BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 71(1), 2023, Article number: e144584
DOI: 10.24425/bpasts.2023.144584

A few-shot fine-grained image recognition method

Jianwei WANG[1,2] **and Deyun CHEN**[1*]

[1] College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[2] College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin 150050, China

Abstract. Deep learning methods benefit from data sets with comprehensive coverage (e.g., ImageNet, COCO, etc.), which can be regarded as a description of the distribution of real-world data. The models trained on these datasets are considered to be able to extract general features and migrate to a domain not seen in downstream. However, in the open scene, the labeled data of the target data set are often insufficient. The depth models trained under a small amount of sample data have poor generalization ability. The identification of new categories or categories with a very small amount of sample data is still a challenging task. This paper proposes a few-shot fine-grained image recognition method. Feature maps are extracted by a CNN module with an embedded attention network to emphasize the discriminative features. A channel-based feature expression is applied to the base class and novel class followed by an improved cosine similarity-based measurement method to get the similarity score to realize the classification. Experiments are performed on main few-shot benchmark datasets to verify the efficiency and generality of our model, such as Stanford Dogs, CUB-200, and so on. The experimental results show that our method has more advanced performance on fine-grained datasets.

Key words: few-shot learning; attention metric; CNN (convolutional neural network); feature expression.

1. INTRODUCTION

in the application field based on deep learning, a backbone is usually trained on large datasets, such as ImageNet and COCO. Then the backbone is fine-tuned on the training set of another new dataset, such as Cifar and Cub, and the model is tested on the test set [1, 2]. However, in many cases, the image data in the training dataset are different from those in the fine-tuning dataset not only in the domain but also in the category. Due to different categories, the original network classification layer cannot be used during fine-tuning. Because of the different domains, the feature extracted by the backbone is not discriminative enough. Few-shot learning aims to learn new knowledge on the base of a few labeled data, which has attracted researchers' attention in the past two years for its application requirements in the real world [3].

For few-shot learning, a model is trained to recognize an object with a small amount of labeled data. Many methods have been proposed to improve the performance of few-shot learning, such as Siamese neural networks [4], prototypical networks [5], meta-learning [6–8], metric learning [9], and so on. Siamese neural networks are twin networks with shared weight matrices at each layer and are trained to discriminate between a collection of the same/different pairs. Then it is generalized to evaluate new categories based on learned feature mappings. Prototypical networks aim to learn a metric space and classify it by calculating the distance from the prototype representation of each category. Meta-learning tries to learn how to deal with new tasks by learning multiple tasks. Metric learning focuses on learning a good feature representation or relation measure [10]. Recently, few-shot learning methods based on metric learning have drawn more attention for their simplicity and effectiveness. Peng *et al*. proposed a novel Knowledge Transfer Network architecture (KTN) for few-shot image recognition [11]. They learn the metric model on the base of plentiful samples to spur query samples to be close to the supporting samples and generalize it to novel classes [12–14]. Although research [15, 16] about few-shot fine-grained recognition has been carried on, the few-shot classification on fine-grained data is still a difficult problem, which is shown in Fig. 1.

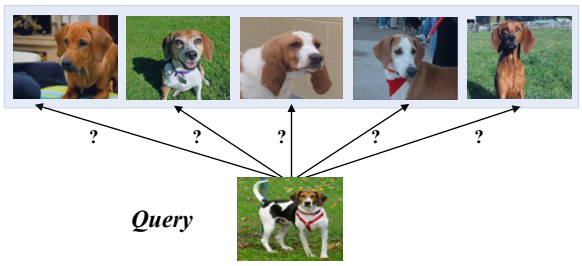

Fig. 1. Problems with fine-grained image recognition. The dog images in support set are from different subdivided species and they are highly similar except for some detail local features, such as ears, eyes, mouths

The visual representation and metric methods are the crucial techniques of metric learning-based few-shot learning methods. A visual representation with strong generalization ability makes a model still perform well when encountering extremely strange or a small amount of labeled data. Recently, employing deep feature representations (i.e., Conv-4 and ResNet-12)

*e-mail: chendeyunahrbust.edu.cn

Manuscript submitted 2022-07-22, revised 2022-12-28, initially accepted for publication 2023-01-10, published in February 2023.

for few-shot learning tasks has been verified to be more expressive and effective than using global features. Although these methods have achieved promising performance, the visual feature description methods are inadequately considered because some more important local features are not emphasized, especially for some fine-grained images. Since local features can provide discriminative information across categories, which is important for image classification in the few-shot scenario, a desirable metric-based algorithm should have the ability to utilize the discriminative representations for metric learning and minimize the impact caused by the irrelevant regions.

Metric learning methods embed samples into vector space and compute the similarity score by a defined similarity function for classification. The similarity function is usually defined according to the distance between the embeddings of the test images and training images (e.g., Euclidean distance and Cosine distance, etc.). It is very important to define a similarity function that is suitable for specific tasks to improve the performance of the model. Snellet *et al*. analyzed the underlying distance function used in order to justify the use of sample means as prototypes [5]. They found Euclidean distance performed better than the more traditional cosine metric. However, the choice of the Euclidean metric was based on assumptions of uncorrelated feature dimensions and uniform variance. Recently, researchers suggest that it is problematic that Euclidean distance is insensitive to the distribution of within-class samples with respect to their prototype [5, 17].

Since visual representation and the metric learning methods are essential for few-shot image classification, especially for fine-grained image classification, this paper focuses on feature extraction and the metric learning methods of few-shot image classification. in this paper, we propose a few-shot fine-grained image recognition method, which extracts the discriminative features of the query image and the support set, aiming to obtain the most related features to the task. Calculate the similarity scores and then renew the weights of channels to form the final similarity scores. The main contributions and works are as follows:

- This paper provides a few-shot fine-grained image recognition method based on attention and metric learning to improve the performance of fine-grained image classification and recognition in the case of a small number of samples.
- A channel-based feature expression is applied by embedding an attention network to emphasize the discriminative features of the base class and novel class. The proposed metric measurement not only pays attention to the relationship of image context but also emphasizes the importance of local features.
- Experiments on fine-grained image datasets (i.e., Stanford Dogs, CUB-200, and so on) show that our proposed method achieved outperformance compared with the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 introduces the proposed method. Section 4 shows the experiment results. Section 5 shows the experimental analysis, and Section 6 summarizes our conclusions.

2. RELATED WORKS

2.1. Visual representation

Before deep learning was widely used, researchers designed visual representation methods based on image gradient information, such as SIFT (scale-invariant feature transform) [18], HOG (histogram of oriented gradient) [19], and so on, which are still widely used nowadays. Since 2012, the visual information representation obtained through a deep neural network called Imagenet has been widely used in many tasks, including image segmentation, object tracking, human pose estimation or extraction, geometric information extraction, and even medical image processing [20]. The new visual information representation learned from the classification task replaces the original visual information representation carefully designed for specific tasks, and achieves very high accuracy. Recently, channel-level and pixel-level representation are considered together in deep learning to achieve a comprehensive visual representation recently [21].

2.2. Visual attention

The attention mechanism is used in deep learning to simulate the characteristics of human attention to things [22]. It can be broadly understood as focusing on part of the input for a specific task rather than seeing the entire input [23]. in 2018, Wang *et al*. proposed a method that presented non-local operations as a generic family of building blocks for capturing long-range dependencies [24]. Hu *et al*. proposed a novel architectural unit termed the "Squeeze-and-Excitation" (SE) block, which adaptively recalibrated channel-wise feature responses by explicitly modelling interdependencies between channels. It produced significant performance improvements at a minimal additional computational cost [25]. Woo *et al*. proposed a convolutional block attention module (CBAM), a simple and effective attention module that can be integrated with any feed-forward convolutional neural network [26]. Unlike previous works that capture contexts by multi-scale feature fusion, Fu *et al*. proposed a dual attention network (DANet) to adaptively integrate local features with their global dependencies [27]. Jiang *et al*. devised a simple and efficient meta-reweighting strategy to adapt the sample representations and generated soft attention to refining the representation such that the relevant features from the query and support samples can be extracted for a better few-shot classification [21].

2.3. Metric learning

Distance metric learning is to learn a distance metric for the input space of data from a given collection of pairs of similar/dissimilar points that preserves the distance relation among the training data [21]. Learning a good distance metric in feature space is crucial to image classification tasks of real-world applications. Tang *et al*. proposed a new generation operator BlockMix by integrating interpolation on the images and labels within metric learning [28]. Common distance functions include Euclidean distance, standardized Euclidean distance, Mahalanobis distance, Cosine similarity, and so on. Global distance metric learning attempts to learn metrics that keep all the data points within the same classes close while separat-

ing all the data points from different classes far apart. in [29], a global distance metric was learned to minimize the distance between the data pairs in the equivalence constraints subject to the constraint and separate the data pairs in the inequivalence constraints. in addition to general-purpose algorithms for distance metric learning, some approaches tried to find feature weights that are adapted to individual test examples. Vinyals *et al*. used cosine distance to measure the gap between features [30]. Zhang *et al*. divided the image into multiple blocks and then introduces the earth mover's distance (EMD) [31].

3. THE METHOD

In this section, we first provide the problem formulations and then present a framework of our network, introducing the visual feature representation, the channel attention module which captures discriminative information in the channel dimension, and the loss function.

The few-shot classification model is required to acquire knowledge from the support set and classify the query samples accurately. Given a few-shot classification task denoted as T, there are a set of support samples denoted as S and a batch of query samples denoted as Q. Train an N-way classifier on a K-shot support set S, where $K$is a small number of training samples per class (e.g., $K = 1$, $K = 5$ or $K = 10$) and N is the number of classes in S. Then test the classifier on a query set Q.

In order to accurately extract the local detail features of fine-grained image data with strong distinguished ability and improve the classification performance, we designed a few-shot classification framework, which is shown in Fig. 2. Visual feature representation based on channel with cosine similarity metric was applied.

We designed a channel attention module to capture the channel dependencies between any two-channel maps by using a similar self-attention mechanism and updating each channel map with a weighted sum of all channel maps. It significantly improved the performance of classification by modeling rich contextual dependencies over local features.

Local feature description was proved to be an efficient method of obtaining the essential representations of a given class of images [25, 32]. Since translating the local features of an image into a compact image-level representation could lose considerable discriminative information [33], local feature description with image-to-class measure was applied.

3.1. Visual feature representation

Feature vectors of images extracted through a feature extractor are expressed as $H \times W \times C$ tensor. A N-way K-shot problem means there are N classes with K samples for each class in the support set. Given a support set denoted as S_n, where $n = \{1,\ldots,N\}$, the feature can be represented as F_{S_n}, and $F_{S_n} \in \mathbb{R}^{K\times C\times P}$, C is the number of channels, H and W is height and width of feature maps, respectively: $P = H \times W$.

Each channel map of features can be regarded as a class-specific response, and different semantic responses are associated with each other [28]. For the few-shot image classification problem, each channel map of features is corresponding to the local semantic representation of the image. The feature vectors of the query set can be represented as F^C_{query}, where C is the number of channels, $F^C_{\text{query}} = [X_1, X_2, \cdots, X_k] \in \mathbb{R}^{P\times C}$, and X_k is the k-th channel feature descriptor.

Channel-based feature of the support set is denoted as $F^C_{S_n}$, where $F^C_{S_n} = [Y_1, Y_2, \cdots, Y_k] \in \mathbb{R}^{P\times KC}$, Y_k is the k-th channel feature descriptor of the support class S_n. We embedded the SENet into the framework to obtain the discriminative image features, which are shown in Fig. 3.

The SE module determines the channel weight by the squeeze and excitation operations. Squeeze operation compresses the feature in the spatial dimension to get a $1 \times 1 \times C$ channel description. The feature map has a global receptive field.

We compress the entire spatial information on a channel into a global feature, and finally get C global features, which are implemented by global average pooling. The formula is:

$$Z_k = \frac{1}{P}\sum_{i=1}^{H}\sum_{j=1}^{W} X_k(i,j). \tag{1}$$

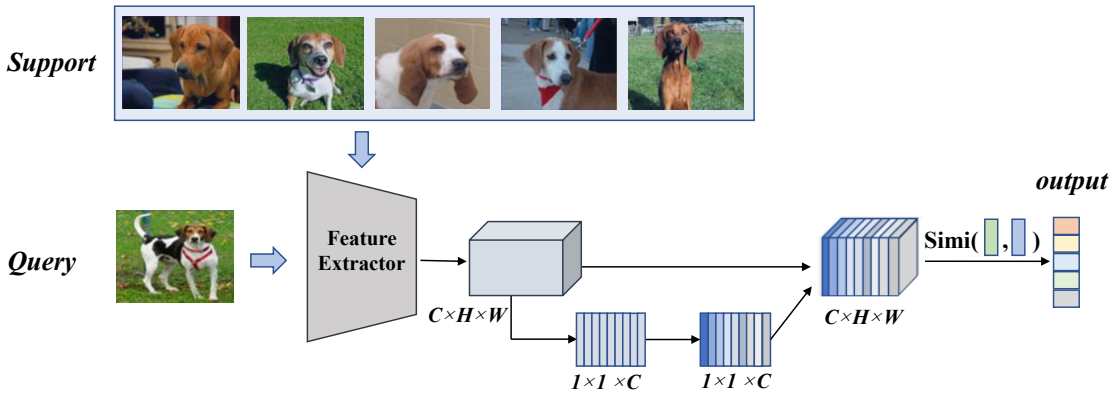

Fig. 2. Few-shot fine-grained image recognition method based on attention and cosine similarity-based metric. It provides an example of few-shot classification of fine-grade data. There are images from 5 subdivided dog species in a support set and 5 or 10 samples in a query set. An attention module is embedded in the model to enhance feature representation. Cosine similarity-based metric is used to evaluate the similarity between the queried sample and samples in the support set

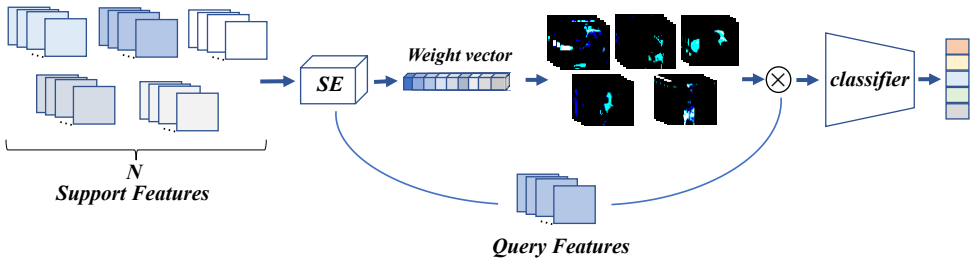

Fig. 3. Weighting feature by the squeeze and excitation operations. Each channel gets a scalar through the global average pool, and C channels get C scalars. C scalars between 0 and 1 are used as the weight of the channel after a full connection, ReLU, full connection and Sigmoid operations. The weight corresponding to each channel of the original output channel is weighted to obtain the new weighted feature maps

Learn the relationship between channels through a full connection, ReLU, another full connection, and sigmoid operations. The first full connection compresses $1 \times 1 \times C$ into $1 \times 1 \times C/r$ (the best effect is when r is 16), and the second full connection is expanded into $1 \times 1 \times C$. Then it is applied to activate the function sigmoid (make the value between 0–1) to get the weight matrix.

We multiply the learned weight coefficients of each channel by all the elements of the corresponding channel to enhance the important features, weaken the unimportant features, and make the extracted features more directional.

3.2. Cosine similarity-based metric

Learning similarity aims to develop a well-defined similarity metric which can fit the maps well. For the d-dimensional input space, two arbitrary patterns are denoted as x and y, and the class labels of x and y are l_x and l_y, respectively. $S(x,y)$ is the similarity function. The intrinsic model of a similarity learning problem can be defined as a map. If $S(x,y) \to 1$, it is determined that x and y are similar. Otherwise, x and y are dissimilar.

Cosine measures similarity as the angle between two vectors, which is shown in Fig. 4. It has the advantage of not being sensitive to magnitudes and it is particularly used in high-dimensional positive spaces to perform tasks such as information retrieval and data mining.

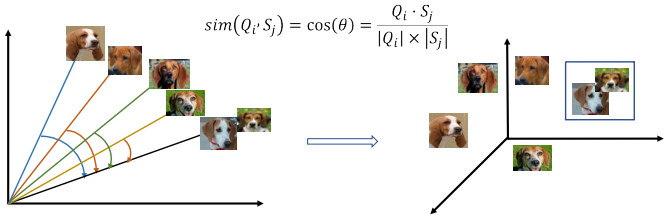

Fig. 4. The cosine similarity measure. The cosine similarity between two patterns belonging to the same category is tend to 1

The cosine similarity of two patterns x and y is defined by:

$$\cos(\theta) = \frac{\sum_{i=1}^{d} x_i \times y_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \times \sqrt{\sum_{i=1}^{d} y_i^2}}, \tag{2}$$

where θ is the angle between x and y. The similarity between these patterns increases as $\cos(\theta)$ increases. in order to make full use of the cosine similarity metric, it is advised to take the different scales between the two patterns into consideration and subtract the corresponding average from each pattern. The improved cosine similarity of x and y is defined as:

$$\cos'(\theta) = \frac{\sum_{i=1}^{d} (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{d} (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{d} (y_i - \bar{y})^2}}, \tag{3}$$

where $\bar{x} = (1/d)\sum_{i=1}^{d} x_i$ and $\bar{y} = (1/d)\sum_{i=1}^{d} y_i$. The similarity between the two images is independent of light intensity by applying the adjusted cosine similarity metric.

We calculate the cosine similarity between the query image and the k-th sample images of the j-th category of the support set. The cosine similarity metric method is defined as:

$$C_k^j(X, Y_j) = \frac{V_w(X - \overline{X})^T \cdot M(Y_j - \overline{Y_j})}{\|V_w(X - \overline{X})\| \|M_k(Y_j - \overline{Y_j})\|}. \tag{4}$$

$V_w(X - \overline{X})$ is the description of the query image by weighted channel features, $M(Y_j - \overline{Y_j})$ describes the k-th sample images of the j-th category of the support set by weighted channel features, $k \in [1, \cdots, K]$ and $j \in [1, \cdots, N]$. $V_w(X - \overline{X})$ and $M_k(Y_j - \overline{Y_j})$ are defined as:

$$V_w(X - \overline{X})^T = \left[w_1(X_1 - \overline{X}), \cdots, w_i(X_i - \overline{X})\right], \tag{5}$$

and

$$M_k\left(Y_j - \overline{Y_j}\right)^T = \left[w_1\left(Y_{11} - \overline{Y_{1k}}\right), \cdots, w_i\left(Y_{i1} - \overline{Y_{ik}}\right)\right], \tag{6}$$

where X_i is the description of the query image by the i-th channel feature, Y_{ij} is the description of the k-th sample images in support set by the i-th channel feature, $k \in [1, \cdots, K]$ and $i \in [1, \cdots, C]$.

$\|V_w(X - \overline{X})\|$ and $\|M_k(Y_j - \overline{Y_j})\|$ are definedas:

$$\left\|V_w\left(X - \overline{X}\right)\right\| = \sqrt{V_w\left(X - \overline{X}\right)^T \cdot V_w\left(X - \overline{X}\right)}, \tag{7}$$

and

$$\left\|M_k\left(Y_j-\overline{Y_j}\right)\right\|=\sqrt{M_k\left(Y_j-\overline{Y_j}\right)^T\cdot M_k\left(Y_j-\overline{Y_j}\right)}. \quad (8)$$

The similarity between the query image and the images of the j-th category of the support set is calculated by:

$$C^j(X,Y_j)=\frac{1}{K}\sum_{k=1}^{K}C_k^j(X,Y_j). \quad (9)$$

For each X, there is Y_j that is most similar to X, and $C^j(X,Y_j)\to 1$.

4. EXPERIMENT

To evaluate the proposed method, we carry out comprehensive experiments on MiniImageNet [30], CUB-200 [34], Stanford Dogs [35], and Stanford Cars [36]. We report our results on these four typical databases in the next subsections after introducing the datasets and implementation details.

4.1. Datasets

Figure 5 shows a few samples from MiniImageNet, CUB-200, Stanford Dogs, and Stanford Cars.

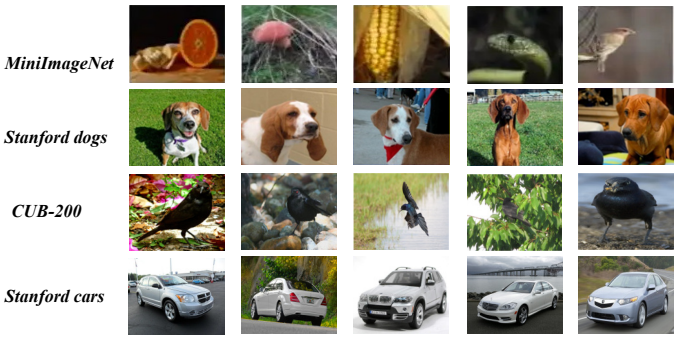

Fig. 5. Samples from MiniImageNet, CUB-200, Stanford Dogs, and Stanford Cars. Unlike MiniImageNet, Stanford Dogs, CUB-200 and Stanford Cars are fine-grained datasets used for few-shot classification

- **MiniImageNet.** The dataset has 60,000 images from 100 different categories and each category has 600 samples. There are 64 categories in the training set, 16 categories in the validation set, and 20 categories in the testing set.
- **CUB-200.** It is a fine-grained benchmark dataset. The dataset has 11 788 images from subdivided species of birds. There are 100 categories in the training set, 50 categories in the validation set, and 50 categories in the testing set.
- **Stanford Dog.** It is a fine-grained benchmark dataset. The dataset has 20 580 images from 120 subdivided species of dogs. There are 70 categories in the training set, 20 categories in the validation set, and 30 categories in the testing set.
- **Stanford Cars.** It is a fine-grained benchmark dataset. The dataset has 16 185 images from 196 different types of cars according to brands, models, and years. There are 130 categories in the training set, 17 categories in the validation set, and 49 categories in the testing set.

4.2. Implementation details

The model is verified on 5-way 1-shot and 5-way 5-shot tasks. We implement our model by pytorch. Cross entropy loss is used to train our model. During the training phase, the episode training mechanism is used to realize the end-to-end training. The training epochs are set to 180 and the batch size is set to 8. The initial learning rate is set to 0.01, and reduced by half for every 10 epochs. The number of query samples per class is set to 10.

4.3. Experimental results

- **Experimental results on MiniImageNet**

The experimental results on MiniImageNet are shown in Fig. 6.

The results of other algorithms are shown in Table 1, where the confidence intervals are 95%.

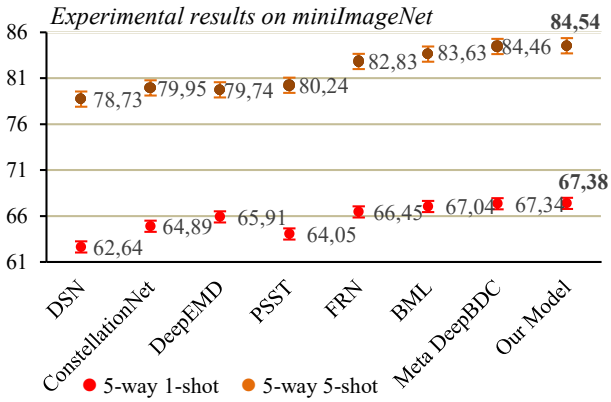

Fig. 6. The experimental results on MiniImageNet. Comparison with state-of-the-art methods (i.e., DSN [37], ConstellationNet [38], DeepEMD [31], PSST [39], FRN [40], BML [41] and MetaDeepBDC [42]) with 95% confidence intervals

Table 1
Experimental results on MiniImageNet

Method	5-way 1-shot	5-way 5-shot
DeepEMD [31]	65.91 ± 0.82	79.74 ± 0.56
PSST [39]	64.05 ± 0.49	80.24 ± 0.45
FRN [40]	66.45 ± 0.19	82.83 ± 0.13
BML [41]	67.04 ± 0.63	83.63 ± 0.29
MetaDeepBDC [42]	67.34 ± 0.43	84.46 ± 0.28
Our model	67.38 ± 0.39	84.54 ± 0.42

We applied ResNet-12 as the backbone which has 4 residual blocks with 3 convolutional layers in each residual block. Each of the first three residual blocks is followed by a maximum pool layer with a kernel size of 2×2, and a global average pool layer follows the last residual block. Figure 6 shows that our model achieved 67.38% accuracy on a 5-way 1-shot task and 84.54 % accuracy on a 5-way 5-shot task. Compared with other methods, the accuracy is almost the same with the MetaDeepBDC, which has achieved fairly good performance on the 5-way 1-shot and 5-way 5-shot tasks.

• **Experimental results on Stanford Dogs**

Stanford Dogs is a commonly used fine-grained dataset. The dog images in the support set are from different subdivided species and they are highly similar excepting some details of local features, such as the ears, eyes mouths, which are shown in Fig. 1. The discriminative fine-grained features, such as gender, age, posture, and so on, are difficult to be extracted because they are too similar to be recognized even by human beings. We applied the Conv-4 as the backbone of the feature extractor, and re-implement the codes of ProtoNet, MatchingNet, RelationNet, GNN, MAML, DN4, and CovaMNet. The confidence intervals are 95%. The experimental results on Stanford Dogs are shown in Table 2.

Table 2
Experimental results on Stanford Dogs

Method	5-way 1-shot	5-way 5-shot
ProtoNet	36.42 ± 0.67	50.22 ± 0.34
MatchingNet	38.62 ± 0.58	47.31 ± 0.37
RelationNet	42.89 ± 0.62	57.01 ± 0.40
GNN	44.62 ± 0.76	46.98 ± 0.27
MAML	44.70 ± 0.98	61.11 ± 0.67
DN4	44.83 ± 0.65	63.02 ± 0.56
CovaMNet	48.21 ± 0.62	62.93 ± 0.73
Our model	56.78 ± 0.89	69.23 ± 0.25

Although the experiments are carried out on a shallow backbone network, the effectiveness of our model can still be proved. The accuracy of our model carried on the Stanford Dogs dataset on 5-way 1-shot task and 5-way 5-shot task are 56.78% and 69.23%, respectively, which are significantly higher than other methods. Because there are more samples of each class in the 5-way 5-shot task than in the 5-way 1-shot task, the accuracy obtained from the 5-way 5-shot task is higher than that obtained from the 5-way 1-shot task by 21.93%.

• **Experimental results on CUB-200**

CUB-200 is another commonly used fine-grained dataset and is also a challenging fine-grained dataset. Parts of the images in CUB-200 are full of noise, occlusion, or light problems, which are shown in Fig. 5. We applied the Conv-4 as the backbone of the feature extractor, and re-implement the codes of ProtoNet, GNN, DN4, MAML, MatchingNet, CovaMNet, and RelationNet. The confidence intervals are 95%. The experimental results on CUB-200 are shown in Table 3.

Table 3 shows that our model achieved better performance than the other 7 mainstream methods on both the 5-way 1-shot task and the 5-way 5-shot task. Since the improved cosine similarity metric is independent of light intensity and the channel pays more attention to extracting the discriminative features, our model shows strong advantages in lighting, occlusion, and other problems. For the 5-way 5-shot task, since there are more samples of each class, our model achieved 80.06% accuracy in the case of noise.

Table 3
Experimental results on CUB-200

Method	5-way 1-shot	5-way 5-shot
ProtoNet	45.64 ± 0.64	71.73 ± 0.48
GNN	51.83 ± 0.52	64.59 ± 0.45
DN4	54.65 ± 0.61	78.64 ± 0.47
MAML	55.24 ± 0.49	72.18 ± 0.47
MatchingNet	58.73 ± 0.18	66.73 ± 0.74
CovaMNet	59.64 ± 0.87	72.15 ± 0.77
RelationNet	60.45 ± 0.73	76.38 ± 0.59
Our model	66.28 ± 0.66	80.06 ± 0.59

• **Experimental results on Stanford Cars**

Stanford Cars are proposed for fine-grained recognition in the fully supervised setting and recently applied to the challenging fine-grained few-shot classification. We applied the Conv-4 as the backbone of the feature extractor, and re-implement the codes of ProtoNet, MatchingNet, RelationNet, MAML, GNN, CovaMNet, and DN4. The confidence intervals are 95%. The experimental results on Stanford Cars are shown in Table 4.

Table 4 shows that our model achieved a great performance, especially on the 5-way 5-shot task. The variances in Stanford Cars are small. Since the model is trained to learn to find more accurate evidence to make a decision, it is more challenging than the generic datasets. Even for the difficult 5-way 1-shot task, the accuracy of our model on Stanford Cars is improved greatly and we achieved 69.98% accuracy.

Table 4
Experimental results on Stanford Cars

Method	5-way 1-shot	5-way 5-shot
ProtoNet	30.39 ± 0.35	58.22 ± 0.41
MatchingNet	36.47 ± 0.25	54.31 ± 0.67
RelationNet	46.19 ± 0.47	57.21 ± 0.38
MAML	46.67 ± 0.52	60.31 ± 0.52
GNN	55.53 ± 0.48	68.98 ± 0.34
CovaMNet	54.36 ± 0.61	71.05 ± 0.68
DN4	59.21 ± 0.83	88.14 ± 0.69
Our model	69.98 ± 0.78	89.84 ± 0.63

• **Ablation study**

The above experimental results show that our model is meaningful, especially for the fine-grained recognition tasks. We perform a set of ablation studies to further investigate the effect of each component in our proposed model. Conv-4 is applied as the backbone of the feature extractor, and the confidence intervals are 95%. The results are summarized in Table 5.

Table 5
Ablation study of our model

datasets	components			5-way 1-shot	5-way 5-shot
	CA	ICM	CM		
Stanford Dogs			√	49.67 ± 0.22	66.23 ± 0.41
		√		49.89 ± 0.45	66.85 ± 0.19
	√		√	53.02 ± 0.63	68.11 ± 0.53
	√	√		56.78 ± 0.89	69.23 ± 0.25
CUB-200			√	59.84 ± 0.34	75.89 ± 0.29
		√		60.03 ± 0.31	75.95 ± 0.28
	√		√	64.62 ± 0.52	79.24 ± 0.49
	√	√		66.28 ± 0.66	80.06 ± 0.59
Stanford Cars			√	62.98 ± 0.33	86.03 ± 0.46
		√		63.11 ± 0.47	86.08 ± 0.39
	√		√	66.67 ± 0.62	88.21 ± 0.58
	√	√		69.88 ± 0.78	89.84 ± 0.63

In Table 5, CA, ICM and CM are the abbreviations of the channel-attention component, improved cosine metric component and cosine metric component, respectively. in ablation study, we applied Conv-4 as the backbone to extract features. When we replaced the cosine similarity metric with the improved cosine similarity metric and directly applied them on feature maps, the experimental results on these three fine-grained datasets show that the improved cosine similarity metric has better performance than the cosine similarity metric although the improvement of accuracy is insignificant. The accuracy of our model is significantly improved when we apply channel attention and improved cosine similarity metric, and we achieved 56.78%, 66.28%, and 69.88% accuracy on Stanford Dogs, CUB-200 and Stanford Cars, respectively, for the difficult 5-way 1-shot task.

- **Visualized features**

We visualized the features by the Grad-Cam method in Fig. 7. The input images are shown in the first line and the visualized features by the Grad-Cam method are shown in the second line.

Fig. 7. The visualized results

As shown in Fig. 7, our method pays more attention to the discriminative features, such as the dog's eyes, ears, and feet.

5. EXPERIMENTAL ANALYSIS

We carried out validation experiments on four data sets for both 5-way 1-shot and 5-way 5-shot tasks. Figure 6 reports a few-shot classification performance on MiniImageNet. We applied ResNet-12 as the backbone and compared the results with those of the state-of-the-art models. Our model achieved comparable performance with MetaDeepBDC. The rest of the experiments are carried out on the more difficult fine-grained datasets with the Conv-4 as the backbone of the feature extractor. Our model achieved better performance compared with the state-of-the-art models on the 5-way 5-shot task and the accuracy is significantly improved on the 5-way 1-shot task. in the fine-grained datasets, the discriminative fine-grained features, such as ears, tails, eyes of dogs, eyes, and feathers of birds, brands and lights of cars are difficult to be extracted. The channel-attention module is applied in our model to effectively extract the discriminative fine-grained features. Figure 5 shows that some of the images in fine-grained datasets are full of noise, occlusion, or light problems, especially in CUB-200. We applied the improved cosine similarity as a metric in our model. The ablation study demonstrates the effectiveness of the channel attention and improved cosine similarity components on Stanford Dogs, CUB-200, and Stanford Cars for both 5-way 1-shot and 5-way 5-shot tasks.

6. CONCLUSIONS

in this work, we proposed a few-shot fine-grained image recognition method. Specifically, we embedded a channel-attention module in the feature extraction net and applied an improved cosine similarity metric to measure the similarity between query images and images in the support set. Since the variances in fine-grained datasets are small and each class contains only a few images, they are more challenging than the generic datasets like MiniImageNet and the recognition model needs to learn the more discriminative features. This paper designed a channel-attention module to emphasize the channels with more discriminative features. To lower the dependence on the light intensity, the corresponding average is subtracted from each pattern in the improved cosine similarity metric. Experimental results demonstrate that our method achieves state-of-the-art performance, especially on three fine-grained benchmark datasets.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *ACL 2018 – 56th Annual Meeting of the Association for Computational Linguistics*, 2018, doi: 10.18653/v1/p18-1031.

[2] S. Kornblith,, J. Shlens, and Q.V. Le. "Do better imagenet models transfer better?" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2661–2671, doi: 10.48550/arXiv.1805.08974.

[3] K. Cao, M. Brbic, and J. Leskovec, "Concept learners for few-shot learning," in *ICLR 2021*, 2021, doi: 10.48550/arXiv.2007.07375.

[4] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, 2015, [online]. [Available]: http://www.cs.toronto.edu/~gkoch.

[5] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural information Processing Systems*, 2017, p. 30, doi: 10.48550/arXiv.1703.05175.

[6] T. Yu *et al.*, "One-shot imitation from observing humans via domain-adaptive meta-learning," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1802.01557.

[7] H.S. Behl *et al.*, "Meta-Learning Deep Visual Words for Fast Video Object Segmentation," in *2020 IEEE/RSJ international Conference on intelligent Robots and Systems (IROS)*, 2020, pp. 8484–8491, doi: 10.48550/arXiv.1812.01397.

[8] K. Hsu, S. Levine, and C. Finn, "Unsupervised learning via meta-learning," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1810.02334.

[9] J. Lu *et al.*, "Learning from very few samples: A survey," *arXiv preprint*, 2020, doi: 10.48550/arXiv.2009.02653.

[10] H. Chen *et al.*, "Sparse spatial transformers for few-shot learning," *arXiv preprint*, 2021, doi: 10.48550/arXiv.2109.12932.

[11] Z. Peng *et al.*, "Few-shot image recognition with knowledge transfer," in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2019, pp.441–449, doi: 10.1109/ICCV.2019.00053.

[12] F. Hao *et al.*, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2019, pp. 8460–8469, doi: 10.1109/ICCV.2019.00855.

[13] F. Wu *et al.*, "Attentive prototype few-shot learning with capsule network-based embedding," in *European Conference on Computer Vision*, 2020, pp. 237–253, doi: 10.1007/978-3-030-58604-1_15.

[14] D. Kang *et al.*, "Relational Embedding for Few-Shot Classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8822–8833, doi: 10.48550/arXiv.2108.09666.

[15] H. Tang *et al.*, "Learning Attention-Guided Pyramidal Features for Few-shot Fine-grained Recognition," *Pattern Recognit.*, vol. 130, p. 108792, 2022, doi: 10.1016/j.patcog.2022.108792.

[16] S. Tian, H. Tang, and L. Dai, "Coupled Patch Similarity Network FOR one-Shot Fine-Grained Image Recognition," in *2021 IEEE international Conference on Image Processing (ICIP)*, 2021, pp. 2478–2482, doi: 10.1109/ICIP42928.2021.9506685.

[17] B. Oreshkin, P. Rodríguez López, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," *Advances in Neural information Processing Systems,* 2018, p. 31, doi: 10.48550/arXiv.1805.10123.

[18] P.C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, 2003, doi: 10.1093/nar/gkg509.

[19] N. Datal, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision And Pattern Recognition (CVPR'05)*, 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[20] V. Devisurya, R. Devi Priya, and N. Anitha, "Early detection of major diseases in turmeric plant using improved deep learning algorithm," *Bull. Pol. Acad. Sci. Tech. Sci.*, vol. 70, no. 2, p. e140689, 2022, doi: 10.24425/bpasts.2022.140689.

[21] Z. Jiang *et al.*, "Few-shot classification via adaptive attention," *Computer Vision and Pattern Recognition*, 2020, doi: 10.48550/arXiv.2008.02465.

[22] D. Wang *et al.*, "Learning a tree-structured channel-wise refinement network for efficient image deraining," in *2021 IEEE international Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428187.

[23] J.S. Lim *et al.*, "Small object detection using context and attention," in *2021 international Conference on Artificial intelligence in information and Communication (ICAIIC)*, 2021, pp. 181–186, doi: 10.48550/arXiv.1912.06319.

[24] X. Wang *et al.*, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141, doi: 10.48550/arXiv.1709.01507.

[26] S. Woo *et al.*, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19, doi: 10.48550/arXiv.1807.06521.

[27] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition*, 2019, pp. 3146–3154, doi: 10.1109/CVPR.2019.00326.

[28] H. Tang *et al.*, "Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning," in *Proceedings of the 28th ACM international Conference on Multimedia*, 2020, pp. 610–618, doi: 10.1145/3394171.3413884.

[29] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," Department of Computer Science and Engineering, Michigan State Universiy, vol. 2, 2006.

[30] O. Vinyals *et al.*, "Matching networks for one shot learning," *Advances in Neural information Processing Systems,* 2016, p. 29, doi: 10.48550/arXiv.1606.04080.

[31] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Differentiable Earth Mover's Distance for Few-Shot Learning," *IEEE Trans. Pattern Anal. Mach. intell.*, 2022, doi: 10.1109/TPAMI.2022.3217373.

[32] H. Huang *et al.*, "Local descriptor-based multi-prototype network for few-shot Learning," *Pattern Recognit.*, vol. 116, p. 107935, 2021, doi: 10.1016/j.patcog.2021.107935.

[33] W. Li *et al.*, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7260–7268, doi: 10.48550/arXiv.1903.12290.

[34] C. Wah *et al.*, "The caltech-ucsd birds-200-2011 dataset," [online]. [Available]: http://www.vision.caltech.edu/visipedia/CUB-200.html.

[35] A. Khosla *et al.*, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011, vol. 2. no. 1.

[36] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proceedings of the IEEE international Conference on Computer Vision Workshops*, 2013, pp. 554–561, doi: 10.1109/ICCVW.2013.77.

[37] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive Subspaces for Few-Shot Learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, 2020, pp. 4135–4144, doi: 10.1109/CVPR42600.2020.00419.

[38] W. Xu *et al.*, "Attentional constellation nets for few-shot learning," in *International Conference on Learning Representations*, 2021, https://par.nsf.gov/servlets/purl/10278170.

[39] Z. Chen *et al.*, "Pareto self-supervised training for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13663–13672, doi: 10.1109/CVPR46437.2021.01345.

[40] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8012–8021, doi: 10.1109/CVPR46437.2021.00792.

[41] Z. Zhou *et al.*, "Binocular mutual learning for improving few-shot classification," in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2021, pp. 8402–8411, doi: 10.1109/ICCV48922.2021.00829.

[42] J. Xie *et al.*, "Joint Distribution Matters: Deep Brownian Distance Covariance for Few-Shot Classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7972–7981, doi: 10.48550/arXiv.2204.04567.