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

Haoxing Chen, Huaxiong Li, Yaohui Li, Chunlin Chen
Department of Control and Systems Engineering, Nanjing University, Nanjing, China
Email: haoxingchen@smail.nju.edu.cn, huaxiongli@nju.edu.cn, yaohuili@smail.nju.edu.cn, clchen@nju.edu.cn

***Abstract*—Few-shot learning has aroused considerable interest in recent years, which aims to recognize unseen categories by using a few labeled samples. In various few-shot methods, pixel-level 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 local representations among the entire task. Finally, a similarity-to-class module is adapted 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. Our source code is available at: https://github.com/chenhaoxing/MATANet.**

I. Introduction

Deep learning based computer vision method has achieved great success in many practical problems [1]–[3]. Most of these methods require a lot of labeled data for training. However, collecting large amounts of label data is time-consuming and labor-intensive. Besides, some fine-grained data [4]–[6] require expert knowledge to be accurately labeled. How to tackle the image classification under the few-shot learning setting accurately remains an open problem [7], [8].

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 [9]–[21] have been proposed. Meta-learning and metric-based learning are two kinds of mainstream few-shot learning methods. Meta-learning based methods [10], [12], [15], [19] focus on how to find a satisfying parameter initialization or optimizers. Metric-learning based methods [11]–[14], [16]–[18], [22] aims to find a more discriminative distance measure to distinguish different categories of samples. However, most of these metric learning methods [11]–[14], [18], [22] 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. Instead, many methods [17], [18] based on pixel-level features were proposed, i.e., Local Descriptors (LDs) 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 [16].

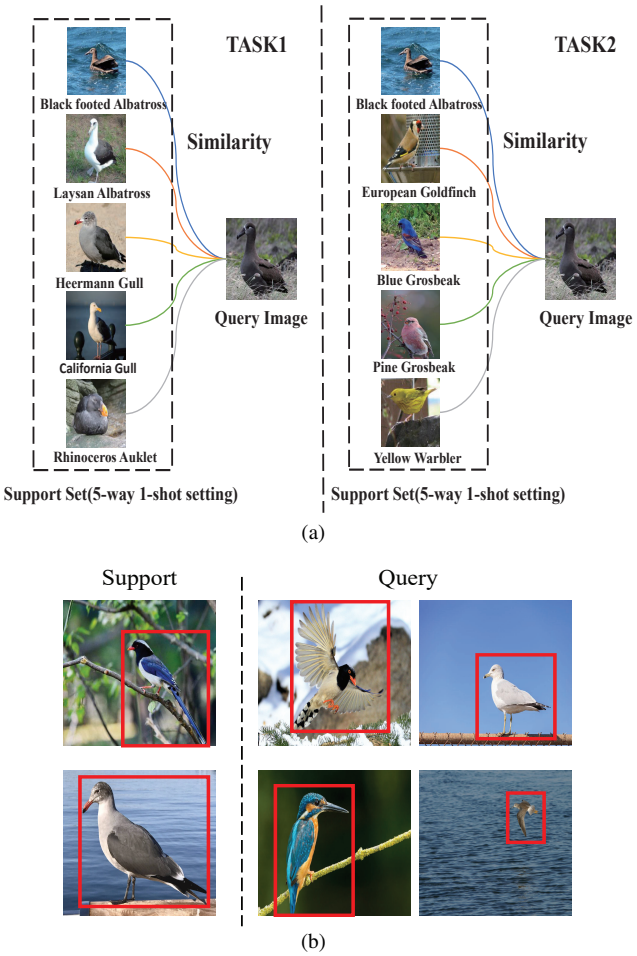

Fig. 1. The two main limitations of the previous local representation based methods. (a) In different tasks, the most discriminative features are different. In task 1, the beak is the key distinguishing feature, while the most critical feature is the wing in task 2. (b) The scales of the dominant objects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

However, these methods ignore the context information of the whole task, which cannot make full use of the representation ability of local feature descriptors. When humans judge the similarity between images, 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, it is required 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 highly 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 LD varies from task to task.

Although many existing methods [16]–[18] can extract the relation between the query image and each support set independently, they do not consider the importance of each LD under the whole task, and all the LDs are weighted equally, *and we argue that the task-relevant LDs should enjoy the higher weights*. Moreover, these methods can only calculate their similarity on 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 feature representations used in [16]–[18] are not discriminative, since CNN treats all features equally [23], [24].

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 LDs 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 query image and support set by calculating the semantic relation matrix. Afterward, we employ an adaptive task attention module to select the most distinguishing feature of the current task. Third, to further make full use of LDs, 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 the 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 current task, aiming to learn task-relevant features.
- Finally, we conduct sufficient experiments on four benchmark datasets to verify the advancement of our model, and the performance of our model achieves state-of-the-art.

II. 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 [11], [13], [14], [16]–[18], [20], [22], [25]. Koch *et al.* [26] migrated the Siamese Network to the one-shot learning task. Snell *et al.* [13] 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.* [14] 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 [18], DN4 [16] and CovaMNet [17] shows that the rich low-level features (i.e., LDs) have better representation capabilities. Hao *et al.* [18] proposed the Semantic Alignment Metric Learning method to find semantic relevance between query images and support sets. Li *et al.* [16] proposed a Deep Nearest Neighbor Network to find the similarity between local feature descriptors. Li *et al.* [17] 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.* [27] 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 [12], [28]. 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.

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.

III. The proposed Method

A. 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 standard definition of few-shot learning image classification task [13]. Given a N-way

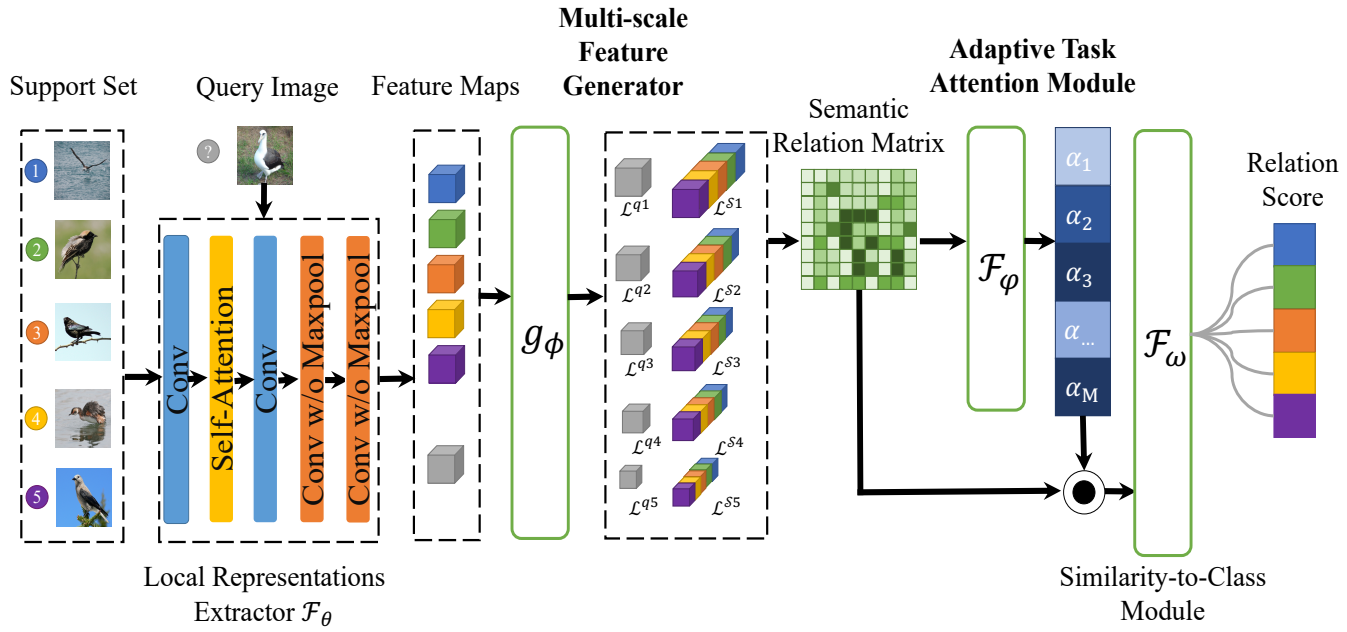

Fig. 2. The architecture of the proposed MATANet. The model consists of four parts: the local representatios extractor $\mathcal{F}_\theta$ to learn local representations and enhance the discriminant ability of feature representations, the multi-scale feature generator g_ϕ to generate multiple features at different scales, the adaptive task attention module $\mathcal{F}_\varphi$ to generate adaptive task attention mask for selecting more important elements of semantic relation matrix, and the similarity-to-class module $\mathcal{F}_\omega$ to get a similarity score to determine which support class the query image belongs to. (Best view in color.)

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 a meta-training set to train a model to learn meta-knowledge. Models are trained by the episodic training mechanism [11]. In each episode, a new task is randomly constructed in $\mathcal{A}$, and each task consists of two subsets: support set $\mathcal{A}_\mathcal{S}$ and 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 LDs. In practice, we choose 4-layer CNN as our feature extractor [12], [16], [28] and we embed the self-attention module after the first block, to enhance the discriminant ability of feature representations for few-shot learning [29], [30]. Then the multi-scale feature generator g_ϕ generates multiple features at different scales. Afterward, semantic relation matrixes are calculated to measure semantic relevance. The adaptive task attention module $\mathcal{F}_\varphi$ learns a task attention mask that can adaptively calculate the importance of each LD in the current task. We use task attention masks to weight the semantic relation matrix to prominently display task-relevant elements. Finally, the weighted semantic relation matrix is processed by the 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.

B. Local Representations Extractor

As some recent studies [16], [17] on few-shot learning have proved, LDs show richer representation ability and can alleviate the problem of sample scarcity in few-shot learning. Therefore, we use LDs 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 the local representations extractor, i.e., Squeeze-and-Excitation Module (SEM) [29] and Convolutional Block Attention Module (CBAM) [30]. We use MATANet-SEM and MATANet-CBAM to represent the models using SEM and CBAM as self-attention modules respectively.

C. 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 shown in Fig. 3, we use five different pooling operations to get multi-scale features. Note that all five components share the 1×1 convolutional layer as a transformation layer.

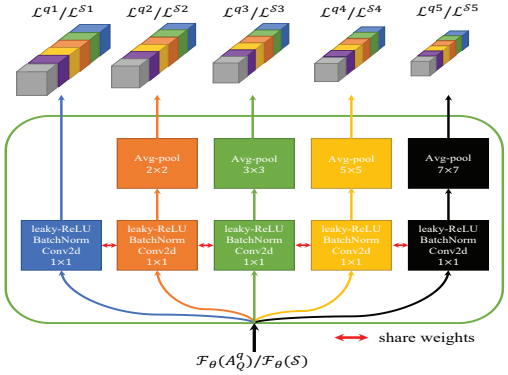

Fig. 3. The architecture of the multi-scale feature generator g_ϕ.

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 LDs

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

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

$$\mathcal{L}^{Sz} = [x_1, ..., x_{NKH_zW_z}] \in \mathbb{R}^{C\times NKH_zW_z} \tag{2}$$

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 $\mathcal{L}^S$

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

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

D. 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 LDs. 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 LD of the query image and the j-th LD of support set and $\cos(\cdot,\cdot)$ is the cosine distance measure function.

Each row in $\mathcal{R}$ represents the semantic similarity of each LD in the query image to all support LDs, i.e., semantic relation vector $\mathcal{R}_i$ represent the relation between the i-th LD of query image q to all NKM LDs 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 [16]. In this module, for each LD of the query image, we find the k most similar LDs of all support LDs for class n. Then, we sum kM selected LDs 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 $\mathrm{Topk}(\cdot)$ 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 $\mathcal{P} \in \mathbb{R}^N$.

IV. Experiments

In this section, we first perform experiments on several popular datasets to verify the effectiveness of the proposed MATANet. Then, we further test the superiority of our approach on a more difficult cross-domain task. Finally, we conducted some comparative experiments and visualization experiments to verify the effectiveness of our proposed method.

A. Datasets

***mini*ImageNet.** As a small subset of ImageNet [33], the dataset consists of 100 categories, each containing 600 images. We use common splits as in [12], which devides the dataset into training, validation and evaluation dataset with 64/16/20 classes respectively.

***tiered*ImageNet.** A subset of ImageNet [33], proposed by [34], the dataset consists of 608 categories, and has a hierarchical structure of categories. We use common splits as in [34], which takes 351, 97 and 160 classes for training, validation and evaluation, respectively.

CUB Birds [4] is composed of 11, 788 images of 200 birds species. We use 100/50/50 categories for training, validation, and evaluation respectively.

TABLE I

AVERAGE CLASSIFICATION ACCURACY OF 5-WAY 1-SHOT AND 5-WAY 5-SHOT TASKS WITH 95% CONFIDENCE INTERVALS ON *mini*IMAGENET AND *tiered*IMAGENET. * RESULTS REPORTED BY THE ORIGINAL WORK. (RED/BLUE IS BEST/SECOND BEST PERFORMANCES)

Model	Backbone	Type	*mini*ImageNet		*tiered*ImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
MAML* [12]	Conv-32F	Meta	48.70±1.84	63.11±0.92	51.67±1.81	70.30±1.75
BOIL [31]	Conv-64F	Meta	49.61±0.16	66.45±0.37	49.35±0.26	69.37±0.12
Matching Nets* [11]	Conv-64F	Metric	43.56±0.84	55.31±0.73		
Prototypical Nets* [13]	Conv-64F	Metric	49.42±0.78	68.20±0.66	48.67±0.87	69.57±0.75
Relation Nets* [14]	Conv-64F	Metric	50.44±0.82	65.32±0.70	54.48±0.93	71.32±0.78
CovaMNet* [17]	Conv-64F	Metric	51.19±0.76	67.65±0.63	54.98±0.90	71.51±0.75
DN4* [16]	Conv-64F	Metric	51.24±0.74	71.02±0.64	53.37±0.86	74.45±0.70
CTM* [27]	Conv-64F	Metric	41.62±0.00	58.77±0.00	-	-
SAML* [18]	Conv-64F	Metric	52.22±0.00	66.49±0.00	-	-
DSN* [20]	Conv-64F	Metric	51.78±0.96	68.99±0.69	-	-
Align(Centroid)* [21]	Conv-64F	Metric	53.14±1.06	71.45±0.72	-	-
Neg-Margin* [32]	Conv-64F	Others	52.68±0.76	70.41±0.66	-	-
MATANet-SEM (K=3)	Conv-128F	Metric	53.36±0.57	73.68±0.51	56.35±0.68	75.89±0.52
MATANet-SEM (K=7)	Conv-128F	Metric	53.97±0.61	74.01±0.48	56.12±0.64	72.14±0.51
MATANet-CBAM (K=3)	Conv-128F	Metric	53.82±0.63	73.78±0.52	56.66±0.70	76.15±0.55
MATANet-CBAM (K=7)	Conv-128F	Metric	54.14±0.62	74.20±0.51	56.21±0.53	72.32±0.54

Stanford Dogs [5] is contains 120 categories of dogs and 20, 480 images. We use 70/20/30 categories for training, validation, and evaluation respectively.

Stanford Cars [6] consists of 196 categories of cars with 16, 185 images. We use 130/17/49 categories for training, validation, and evaluation respectively.

B. Network Architecture

Normally, using deeper or pre-trained feature extractors can achieve better performance ability. To make a fair comparison with other works, we use shallow Conv-64F as our local representations extractor $\mathcal{F}_\theta$. $\mathcal{F}_\theta$ 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 LDs 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 LDs for an 84×84 input image. In contrast, if we only use two max-pooling layers, we will get 441 LDs, which will be helpful for us to find local semantic relations.

C. 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 training stage, we randomly construct 300,000 episodes to train our MATANet. In each episode, we select 15 query images from each class, i.e., in a 5-way 1-shot task, we have 75 query images and 5 support images. We adopt the Adam algorithm [35] with the cross-entropy (CE) loss to train the network. Also, the initial learning rate is set to 0.001 and decay 0.5 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. We used pytorch [36] to implement all the experiments.

D. Comparison Against Related Approaches

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

Results on *mini*ImageNet. The experimental results on *mini*ImageNet are reported in Table I. 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 [21] 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 [21]. Note that, our model gains 5.7% and 4.5% improvements over the most relevant work [16] 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 *tiered*ImageNet. The experimental results on *tiered*ImageNet are reported in Table I. As it can be seen, our proposed MATANet-CBAM (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 9.7%/8.3%, 16.4%/9.5%, 4.0%/6.8%, 3.1%/6.5%, 6.2%/2.3%, 14.8%/9.8% better than MAML [12], Prototypical Nets [13], Relation Nets [14], CovaMNet [17], DN4 [16], BOIL [31] under the 1-shot/5-shot setting, respectively.

Results on fine-grained datasets. From Table II, 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-

TABLE II
RESULTS ON STANFORD DOGS, STANFORDCARS AND CUB BIRDS. (RED/BLUE IS BEST/SECOND BEST PERFORMANCES)

Model	Stanford Dogs		Stanford Cars		CUB Birds	
	1-shot	5-shot	1-shot	5-shot	5-shot	5-shot
Matching Nets [11]	35.80±0.99	47.50±1.03	34.80±0.98	44.70±1.03	61.16±0.89	72.86±0.70
Prototypical Nets [13]	37.59±1.00	48.19±1.03	40.90±1.01	52.93±1.03	51.31±0.91	70.77±0.69
GNN [25]	46.98±0.98	62.27±0.95	55.85±0.97	71.25±0.89	51.83±0.98	63.69±0.94
MAML [12]	44.81±0.34	58.68±0.31	47.22±0.39	61.21±0.28	55.92±0.95	72.09±0.76
adaCNN [37]	41.87±0.42	53.93±0.44	42.14±0.41	50.12±0.34	56.57±0.47	61.21±0.42
PCM [38]	28.78±2.33	46.92±2.00	29.63±2.38	52.28±1.46	42.10±1.96	62.48±1.21
Relation Nets [14]	43.33±0.42	55.23±0.41	47.67±0.47	60.59±0.40	62.45±0.98	76.11±0.69
CovaMNet [17]	49.10±0.76	63.04±0.65	56.65±0.86	71.33±0.62	60.58±0.69	74.24±0.68
DN4 [16]	45.41±0.76	63.51±0.62	59.84±0.80	88.65±0.44	52.79±0.86	81.45±0.70
$\text{PABN}_{+\text{cpt}}$ [39]	45.65±0.71	61.24±0.62	54.44±0.71	67.36±0.61	63.56±0.79	75.35±0.58
$\text{LRPABN}_{+\text{cpt}}$ [39]	45.72±0.75	60.94±0.66	60.28±0.76	73.29±0.58	63.63±0.77	76.06±0.58
Temperature Net [40]	49.53±0.00	63.37±0.00	57.87±0.00	73.84±0.00	-	-
Align(Adversarial) [21]	-	-	-	-	63.30±0.94	81.35±0.67
MATANet-SEM (K=5)	58.75±0.65	79.83±0.48	77.95±0.60	93.21±0.29	65.56±0.68	83.43±0.43
MATANet-CBAM (K=5)	59.28±0.67	80.11±0.49	78.11±0.62	94.00±0.25	65.84±0.69	83.78±0.45

shot task, our method achieves 30.5%, 30.5%, and 24.7% gains over the most relevant work [16] 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 [16] on three datasets.

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

E. 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 component is indispensable. We remove g_ϕ, $\mathcal{F}_\varphi$ and Self-Attention Module simultaneously as the baseline method. As seen in Table III, 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 our adaptive task attention module can generate 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.

TABLE III
THE ABLATION STUDY ON *mini*IMAGENET FOR THE PROPOSED MATANET. (RED/BLUE IS BEST/SECOND BEST PERFORMANCES)

Model	5-Way Accuracy(%)	
	1-shot	5-shot
baseline	50.85±0.63	68.67±0.52
w/o g_ϕ	53.41±0.61	73.45±0.53
w/o $\mathcal{F}_\varphi$	52.43±0.62	72.17±0.51
w/o Self-Attention Module	53.32±0.61	73.19±0.49
MATANet-CBAM (K=7)	54.14±0.62	74.20±0.51

TABLE IV
THE NUMBER OF TRAINABLE PARAMETERS CONTAINED IN DIFFERENT METRIC-LEARNING BASED METHODS AND THE CORRESPONDING CLASSIFICATION ACCURACIES OF 5-WAY 1-SHOT TASK ON *mini*IMAGENET. (BEST PERFORMANCES ARE IN BOLD)

Model	Params	Accuracy(%)
Prototypical Net	**0.113M**	49.42±0.78
Relation Net	0.229M	50.44±0.82
GNN	1.619M	50.33±0.36
DN4	**0.113M**	51.24±0.74
MATANet-CBAM (K=7)	0.117M	**54.14±0.83**

F. Complexity Analysis

As shown in Table IV, MATANet does not introduce any extra trainable parameters except for the self-attention module. However, the proposed MATANet only introduces a small number of the trainable parameters, while achieves a better result than the methods above.

V. CONCLUSION

In this paper, we propose a novel Multi-scale Adaptive Task Attention Network (MATANet) for few-shot learning, aiming to generate task-relevant local descriptors at different scales by generating multiple features at different scales and looking at the context of the entire task. By capturing the context information of the current task, our method is able to adaptively select the most discriminative local representations in the current task at different scales. Extensive experiments on five benchmark datasets demonstrate the effectiveness and advantages of the proposed MATANet.

ACKNOWLEDGEMENTS

This work was supported partially by the National Natural Science Foundation of China (Nos. 62176116, 62073160, 71732003), and the National Key Research and Development Program of China (Nos. 2018YFB1402600).

REFERENCES

[1] C. Zhang, H. Li, Y. Qian, C. Chen, and X. Zhou, "Locality-constrained discriminative matrix regression for robust face identification," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 3, pp. 1254–1268, 2022.

[2] C. Zhang, H. Li, C. Chen, Y. Qian, and X. Zhou, "Enhanced group sparse regularized nonconvex regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2438–2452, 2022.

[3] Q. Zhang, R. Cong, C. Li, M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[5] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshops*, vol. 2, no. 1, 2011.

[6] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proc. ICCV Workshops*, 2013, pp. 554–561.

[7] M. Chen, Y. Fang, X. Wang, H. Luo, Y. Geng, X. Zhang, C. Huang, W. Liu, and B. Wang, "Diversity transfer network for few-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10 559–10 566.

[8] V. N. Nguyen, S. Løkse, K. Wickstrøm, M. Kampffmeyer, D. Roverso, and R. Jenssen, "SEN: A novel feature normalization dissimilarity measure for prototypical few-shot learning networks," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12368, 2020, pp. 118–134.

[9] F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, 2006.

[10] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[11] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3630–3638.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1126–1135.

[13] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[14] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[15] Q. Sun, Y. Liu, T. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 403–412.

[16] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7260–7268.

[17] W. Li, J. Xu, J. Huo, L. Wang, Y. Gao, and J. Luo, "Distribution consistency based covariance metric networks for few-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8642–8649.

[18] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8459–8468.

[19] M. A. Jamal and G. Qi, "Task agnostic meta-learning for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 719–11 727.

[20] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4135–4144.

[21] A. Afrasiyabi, J. Lalonde, and C. Gagné, "Associative alignment for few-shot image classification," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12350, 2020, pp. 18–35.

[22] K. R. Allen, E. Shelhamer, H. Shin, and J. B. Tenenbaum, "Infinite mixture prototypes for few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 232–241.

[23] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5219–5227.

[24] W. Luo, X. Yang, X. Mo, Y. Lu, L. Davis, J. Li, J. Yang, and S. Lim, "Cross-x learning for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8241–8250.

[25] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[26] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Workshops*, vol. 2, 2015.

[27] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1–10.

[28] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[30] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, vol. 11211, 2018, pp. 3–19.

[31] J. Oh, H. Yoo, C. Kim, and S.-Y. Yun, "Boil: Towards representation change for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[32] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12349, 2020, pp. 438–455.

[33] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[34] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[37] T. Munkhdalai, X. Yuan, S. Mehri, and A. Trischler, "Rapid adaptation with conditionally shifted neurons," in *Proc. Int. Conf. Mach. Learn.*, J. G. Dy and A. Krause, Eds., vol. 80, 2018, pp. 3661–3670.

[38] X. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6116–6125, 2019.

[39] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, "Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification," *IEEE Trans. Multim.*, vol. 23, pp. 1666–1680, 2021.

[40] W. Zhu, W. Li, H. Liao, and J. Luo, "Temperature network for few-shot learning with distribution-aware large-margin metric," *Pattern Recognit.*, vol. 112, p. 107797, 2021.

[41] P. Tokmakov, Y. Wang, and M. Hebert, "Learning compositional representations for few-shot recognition," in *ICCV*, 2019, pp. 6371–6380.