Multi-level Feature Representation and Multi-layered Fusion Contrast for Few-Shot Classification

Wenqian Qin, Wencang Zhao, Ming Li

Abstract—**Training a model that can quickly adapt to new tasks keeps a crucial challenge for few-shot learning. Approaches based on metric-learning are very popular and promising. However, the existing such approaches usually only rely on the feature information obtained in the last layer of the feature extraction backbone for similarity metric, and do not consider that the feature information obtained in multi-layered can be fully utilized. Furthermore, the targeted differentiation of features and the selection and construction of loss functions are usually ignored by these approaches, which will become important factors that limit the performance of the model. Therefore, a few-shot learning approach MFR-MFC with attention mechanism based on multi-level feature representation and multi-layered fusion contrast is proposed in this paper. First, multi-level feature representations are introduced when extracting features, feature information of multi-layers are used to perform information fusion, and then this information is utilized for subsequent metric. Then, when training the model, multi-level features are also used to introduce multi-layered fusion contrast. Additionally, an attention module is introduced in the feature extraction process to make the model obtain more discriminative information. Experiments have shown that the approach proposed in this paper has achieved excellent performance in few-shot classification and has significant advantages compared with advanced technologies.**

Index Terms—**few-shot learning, image classification, multi-level feature, contrastive loss, attention mechanism**

I. Introduction

At present, deep neural networks have been widely and successfully used in various computer vision tasks, and a major factor of these successes is the availability of a large amount of high-quality data as a prerequisite. However, in real life, it is often difficult to obtain a huge amount of data. Humans have the cognitive ability to effectively learn from the data, even if a small amount of data. In order to allow machines to imitate this ability of humans, few-shot learning (FSL) [1][2] has been proposed, and has become a challenging research field with good development prospects.

The current technologies used to solve FSL can be roughly divided into approaches based on metric-learning and based on meta-learning. The former aims at learning the similarity expression and comparison of information, that is, feature extraction and similarity metric. The latter mainly learns a good optimizer or initialization to generalize the model to new tasks that have not been seen before. Among them, approaches based on metric-learning has attracted the attention of many researchers and achieved rapid development due to its simple and effective advantages. However, this type of approaches often lack the ability to highlight the valuable areas of the feature and weaken the irrelevant areas when performing feature extraction, which is a limitation for the acquisition of effective information. This also affect the structure of the embedding space, which will result in the similarity metric performed in the embedding space does not reach the desired effect. The attention mechanism [3] has been effectively used in some areas of deep learning. It helps to distinguish the effectiveness of features and tends to learn more informative and robust representations. Therefore, the attention mechanism is introduced into the model. By embedding an attention module in the feature extraction backbone to obtain attention weights on different channels of the feature map, the features can be more refined to highlight effective information and suppress irrelevant information. The introduction of attention is helpful to utilize these more valuable information in the follow-up similarity metric stage.

Further, FSL approaches based on metric-learning usually only perform a single-path similarity metric. That is, only the sample feature information extracted by the last layer of the feature extraction network is used, and it is directly sent to the metric stage for similarity discrimination and image classification. This idea seems simple and efficient, but when the background of the target image is more complicated or there are too many interference factors, only using a single level of feature may make the information of the sample not ideally represented. This will limit the performance of the model and adversely affect the classification results. Therefore, it is very necessary to express features from multi-level. In this work, multi levels of the feature extraction network is considered and multi-level feature representation is introduced. It can transmit different levels of information by extracting multi levels of sample features. And based on the idea of integration, using the characteristics of each level,

Manuscript received August 06, 2021; revised February 09, 2022. This work was supported by the National Natural Science Foundation of China under Grant (No.61171131) and Key R&D Plan of Shandong Province under Grant (No.YD01033).

Wenqian Qin is a postgraduate student in the College of Automation and Electronic Engineering at Qingdao University of Science and Technology, Qingdao, 266061, China. (e-mail: qinwenqian0131@foxmail.com).

Wencang Zhao is a professor and doctoral supervisor at the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, 266061, China. (e-mail: w.zhao@qust.edu.cn).

Ming Li is a postgraduate student in the College of Automation and Electronic Engineering at Qingdao University of Science and Technology, Qingdao, 266061, China. (e-mail: Leonalm529@163.com).

the similarity relationship on each level is judged, and the corresponding similarity score is obtained. Combining these similarity discrimination scores, the comprehensive similarity discrimination results obtained after multi-level feature representation are further obtained, which makes the similarity discrimination relationship more convincing. At the same time, it also avoids the limitation that each sample has only one set of feature vectors for characterization.

In addition, query-centered classification loss usually be used as the final training loss to train the model by traditional approaches. The importance of the loss function is rarely considered by them, especially in the current rapid development of FSL approaches based on metric-learning. To further improve the classification performance of the model, the selection or formulation of the loss is also worthy of our attention. The idea of contrastive learning [4] is to use contrastive loss by expanding the sample, so that the positive samples of the same class are gathered as much as possible, and the negative samples that do not belong to the same class are scattered as much as possible. Inspired by the idea of contrastive learning, in this work, multi-layered fusion contrastive loss is introduced in the model, data expansion is used to increase intra-class variation, and the distance between samples of the same class is narrowed and the distance between samples of different classes is widened through the contrastive loss. For FSL tasks, combining the contrastive loss based on the support set and the classification loss based on the query set can make full use of the limited support set and query set samples. Meanwhile, a feature extractor with good generalization ability is obtained through training, where the performance of the model on few-shot classification tasks is improved. Moreover, in the FSL task, unlike the traditional metric methods based on the Euclidean domain, a similarity metric method based on Convolutional Neural Network (CNN) is chose for classification.

To summarize, our work and contributions are as follows:

1. In the process of feature extraction, different levels of feature information are obtained by introducing multi-level feature representation. At the same time, this information is fully utilized based on the idea of integration, and a comprehensive and more robust similarity metric result is obtained.

2. The attention mechanism is introduced into the feature extraction backbone, and more distinguishable and valuable feature content is obtained by learning the attention weights of different channels.

3. Based on the idea of contrastive learning, the introduction of multi-layered fusion contrastive loss in the model pre-training is more helpful to improve the model's ability to extract similar features and generalization capabilities.

II. Related Work

A. Metric-learning based approaches

The basic idea of these approaches is to learn the embedding space, and measure the similarity (distance) between the query sample and the support sample in the embedding space to complete the classification. Siamese neural network [5] is a two-way network that measures the similarity of query samples and each support set samples to achieve classification. This approach contains the idea of contrastive learning to some extent, and belongs to the early FSL approach based on metric-learning. In [6], PrototypicalNet was proposed, which is classified by finding the class representation of each type of sample in the feature space, that is, the prototype, and calculating the distance between the query sample and each type of prototype. The idea of RelationNet [7] is to use a neural network to distinguish the degree of similarity between samples and obtain a similarity score for classification. We continue this idea in this work, which can be understood as constructing a non-linear classifier. Intra-class commonality and inter-class uniqueness of features were emphasized in [8], whereby a category traversal module (CTM) was proposed to better extract features. The goal of TADAM [9] is to increase task adaptability, which relies on conditional batch normalization to learn task-related metric space. In [10] and [11], more advanced approaches based on subspace ideas were proposed. Their idea is to classify samples by constructing projection subspaces, not just confined to the original feature space.

B. Meta-learning based approaches

This type of approaches aim to learn a cross-task meta-learner to improve the generalization ability of the model. MAML [12] hopes that the meta-learner has an ideal parameter initialization, so as to better adapt to new tasks and have good performance on new tasks. MM-Net [13] uses an external memory module and uses a Recurrent Neural Network (RNN) to generate weight parameters for the query set encoder. In [14], a differentiable quadratic programming (QP) solver was incorporated in the model. The idea of [15] is to train the model in a high-dimensional parameter space, and only need a few updates in the low-data area to have good performance.

C. The attention mechanism related approaches

The attention mechanism [3] has played a good role in many machine vision tasks. SENet was proposed in [16]. The idea is that the importance of feature maps of different channels is different. A Squeeze-and-Excitation block is used to learn the weights of different feature maps to obtain feature representations with attention weights. The role of the attention mechanism in FSL is usually to help the model better extract useful feature information. MatchingNet [17] combines the attention mechanism with a nearest-neighbor classifier. MADN4 [18] is based on DN4 [19], combining local descriptors and attention mechanism to improve the extracted features and achieve effective performance improvements. In this work, we embed the idea of SENet into the feature extraction network in the form of channel attention module, and by introducing the attention mechanism, the model can focus on important features while suppressing unimportant features.

D. Contrastive learning related approaches

Contrastive learning learns the feature representation of samples by constructing positive and negative samples, comparing between samples in the feature space. Maximize the similarity of samples of the same class, that is, maximize the similarity between different transformed views of the same image, and minimize the similarity of samples of different classes. It has been widely used in self-supervised

tasks [20][21]. For FSL tasks, [22] used the contrastive self-supervised learning to obtain feature extractors with good generalization capabilities. In [23], supervision was incorporated in the instance discrimination based contrastive self-supervised learning framework and formed two approaches, S-SimCLR and S-MoCo, to solve the FSL problem.

III. Proposed Approach

A. Problem formulation

In FSL problem, usually given a training set D_{train}, a validation set D_{val} and a test set D_{test}, the three sets are disjoint. Usually, the support set $D_{support}$ and query set D_{query} are randomly extracted from the training set, and the model is optimized for training by predicting which class the query sample belongs to in the support set. After the training is completed, the model is generalized to new test tasks. We still use the typical N-way K-shot setting when evaluating the model for FSL tasks, that is, in each episode, N classes of samples are randomly selected as the support set, and each class has K samples. We follow the most popular evaluation settings, namely 5-way 1-shot and 5-way 1-shot settings.

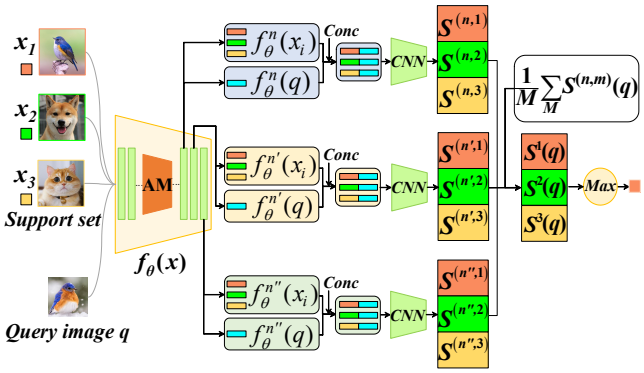


Fig. 1. The process of obtaining the multi-level feature representation.

B. Multi-level feature representation for FSL

We introduced the multi-level feature representation of the sample in this work, and the process of obtaining it is shown in Fig.1. Let x_1, x_2 and x_3 denote the three classes of support set images, and q be the query sample. Input them into the feature extraction network $f_\theta(x)$, the figure shows some convolutional layers and the attention module (AM) embedded in $f_\theta(x)$. Then for each support sample x_i ($i \in \{1,2,3\}$ in the figure) and query sample q, the feature vectors obtained after they pass the nth convolution layer of $f_\theta(x)$ can be expressed as $f_\theta^n(x_i)$ and $f_\theta^n(q)$. Concatenate the representation $f_\theta^n(q)$ of the query sample q with the support image representations $f_\theta^n(x_1)$, $f_\theta^n(x_2)$, and $f_\theta^n(x_3)$ extracted from the corresponding layer, and input them into a CNN $\phi(\cdot)$ for similarity metric to obtain the similarity score between $f_\theta^n(q)$ and the corresponding support sample representation:

$$S^{(n,i)} = \phi(Conc(f_\theta^n(q), f_\theta^n(x_i))) \quad (1)$$

where $Conc$ represents the concatenation between the representations. The above is the 1-shot case. When there is more than one support sample for each class, there are K samples for each class m. At this point, the feature vectors extracted by the selected layer n are averaged to obtain the representation of the class m:

$$X^{(n,m)} = \frac{1}{K}\sum_j f_\theta^n(x_j^m) \quad (2)$$

Then, concatenate the query representation $f_\theta^n(q)$ and the representation of the class m and enter $\phi(\cdot)$ to obtain the similarity score between the query q and the class:

$$S^{(n,m)}(q) = \phi(Conc(f_\theta^n(q), X^{(n,m)})) \quad (3)$$

Equation (3) represents the similarity score between the query q on the specified level n and the class m. Average the scores of each specified layer to obtain the final similarity integration score between q and class m:

$$S^m(q) = \frac{1}{M}\sum_M S^{(n,m)}(q) \quad (4)$$

where M is the total number of selected layers. Equation (4) represents the average value of M scores. The class with the largest integrated similarity score is the classification result of query q.

C. Attention module

In this work, in order to obtain more representative feature information, we followed the idea of [16] and introduced the attention mechanism into our model in the form of the Attention Module (AM). Specifically, the attention module focuses on the weights of different channels. First, the intermediate feature map $X \in R^{C\times H\times W}$ extracted in the previous stage is input into the attention module, and the global average pooling (GAP) is performed on it to obtain the global average pooling vector $X_G \in R^{C\times 1\times 1}$. This step is equivalent to indicating the value distribution of the feature maps of the C channels. Then, through the two-layer fully connected layer and the *sigmoid* function, the attention weight $C_A \in R^{C\times 1\times 1}$ on the channel is obtained. The function of the fully connected layer is to fuse the information of the feature maps of each channel. Finally, the original intermediate feature map X and the channel attention weight C_A are multiplied channel by channel, and the result is the optimized feature map X' we need, that is, $X' = X \otimes C_A$. The whole process of obtaining the optimized features is shown in Fig.2.

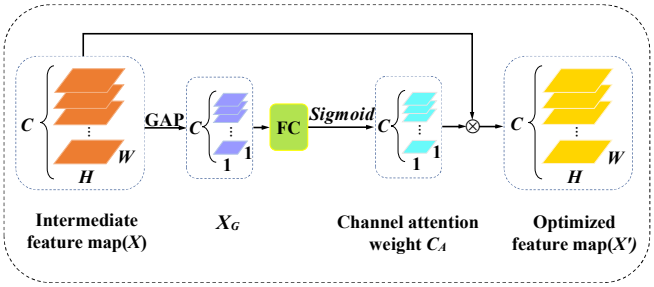


Fig. 2. The structure of the attention module.

D. Multi-layered fusion contrastive loss

The idea of contrastive learning is introduced in the pre-training stage of the model. The similarity comparison is carried out through data expansion and multi-level feature acquisition, and then the contrastive loss is obtained. First, random data augmentation [24] is performed on each input image x_i, and its two expanded samples $\hat{x}_i$ and $\tilde{x}_i$ are obtained. At this time, the number of expanded samples is

twice the original number of samples l, that is, $i \in I$, $I = \{1,...,2l\}$. Then, the two expanded samples are passed through the nth convolutional layer of the feature extraction network $f_\theta(x)$ to obtain features as $f_\theta^n(\hat{x}_i)$ and $f_\theta^n(\tilde{x}_i)$, and then through the projection head ψ_n to obtain features $\psi_n(f_\theta^n(\hat{x}_i))$ and $\psi_n(f_\theta^n(\tilde{x}_i))$. In order to distinguish the expression, in the loss function, the positive sample corresponding to $\hat{x}_i$ is denoted as $\hat{x}_j$. After the sample x_i is expanded, the corresponding contrastive loss obtained after passing through the nth convolutional layer of the feature extraction network is:

$$L^{(n,i)} = -\sum_{i \in I} log \frac{exp(sim(\psi_n(f_\theta^n(\hat{x}_i)), \psi_n(f_\theta^n(\hat{x}_j)))/\tau)}{\sum_{j'} exp(sim(\psi_n(f_\theta^n(\hat{x}_i)), \psi_n(f_\theta^n(\hat{x}_{j'})))/\tau)} \quad (5)$$

where $j' \neq i$, the denominator represents the sum of similarity between $\hat{x}_i$ and the remaining $2l$-1 samples except itself, and τ refers to a temperature parameter. The similarity metric method here uses cosine similarity, namely:

$$sim(a,b) = \frac{a^T b}{\| a \|_2 \| b \|_2} \quad (6)$$

where $\| \|_2$ represents the 2 norm of the vector. For each selected layer that introduced contrastive learning, the corresponding contrastive loss can be calculated. Further, the sum of the contrastive losses obtained after adding contrastive learning in different layers can be obtained, which is what we call the multi-layered fusion contrastive loss:

$$Loss = \sum_{n \in N} \sum_{i \in I} L^{(n,i)} \quad (7)$$

where N is the set of all selected layers, and I is the set of expanded samples. The process of obtaining multi-layered fusion contrastive loss is shown in Fig.3.

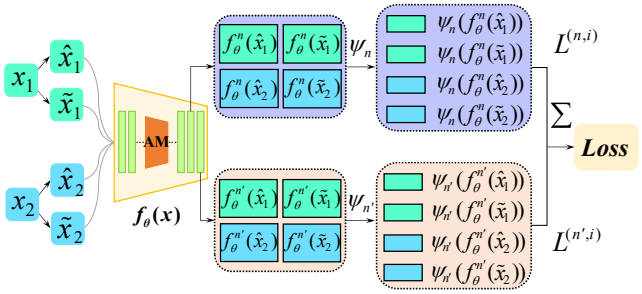


Fig. 3. The process of obtaining the multi-layered fusion contrastive loss.

IV. Experiments

A. Experimental setup

For the calculation of multi-layered fusion contrastive loss, when data augmentation is performed on the support set image, we have chosen four methods: RandomHorizontalFlip, ColorJitter, RandomGrayscale and RandomGaussianBlur. The projection head maps the feature extracted by the feature extraction network to a 128-dimensional vector. When performing FSL similarity metric, the idea of [7] is continued, instead of using artificially prescribed similarity discrimination methods, a CNN is used to measure. It mainly includes 2 convolution layers, each layer has 64 kernels of 3×3 size. Each convolutional layer is followed by a batch normalization layer, a ReLU activation function, and a max-pooling layer. The last pooling layer is followed by the fully connected layer and the *sigmoid* function.

In the selection of the feature extraction backbone, ResNet-12 [25] is used, which is composed of four residual blocks, and the number of channels is 64, 128, 256, and 512, respectively. Each residual block has three convolutional layers and a shortcut connection structure, and the size of the kernel is 3×3. After the convolutional layer is the batch normalization layer and the ReLU activation function, after the last ReLU activation function is the max-pooling layer. The multi-level feature representation consists of the features extracted from the last 5 convolutional layers. The attention module is embedded behind the last convolutional layer in the first two residual blocks, and the contrastive loss is introduced on the last convolutional layer of the last two residual blocks. The specific backbone network architecture is shown in Fig.4.

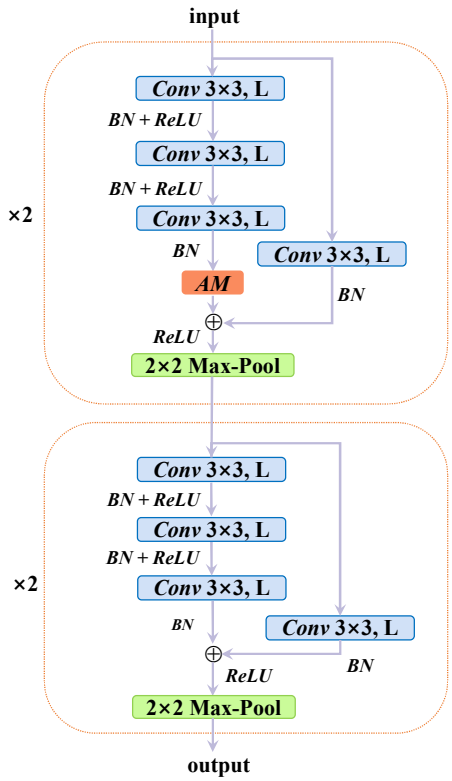


Fig. 4. The backbone network structure.

During pre-training, the loss of the model consists of two parts, namely the classification loss and the multi-layered fusion contrastive loss. The stochastic gradient descent (SGD) [26] is used as the optimizer, the initial learning rate is 10^{-3}. τ is set to 0.1. The batch size is 256. After the pre-training is completed, the samples in the validation set are used to fine-tune the last few layers of the feature extraction network. During the test, 20 query samples are randomly selected for classification in each episode, and a total of 800 episodes are constructed for testing.

B. Datasets

***mini*ImageNet** [17] is a small version of ImageNet [27], with a total of 60000 images divided into 100 classes, each with 600 samples. It is divided into 64 classes for training, 16 classes for verification and 20 classes for testing.

***tiered*ImageNet** [28] also comes from ImageNet, with a total of 779165 images and 608 classes. For this dataset, 351 classes from 20 different categories are used for training, 97

classes from 6 different categories are used for validation, and 160 classes from 8 different categories are used for testing.

C. FSL results on miniImageNet

Table I shows the experimental results of the comparison between our proposed approach and other advanced approaches on *mini*ImageNet. The table lists the experimental data of related approaches when different feature extraction backbones are used. It can be seen that compared with these approaches, our model achieves a better classification accuracy. Compared with the approaches using ResNet-12, the classification accuracy rate on the 5-way 1-shot task achieves the highest advantage of 5.38% (63.88% VS 58.50%), and the 5-way 5-shot task achieves the highest advantage of 3.31% (78.91% VS 75.60%). Compared with TapNet and DSN, we have achieved better classification performance without involving subspace construction, and we have not involved too many complex operations. In addition, a deeper backbone ResNet-18 [25] is used by CTM, while a wider and deeper backbone Wide Residual Network (WRN-28-10) [29] is used by LEO. Compared with them, we do not need such a deep and wide network structure to obtain competitive results. Compared with S-SimCLR, S-MoCo and CSSL-FSL, which involve contrastive learning idea, we have also achieved obvious advantages in performance. Especially compared with CSSL-FSL, this approach involved graph-aggregation, self-distillation and manifold augmentation. Our approach is not required unique work in terms of the above, we can achieve better results.

TABLE I
ACCURACY COMPARISON WITH OTHER APPROACHES ON MINIIMAGENET

Model	Backbone	5-way 1-shot	5-way 5-shot
MatchingNet [17]	Conv-4	43.56 ± 0.84%	55.31 ± 0.73%
PrototypicalNet [6]	Conv-4	49.42 ± 0.78%	68.20 ± 0.66%
RelationNet [7]	Conv-4	50.44 ± 0.82%	65.32 ± 0.70%
MAML [12]	Conv-4	48.70 ± 1.84%	63.11 ± 0.92%
DN4 [19]	Conv-4	51.24 ± 0.74%	71.02 ± 0.64%
TapNet [10]	Conv-4	50.68 ± 0.11%	69.00 ± 0.09%
DSN [11]	Conv-4	51.78 ± 0.96%	68.99 ± 0.69%
MADN4 [18]	Conv-4	53.20 ± 0.52%	71.66 ± 0.47%
PrototypicalNet [6]	ResNet-12	59.25 ± 0.64%	75.60 ± 0.48%
TADAM [9]	ResNet-12	58.50 ± 0.30%	76.70 ± 0.30%
TapNet [10]	ResNet-12	61.65 ± 0.15%	76.36 ± 0.10%
DSN [11]	ResNet-12	62.64 ± 0.66%	78.83 ± 0.45%
MetaOptNet [14]	ResNet-12	62.64 ± 0.61%	78.63 ± 0.46%
CTM [8]	ResNet-18	62.05 ± 0.55%	78.63 ± 0.06%
LEO [15]	WRN-28-10	61.76 ± 0.08%	77.59 ± 0.12%
S-SimCLR [23]	ResNet-18	59.02 ± 0.86%	77.51 ± 0.66%
S-MoCo [23]	ResNet-18	59.94 ± 0.89%	78.17 ± 0.64%
CSSL-FSL [22]	ResNet50-v1	54.17 ± 1.31%	68.91 ± 0.90%
MFR-MFC (ours)	ResNet-12	**63.88** ± **0.36**%	**78.91** ± **0.43**%

D. FSL results on tieredImageNet

Table II shows the experimental results of the comparison between our proposed approach and other approaches on *tiered*ImageNet. Compared with these approaches, our model still achieves a higher accuracy. Compared with the approaches using ResNet-12, our model achieves an advantage of at least 2.75% (68.97% VS 66.22%) in the 5-way 1-shot task and at least 1.72% (84.51% VS 82.79%) in the 5-way 5-shot task. Moreover, our model still performs better than CTM, LEO, S-SimCLR and S-MoCo where more complex backbones are used. It can be seen from the comparison results of the above two datasets that the model we proposed is highly competitive. It is proved that our multi-level feature representation, the introduction of attention mechanism and the idea of multi-layered fusion contrastive loss can make the model achieve outstanding FSL effects.

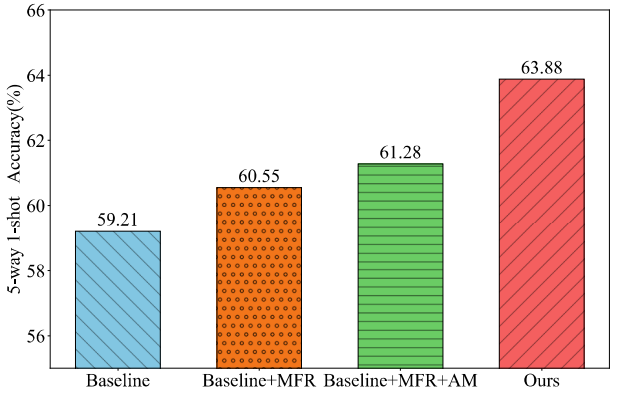
TABLE II
ACCURACY COMPARISON WITH OTHER APPROACHES ON TIEREDIMAGENET

Model	Backbone	5-way 1-shot	5-way 5-shot
MAML [12]	Conv-4	51.67 ± 1.81%	70.30 ± 1.75%
RelationNet [7]	Conv-4	54.48 ± 0.93%	71.32 ± 0.78%
PrototypicalNet [6]	Conv-4	53.31 ± 0.89%	72.69 ± 0.74%
DN4 [19]	Conv-4	53.37 ± 0.86%	74.45 ± 0.70%
TapNet [10]	Conv-4	57.11 ± 0.12%	73.66 ± 0.09%
PrototypicalNet [6]	ResNet-12	61.74 ± 0.77%	80.00 ± 0.55%
TADAM [9]	ResNet-12	62.13 ± 0.31%	81.92 ± 0.30%
TapNet [10]	ResNet-12	63.08 ± 0.15%	80.26 ± 0.12%
DSN [11]	ResNet-12	66.22 ± 0.75%	82.79 ± 0.48%
MetaOptNet [14]	ResNet-12	65.99 ± 0.72%	81.56 ± 0.63%
CTM [8]	ResNet-18	64.78 ± 0.11%	81.05 ± 0.52%
LEO [15]	WRN-28-10	66.33 ± 0.05%	81.44 ± 0.09%
S-SimCLR [23]	ResNet-18	67.60 ± 0.69%	83.74 ± 0.49%
S-MoCo [23]	ResNet-18	68.70 ± 0.72%	84.40 ± 0.64%
MFR-MFC (ours)	ResNet-12	**68.97** ± **0.41**%	**84.51** ± **0.30**%

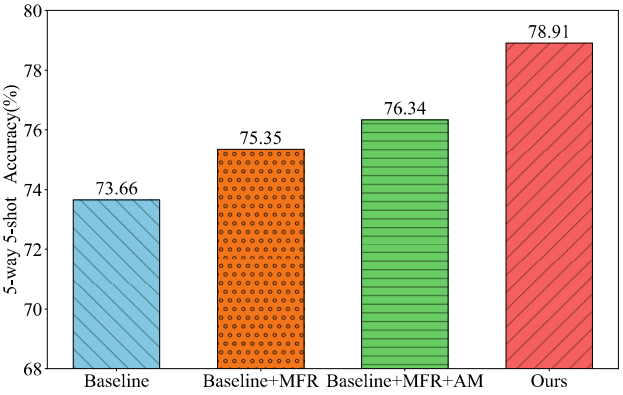
E. Ablation study

In order to study the effectiveness of the ideas of each part of our approach, 5-way 1-shot and 5-way 5-shot FSL ablation experiments is conducted on *mini*ImageNet. First of all, in order to make a better comparison, on the basis of our overall model, without introducing multi-level feature representation, attention module and multi-layered fusion contrastive loss, the model at this time is used as the baseline approach of this section. Secondly, on the basis of this, only multi-level feature representation is introduced, as shown in Fig.5. At this time, the accuracy of the model in the 5-way 1-shot and 5-way 5-shot tasks are increased by 1.34% and 1.69%, respectively. Then, the attention module (AM) is embedded in the feature extraction network, and the accuracy of 5-way 1-shot and 5-way 5-shot tasks are increased by 0.73% and 0.99%, respectively. Finally, on the basis of the previous, the multi-layered fusion contrastive loss is introduced. The classification accuracy of the model on 5-way 1-shot and 5-way 5-shot tasks are increased by 2.6% and 2.57%, respectively. Meanwhile, the model classification accuracy reaches the highest, that is, the whole of the proposed approach MFR-MFC. Through the above ablation studies, the validity of the ideas in this work is further proved. That is, multi-level feature representation, attention module and multi-layered fusion contrastive loss all contribute to the

performance improvement of the model in FSL classification tasks, and can achieve excellent performance.



(a) 5-way 1-shot

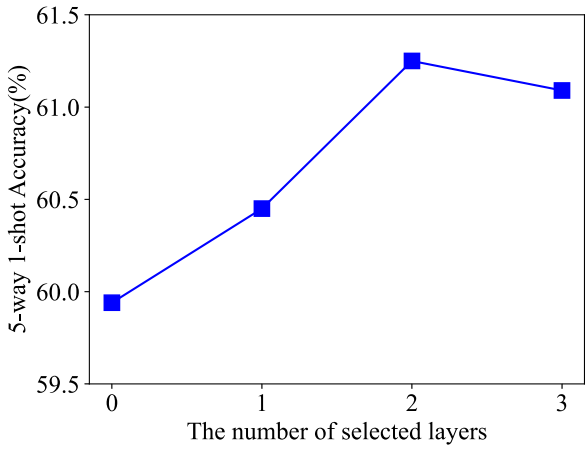


(b) 5-way 5-shot

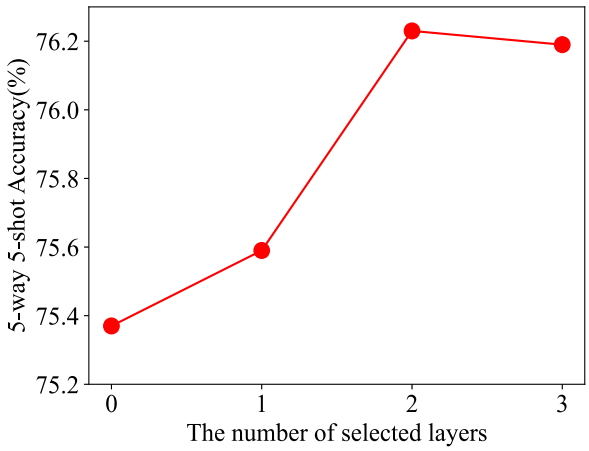
Fig. 5. Ablation study on the *mini*ImageNet dataset.

F. Research on selected layers for contrastive loss

In order to further explore the performance of multi-layered fusion contrastive loss when the selected layers are different, convolutional layers is specified in the backbone network, which are the 9th, 10th and 12th layers of ResNet-12, so as to study the model performance under different settings. For reducing interference factors, the experiments in this section are all carried out without introducing multi-level feature representation and attention module. In this section, research through the 5-way 1-shot and 5-shot FSL image classification tasks are conducted on *mini*ImageNet. As shown in Fig.6 (a), the horizontal axis represents the number of specified layers where the contrastive loss is introduced. From left to right, it can be seen that when no contrastive loss is introduced, the classification accuracy of the model is the lowest, 59.94%. When only the contrastive loss is introduced in the 12th convolutional layer of the backbone, the accuracy is 60.45%, an increase of 0.51%. When the loss is introduced in the 9th and 12th layers, the accuracy is 61.25%, which is 0.8% higher than when it is only introduced in the 12th layer. When the loss is introduced in the 9th, 10th and 12th layers, the accuracy rate is 61.09%, which is slightly lower than the accuracy when the two layers are selected. A similar rule is also shown in Fig.6 (b). The above experiments prove that the idea of introducing contrastive loss in the model is effective, and it also proves that the effect of multi-layered fusion contrastive loss is better than that of introducing contrastive loss on a single layer. In addition, the accuracy is slightly lower when the three-layer contrastive loss is introduced than when the two-layer loss is introduced, but the gap is small, and the effect is still better than the loss without introducing or only introducing a single layer.



(a) 5-way 1-shot



(b) 5-way 5-shot

Fig. 6. Research on selected layers for contrastive loss.

V. Conclusion

In this paper, a novel few-shot learning approach based on metric-learning is proposed. Through the introduction of multi-level feature representation, the information at different levels of the sample is extracted, and information interaction is realized by combining the similarity at each level. By introducing multi-layered fusion contrastive loss, the model can learn a better class representation, and at the same time help it have a stronger generalization ability. Additionally, through the introduction of the attention mechanism, the information interaction between the feature map channels is realized, which helps the model to pay attention to more valuable feature information while suppressing the interference of irrelevant information when performing feature extraction. A series of experiments on the *mini*ImageNet and *tiered*ImageNet datasets have proved the excellent performance of our proposed approach, especially when compared with the existing advanced technologies, it is also very competitive.

References

[1] M. Fink, "Object Classification from A Single Example Utilizing Class Relevance Metrics," *Advances in Neural Information Processing Systems*, no. 17, pp. 449-456, 2005.

[2] F. F. Li, R. Fergus and P. Perona, "One-Shot Learning of Object Categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594-611, 2006.
[3] V. Mnih, N. Heess and A. Graves, "Recurrent Models of Visual Attention," *Advances in Neural Information Processing Systems*, pp. 2204-2212, 2014.
[4] R. Hadsell, S. Chopra and Y. LeCun, "Dimensionality Reduction by Learning An Invariant Mapping," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1735-1742.
[5] G. Koch, R. Zemel and R. Salakhutdinov, "Siamese Neural Networks for One-Shot Image Recognition," *ICML Deep Learning Workshop*, Vol. 2, 2015.
[6] J. Snell, K. Swersky and R. Zemel, "Prototypical Networks for Few-Shot Learning," *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4080-4090.
[7] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr and T. M. Hospedales, "Learning to Compare: Relation Network for Few-Shot Learning," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199-1208.
[8] H. Li, D. Eigen, S. Dodge, M. Zeiler and X. Wang, "Finding Task-Relevant Features For Few-Shot Learning by Category Traversal," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1-10.
[9] B. N. Oreshkin, P. Rodriguez and A. Lacoste, "TADAM: Task Dependent Adaptive Metric for Improved Few-Shot Learning," *In Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 719-729.
[10] S. W. Yoon, J. Seo and J. Moon, "Tapnet: Neural Network Augmented with Task-Adaptive Projection for Few-Shot Learning," *In International Conference on Machine Learning*, 2019, pp. 7115-7123.
[11] C. Simon, P. Koniusz, R. Nock and M. Harandi, "Adaptive Subspaces for Few-Shot Learning," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4136-4145.
[12] C. Finn, P. Abbeel and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *In International Conference on Machine Learning*, 2017, pp. 1126-1135.
[13] Q. Cai, Y. Pan, T. Yao, C. Yan and T. Mei, "Memory Matching Networks for One-Shot Image Recognition," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4080-4088.
[14] K. Lee, S. Maji, A. Ravichandran and S. Soatto, "Meta-Learning with Differentiable Convex Optimization," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10657-10665.
[15] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero and R. Hadsell, "Meta-Learning with Latent Embedding Optimization," *In International Conference on Learning Representations*, 2019.
[16] 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.
[17] O. Vinyals, C. Blundell, T. Lillicrap and D. Wierstra, "Matching Networks for One Shot Learning," *Advances in Neural Information Processing Systems*, no. 29, pp. 3630-3638, 2016.
[18] H. Li, L. Yang and F. Gao, "More Attentional Local Descriptors for Few-Shot Learning," *In International Conference on Artificial Neural Networks*, 2020, pp. 419-430.
[19] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7260-7268.
[20] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729-9738.
[21] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *In International Conference on Machine Learning*, 2020, pp. 1597-1607.
[22] J. Li and G. Liu, "Few-Shot Image Classification via Contrastive Self-Supervised Learning," *arXiv preprint arXiv:2008.09942*, 2020.
[23] O. Majumder, A. Ravichandran, S. Maji, M. Polito, R. Bhotika and S. Soatto, "Revisiting Contrastive Learning for Few-Shot Classification," *arXiv preprint arXiv:2101.11058*, 2021.
[24] X. Cui, V. Goel and B. Kingsbury, "Data Augmentation for Deep Neural Network Acoustic Modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.23, no.9, pp. 1469-1477, 2015.
[25] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.
[26] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 6, pp. 541-551, 1989.
[27] O. Russakovsky, J. Deng, H. Su, et al, "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.
[28] M. Ren, E. Triantafillou, S. Ravi, et al, "Meta-Learning for Semi-Supervised Few-Shot Classification," *In International Conference on Learning Representations*, 2018.
[29] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," *In British Machine Vision Conference*, 2016.
[30] Longlei Cui, and Ying Tian, "Facial Expression Recognition by Regional Attention and Multi-task Learning," *Engineering Letters*, vol. 29, no.3, pp919-925, 2021.
[31] Shumpei Takezaki, and Kazuya Kishida, "Construction of CNNs for Abnormal Heart Sound Detection using Data Augmentation," Lecture Notes in Engineering and Computer Science: *Proceedings of The International MultiConference of Engineers and Computer Scientists 2021*, 20-22 October, 2021, Hong Kong, pp18-23.

Wenqian Qin received his B.Eng. degree in Automation from the College of Engineering of Qufu Normal University, Rizhao, Shandong Province, China, in 2019. He is currently pursuing a master's degree in Control Engineering at the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, Shandong Province, China. His main research directions are: intelligent cognition and understanding, pattern recognition and machine vision, few-shot learning. He has won four first-class scholarships and one second-class scholarship.

Wencang Zhao received the B.Eng. degree in Automation from Qingdao University of Science and Technology, Qingdao, China, in 1995, the M.Eng. degree in Signal and Information Processing from the Shandong University, Jinan, China, in 2002, and the Ph.D. degree in Physical Ocean Science from Ocean University of China, Qingdao, China, in 2005.
He was a Visiting Scholar with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA, in 2016 and 2017. He has been a faculty member with the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, China, since 2005. His current research interests include pattern recognition, image processing, and machine learning.

Ming Li is currently pursuing a master's degree in Control Engineering at the College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, Shandong Province, China. Her main research directions are: intelligent cognition and understanding, pattern recognition and machine vision. She has won first-class scholarship, the second-class national scholarship, and the first-class scholarship of Shandong Province many times.