feature generator g_ϕ generates multiple features at different scales. Afterward, semantic relation matrixes are calculated to measure the semantic relevance between query image and support set at each scale. The adaptive task attention module $\mathcal{F}_\varphi$ learns a task attention mask that can adaptively calculate the importance of each LR in the current task. We use task attention masks to weighting the semantic relation matrix to prominently display task-relevant elements. Finally, the weighted semantic relation matrix is processed by similarity-to-class module $\mathcal{F}_\omega$ to determine which support class the query image belongs to. Our proposed MATANet can be trained in an end-to-end mechanism.

3.2 Local Representations Extractor

As some recent studies [15,16] on few-shot learning have proved, LRs show richer representation ability and can alleviate the problem of sample scarcity in few-shot learning. Therefore, we use LRs to represent the features of each image. Given a query image q, we can get a feature representation $\mathcal{F}_\theta(q) \in \mathbb{R}^{C\times H\times W}$ through $\mathcal{F}_\theta$. Under the N-way K-shot few-shot learning setting, there are K images for each support class in a certain task. Through local representations extractor we can get a feature representation of support set $\mathcal{S}$, which can be denoted as $\mathcal{F}_\theta(\mathcal{S}) \in \mathbb{R}^{NK\times C\times H\times W}$. To enhance the discriminant ability of feature representations, we embed a self-attention module into local representations extractor, i.e., Squeeze-and-Excitation Module (SEM) [7] and Convolutional Block Attention Module (CBAM) [28].

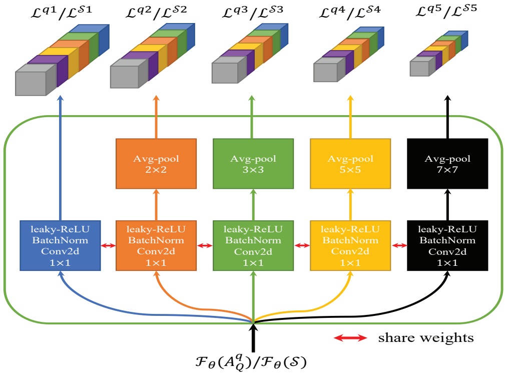
$\mathcal{L}^{q1}/\mathcal{L}^{\mathcal{S}1}$
$\mathcal{L}^{q2}/\mathcal{L}^{\mathcal{S}2}$
$\mathcal{L}^{q3}/\mathcal{L}^{\mathcal{S}3}$
$\mathcal{L}^{q4}/\mathcal{L}^{\mathcal{S}4}$
$\mathcal{L}^{q5}/\mathcal{L}^{\mathcal{S}5}$
Avg-pool 2×2
Avg-pool 3×3
Avg-pool 5×5
Avg-pool 7×7
leaky-ReLU BatchNorm Conv2d 1×1
leaky-ReLU BatchNorm Conv2d 1×1
leaky-ReLU BatchNorm Conv2d 1×1
leaky-ReLU BatchNorm Conv2d 1×1
leaky-ReLU BatchNorm Conv2d 1×1
share weights
$\mathcal{F}_\theta(A_Q^q)/\mathcal{F}_\theta(\mathcal{S})$

Fig. 3. The architecture of the multi-scale feature generator g_ϕ.

3.3 Multi-Scale Feature Generator

The multi-scale feature generator aims to generate multiple features at different scales and eliminate the effect of size on classification. As illustrated in Fig. 3, the multi-scale feature generator consists of five components. Note that, all five components share the 1×1 convolutional layer as a transformation layer.

Through the multi-scale feature generator, for $\mathcal{F}_\theta(q)$, we can get the 3D features $\mathcal{L}^{qz} \in \mathbb{R}^{C\times H_z \times W_z}, z \in \{1,2,3,4,5\}$, which can be regarded as a set of $H_z \times W_z$ C-dimensional LRs

$$\mathcal{L}^{qz} = [x_1, ..., x_{H_z W_z}] \in \mathbb{R}^{C \times H_z W_z} \tag{1}$$

where x_i is the i-th LR. Through multi-scale feature generator we can get the LRs of support set $\mathcal{S}$ as follows

$$\mathcal{L}^{\mathcal{S}z} = [x_1, ..., x_{NKH_z W_z}] \in \mathbb{R}^{C \times NKH_z W_z} \tag{2}$$

Subsequently, we concatenate the local feature descriptors of all scales. Specifically, for every image, the number of local feature descriptors is $M = \sum_{z=1}^{5} H_z \times W_z$. Through concatenation, we can get the set $\mathcal{L}^q$ and the set $\mathcal{L}^S$

$$\mathcal{L}^q = [x_1, ..., x_M] \in \mathbb{R}^{C \times M} \tag{3}$$

$$\mathcal{L}^{\mathcal{S}} = [x_1, ..., x_{NKM}] \in \mathbb{R}^{C \times NKM} \tag{4}$$

3.4 Adaptive Task Attention Module

Under the N-way K-shot few-shot learning setting, we calculate the semantic relation matrix $\mathcal{R}$ between a query image q and support set $\mathcal{S}$ to measure semantic relevance by LRs. Then the $\mathcal{R}$ can be calculated as follows

$$\mathcal{R}_{i,j} = \cos(\mathcal{L}_i^q, \mathcal{L}_j^{\mathcal{S}}) \tag{5}$$

$$\cos(\mathcal{L}_i^q, \mathcal{L}_j^{\mathcal{S}}) = \frac{(\mathcal{L}_i^q)^{\mathrm{T}} \mathcal{L}_j^{\mathcal{S}}}{\|\mathcal{L}_i^q\| \cdot \|\mathcal{L}_j^{\mathcal{S}}\|} \tag{6}$$

where $i \in \{1, ..., M\}, j \in \{1, ..., NKM\}$, $\mathcal{R}_{i,j}$ is the distance between the i-th LR of the query image and the j-th LR of support set and $\cos(\cdot, \cdot)$ is the cosine distance measure function.

Each row in $\mathcal{R}$ represents the semantic similarity of each LR in the query image to all LRs of all images in the support set, i.e., semantic relation vector $\mathcal{R}_i$ represent the relation between i-th LR of query image q to all NKM LRs of support set. $\mathcal{R}$ can be decomposed into N submatrices $\mathcal{R}^n$, $n \in \{1, ..., N\}$ according to columns, representing the semantic relation between the query image and each support class.

Then we can calculate the task attention score of each element of $\mathcal{R}$ for the current task as

$$\alpha_i = \frac{\sum_{j=1}^{NKM} \mathcal{R}_{i,j}}{\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{NKM} \mathcal{R}_{i,j}}} \tag{7}$$

The task attention mask α is consist of all task attention scores α_i, $i \in \{1, ..., M\}$. Afterwards, we use dot-product to weight $\mathcal{R}_i$ by α_i

$$\mathcal{M}_i = \alpha_i \cdot \mathcal{R}_i \tag{8}$$

where $\mathcal{M}_i$ is the i-th row of weighted semantic relation matrix. Thus we can get the weighted semantic relation matrix $\mathcal{M}$, which can be decomposed into N submatrices $\mathcal{M}^n$, $n \in \{1, ..., N\}$ according to columns. While the semantic relations of task-irrelevant regions are suppressed; meanwhile, the semantic relations of task-relevant regions are enhanced.

After the Adaptive Task Attention Module, we use Similarity-to-Class Module to determine which support class the query image belongs to [15]. In this module, for each LR of the query image, we find the k most similar LRs of all support LRs for class n. Then, we sum kM selected LRs as the similarity score between the query image and the n-th support class

$$\mathcal{P}^n = \sum_{i=1}^{kM} \mathrm{Topk}(\mathcal{M}_i^n) \tag{9}$$

where $\mathcal{P}^n$ is the semantic similarity between the query image and support class n, and Topk$(\cdot)$ means collecting the k lagest elements in each row of the weighted semantic relation matrix $\mathcal{M}^n$. Specially, we set k to 7 on the *mini*Imagenet dataset, 3 on the *tiered*Imagenet dataset, and 5 on three fine-grained datasets. Under the N-way K-shot few-shot learning setting, we can get semantic similarity vectors $\mathcal{P} \in \mathbb{R}^N$.

4 Experiments

4.1 Datasets

***mini*ImageNet.** As a small subset of ImageNet [4], the dataset consists of 100 categories, each containing 600 images. We use common splits as in [5], which

Table 1. The splits of three fine-grained datasets.

Dataset	**Stanford Dogs**	**Stanford Cars**	**CUB Birds**
N_{all}	120	196	200
N_{train}	70	130	100
N_{val}	20	17	50
N_{test}	30	49	50

devides the dataset into training, validation and test dataset with 64/16/20 classes respectively.

***tiered*ImageNet.** A subset of ImageNet [4], proposed by [21], the dataset consists of 608 categories, and has a hierarchical structure of categories. We use common splits as in [21], which takes 351, 97 and 160 classes for training, validation and testing, respectively.

We also conduct experiments on few-shot fine-grained image classification tasks. We use three popular fine-grained datasets to conduct experiments.

CUB Birds [27] is composed of 11, 788 images of 200 birds species.

Stanford Dogs [10] is contains 120 categories of dogs and 20, 480 images.

Stanford Cars [13] consists of 196 categories of cars with 16, 185 images. For fair comparisons, we strictly follow the splits used in [15,16] on Stanford Dogs and Stanford Cars, and follow the splits used in [3] on CUB Birds as Table 1 shows.

4.2 Network Architecture

Normally, using deeper or pre-trained feature extractors can achieve better performance ability. We follow the basic feature extractor module which is adopted in previous works, to make a fair comparison with other works. And our local representations extractor $\mathcal{F}_\theta$ consists of four convolutional blocks and one self-attention module. Specifically, each convolutional block consists of a convolutional layer (with 3×3 convolution and 64 (128) filters for the first two blocks (last two blocks)), a norm layer, and a leaky ReLU non-linearity. Moreover, we add a 2×2 max-pooling operation to the first two convolution blocks and embed a self-attention module after the first block. The reason for using only two max-pooling layers is we can get more LRs to capture the semantic relation between them. For example, in a 5-way 1-shot few-shot learning task, if we use four max-pooling layers, we can only get 25 LRs for an 84×84 input image. In contrast, if we only use two max-pooling layers, we will get 441 LRs, which will be helpful for us to find local semantic relations.

4.3 Implementation Details

Our experiments are conducted under the N-way K-shot setting on four benchmarks. All the images in four benchmarks are resized to 84×84. During the

Table 2. The ablation study on *mini*ImageNet for the proposed MATANet-CBAM (K=7). (Red/blue is best/second best performances)

Model	1-shot	5-shot
baseline	50.85	68.67
w/o g_ϕ	53.41	73.45
w/o $\mathcal{F}_\varphi$	52.43	72.17
w/o SAM	53.32	73.19
MATANet	54.14	74.20

Table 3. Average classification accuracy of 5-way 1-shot and 5-way 5-shot tasks with 95% confidence intervals on *tiered*ImageNet. † Results re-implemented in the same setting. (Red/blue is best/second best performances)

Model	Backbone	1-shot	5-shot
Prototypical Nets* [24]	Conv-64F	48.67	69.57
Relation Nets* [25]	Conv-64F	54.48	71.32
DN4* [15]	Conv-64F	53.37	74.45
DeepEMD† [29]	Conv-64F	50.92	66.32
BOIL* [19]	Conv-64F	49.35	69.37
MATANet-SEM (K=3)	Conv-128F	56.35	75.89
MATANet-CBAM (K=3)	Conv-128F	56.66	76.15

training stage, we randomly construct 300,000 episodes to train our MATANet for the *mini*ImageNet, and 300,000 for the other datasets by episodic training mechanism. In each episode, we select 15 or 10 query images from each class for the 1-shot or 5-shot setting, respectively, i.e., in a 5-way 1-shot task, we have 75 query images and 5 support images. We adopt the Adam algorithm [11] with the cross-entropy (CE) loss to train the network. Also, the initial learning rate is set to 0.001 and reduce by half of every 100,000 episodes. 5,000 episodes are constructed from the test set during the test stage. Then the mean accuracy and 95% confidence intervals will be reported simultaneously.

4.4 Ablation Study

To further verify the effectiveness of the multi-scale feature generator, adaptive task-attention module, and similarity-to-class module, we perform an ablation study on *mini*ImageNet. We remove g_ϕ, $\mathcal{F}_\varphi$ and Self-Attention Module from the MATANet respectively to confirm that each part of the model is indispensable. We remove g_ϕ, $\mathcal{F}_\varphi$ and Self-Attention Module simultaneously as the baseline method. As seen in Table 2, the main improvement comes from the adaptive task-attention module $\mathcal{F}_\varphi$. If we remove $\mathcal{F}_\varphi$, the performance will be reduced by 3.2%, 2.7% on 1-shot, 5-shot tasks, respectively. This empirical study proves that the discriminative $\mathcal{F}_\varphi$ gives a performance boost and results in more discriminative features for classification. Similarly, if we remove g_ϕ, the performance will be reduced by 1.3%, 1.0% on 1-shot, 5-shot tasks, respectively. Moreover, if we remove Self-Attention Module, the performance will be reduced by 1.5%, 1.4% on 1-shot, 5-shot tasks, respectively.

4.5 Comparison Against Related Approaches

Our method is compared with related approaches on several popular datasets.

Results on *tiered*ImageNet. The experimental results on *tiered*ImageNet are reported in Table 3. As it can be seen, our proposed MATANet-CBAM

Table 4. Average classification accuracy of 5-way 1-shot and 5-way 5-shot tasks with 95% confidence intervals on *mini*ImageNet. † Results re-implemented in the same setting. (Red/blue is best/second best performances)

Method	Venue	Backbone	Type	1-shot	5-shot
MAML [5]	ICML'17	Conv-32F	Meta	48.70	63.11
BOIL [19]	ICLR'21	Conv-64F	Meta	49.61	66.45
Prototypical Nets [24]	NeurIPS'17	Conv-64F	Metric	49.42	68.20
Relation Nets [25]	CVPR'18	Conv-64F	Metric	50.44	65.32
CovaMNet [16]	AAAI'19	Conv-64F	Metric	51.19	67.65
DN4 [15]	CVPR'19	Conv-64F	Metric	51.24	71.02
SAML [6]	ICCV'19	Conv-64F	Metric	52.22	66.49
DeepEMD† [29]	CVPR'20	Conv-64F	Metric	52.21	65.63
DSN [23]	CVPR'20	Conv-64F	Metric	51.78	68.99
Align(Centroid) [1]	ECCV'20	Conv-64F	Metric	53.14	71.45
Neg-Margin [17]	ECCV'20	Conv-64F	Others	52.68	70.41
MATANet-SEM (K=7)	Ours	Conv-128F	Metric	53.97	74.01
MATANet-CBAM (K=7)	Ours	Conv-128F	Metric	54.14	74.20

(K=3) can consistently perform better than prior art on *tiered*ImageNet under both 5-way 1-shot and 5-way 5-shot settings. Specifically, our method is around 16.4%/9.5%, 4.0%/6.8%, 6.2%/2.3%, 14.8%/9.8% better than Prototypical Nets [24], Relation Nets [25], DN4 [15], BOIL [19] under the 1-shot/5-shot setting, respectively.

Results on *mini*ImageNet. The experimental results on *mini*ImageNet are reported in Table 4. It can be observed that our method significantly outperforms other methods under both 5-way 1-shot and 5-way 5-shot settings. Especially, we are 1.9% better than the second best method [1] under the 5-way 1-shot setting, with an accuracy rate of 54.14%. Similarly, we achieve 74.20% under the 5-way 5-shot setting, with an improvement of 3.8% from the second best method [1]. Note that, our model gains 5.7% and 4.5% improvements over the most relevant work [15] on 1-shot and 5-shot, respectively, which proposes a Deep Nearest Neighbour Network to find the relation at local information level. This improvement verifies the effectiveness of our model, which can adaptively select the most discriminative local features at multiple scales in a certain task.

Results on fine-grained datasets. From Table 5, it can be observed that the proposed MATANet outperforms all other state-of-the-art methods under both 5-way 1-shot and 5-way 5-shot few-shot learning settings. Especially for the 5-way 1-shot task, our method achieves 30.5%, 30.5%, and 24.7% gains over the most relevant work [15] on Stanford Dogs, Stanford Cars, and CUB Birds, respectively. For the 5-way 5-shot task, our method achieves 26.1%, 6.0%, and 2.9% gains over the most relevant work [15] on three datasets.

The reason why our MATANet can outperform other methods is that MATANet can adaptively select the task-relevant LRs at multiple scales for classification.

Table 5. Results on Stanford Dogs, StanfordCars and CUB Birds. For Stanford Dogs and Stanford Cars, † results re-implemented in the same setting for a fair comparison. (Red/blue is best/second best performances)

Model	Stanford Dogs		Stanford Cars		CUB Birds	
	1-shot	5-shot	1-shot	5-shot	1-shot	5shot
Prototypical Nets [24]	37.59	48.19	40.90	52.93	51.31	70.77
GNN [22]	46.98	62.27	55.85	71.25	51.83	63.69
MAML† [5]	44.81	58.68	47.22	61.21	55.92	72.09
Relation Nets† [25]	43.33	55.23	47.67	60.59	62.45	76.11
CovaMNet [16]	49.10	63.04	56.65	71.33	60.58	74.24
DN4 [15]	45.41	63.51	59.84	88.65	52.79	81.45
LRPABN$_{+\text{cpt}}{}^{*}$ [8]	45.72	60.94	60.28	73.29	63.63	76.06
MATANet-SEM (K=5)	58.75	79.83	77.95	93.21	65.56	83.43
MATANet-CBAM (K=5)	59.28	80.11	78.11	94.00	65.84	83.78

5 Conclusion

In this paper, we propose a novel Multi-scale Adaptive Task Attention Network (MATANet) for few-shot learning, aiming to learn more discriminative task-relevant local representations at different scales by generating multiple features at different scales and looking at the context of the entire task. By taking a view of the entire task, our method is able to adaptively select the most discriminative local representations in the current task at different scales. Extensive experiments on four benchmark datasets demonstrate the effectiveness and advantages of the proposed MATANet. More importantly, our method shows great generalize ability on cross-domain few-shot learning tasks.

References

1. Afrasiyabi, A., Lalonde, J., Gagné, C.: Associative alignment for few-shot image classification. In: ECCV. vol. 12350, pp. 18–35 (2020)
2. Allen, K.R., Shelhamer, E., Shin, H., Tenenbaum, J.B.: Infinite mixture prototypes for few-shot learning. In: ICML. vol. 97, pp. 232–241 (2019)
3. Chen, W., Liu, Y., Kira, Z., Wang, Y.F., Huang, J.: A closer look at few-shot classification. In: ICLR (2019)
4. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML. vol. 70, pp. 1126–1135 (2017)
6. Hao, F., He, F., Cheng, J., Wang, L., Cao, J., Tao, D.: Collect and select: Semantic alignment metric learning for few-shot learning. In: ICCV. pp. 8459–8468 (2019)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
8. Huang, H., Zhang, J., Zhang, J., Xu, J., Wu, Q.: Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. arXiv preprint arXiv:1908.01313 (2019)

9. Jiang, W., Huang, K., Geng, J., Deng, X.: Multi-scale metric learning for few-shot learning. TCSVT **31**(3), 1091–1102 (2021)
10. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: CVPR Workshops. vol. 2 (2011)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
12. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Workshops. vol. 2 (2015)
13. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV Workshops. pp. 554–561 (2013)
14. Li, H., Eigen, D., Dodge, S., Zeiler, M., Wang, X.: Finding task-relevant features for few-shot learning by category traversal. In: CVPR. pp. 1–10 (2019)
15. Li, W., Wang, L., Xu, J., Huo, J., Gao, Y., Luo, J.: Revisiting local descriptor based image-to-class measure for few-shot learning. In: CVPR. pp. 7260–7268 (2019)
16. Li, W., Xu, J., Huo, J., Wang, L., Gao, Y., Luo, J.: Distribution consistency based covariance metric networks for few-shot learning. In: AAAI. pp. 8642–8649 (2019)
17. Liu, B., Cao, Y., Lin, Y., Li, Q., Zhang, Z., Long, M., Hu, H.: Negative margin matters: Understanding margin in few-shot classification. In: Proc. Eur. Conf. Comput. Vis. (ECCV). vol. 12349, pp. 438–455 (2020)
18. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. In: ICLR (2019)
19. Oh, J., Yoo, H., Kim, C., Yun, S.Y.: Boil: Towards representation change for few-shot learning. In: ICLR (2021)
20. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
21. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: ICLR (2018)
22. Satorras, V.G., Estrach, J.B.: Few-shot learning with graph neural networks. In: ICLR (2018)
23. Simon, C., Koniusz, P., Nock, R., Harandi, M.: Adaptive subspaces for few-shot learning. In: CVPR. pp. 4135–4144 (2020)
24. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: NeurIPS. pp. 4077–4087 (2017)
25. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR. pp. 1199–1208 (2018)
26. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: NeurIPS. pp. 3630–3638 (2016)
27. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
28. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: ECCV. vol. 11211, pp. 3–19 (2018)
29. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: CVPR. pp. 12200–12210 (2020)