IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 May 2023, accepted 10 June 2023, date of publication 13 June 2023, date of current version 19 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285721

RESEARCH ARTICLE

Three-Dimension Attention Mechanism and Self-Supervised Pretext Task for Augmenting Few-Shot Learning

YONG LIANG, ZETAO CHEN, (Member, IEEE), DAOQIAN LIN, JUNWEN TAN, ZHENHAO YANG, JIE LI, AND XINHAI LI

College of Mechanical and Control Engineering, Guilin University of Technology, Guilin 541006, China

Corresponding author: Zetao Chen (chenzetao2021@163.com)

This work was supported in part by the Science and Technology Program of Guangxi, China, under Grant 2018AD19184; and in part by the Project of the Guilin University of Technology under Grant GLUTQD2017003.

ABSTRACT The main challenge of few-shot learning lies in the limited labeled sample of data. In addition, since image-level labels are usually not accurate in describing the features of images, it leads to difficulty for the model to have good generalization ability and robustness. This problem has not been well solved yet, and existing metric-based methods still have room for improvement. To address this issue, we propose a few-shot learning method based on a three-dimension attention mechanism and self-supervised learning. The attention module is used to extract more representative features by focusing on more semantically informative features through spatial and channel attention. Self-supervised learning mainly adopts a proxy task of rotation transformation, which increases semantic information without requiring additional manual labeling, and uses this information for training in combination with supervised learning loss function to improve model robustness. We have conducted extensive experiments on four popular few-shot datasets and achieved state-of-the-art performance in both 5-shot and 1-shot scenarios. Experiment results show that our work provides a novel and remarkable approach to few-shot learning.

INDEX TERMS Few-shot, self-supervised pretext task learning, deep learning, image classification, attention mechanism.

I. INTRODUCTION

Image classification is a well-known task in the field of computer vision. With the aid of large datasets of images, deep learning has achieved remarkable results. However, in scenarios where data is limited, training a deep neural network using supervised learning methods requires significant manual labeling effort. This may not be feasible in practical situations, such as medical imaging, where obtaining labeled data may require the expertise of professionals. When data is scarce or labels are unavailable, deep networks are prone to overfitting. While data augmentation and regularization techniques can mitigate this issue, it has not been fully resolved. Therefore, few-shot learning for small datasets has become a crucial technology for addressing this challenge.

In recent years, numerous researchers have put forward diverse research methods. The primary method that is currently in use is meta-learning, also known as learning to learn. The concept behind meta-learning is to learn how to learn, with the hope that the trained model will be able to generalize to new categories. Meta-learning comprises three learning paradigms: metric-based [1], [2], [3], [4], [5], [6], [7], model-based [8], [9], and optimization-based [10], [11]. The model-based meta-learning paradigm acquires experiential knowledge by building a meta-model, which is [12], [13], [14], [15], [16] then utilized to evaluate the final classification task. On the other hand, optimization-based meta-learning ensures that the network learns an excellent initialization to make it easier for the model to be fine-tuned for new tasks.

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

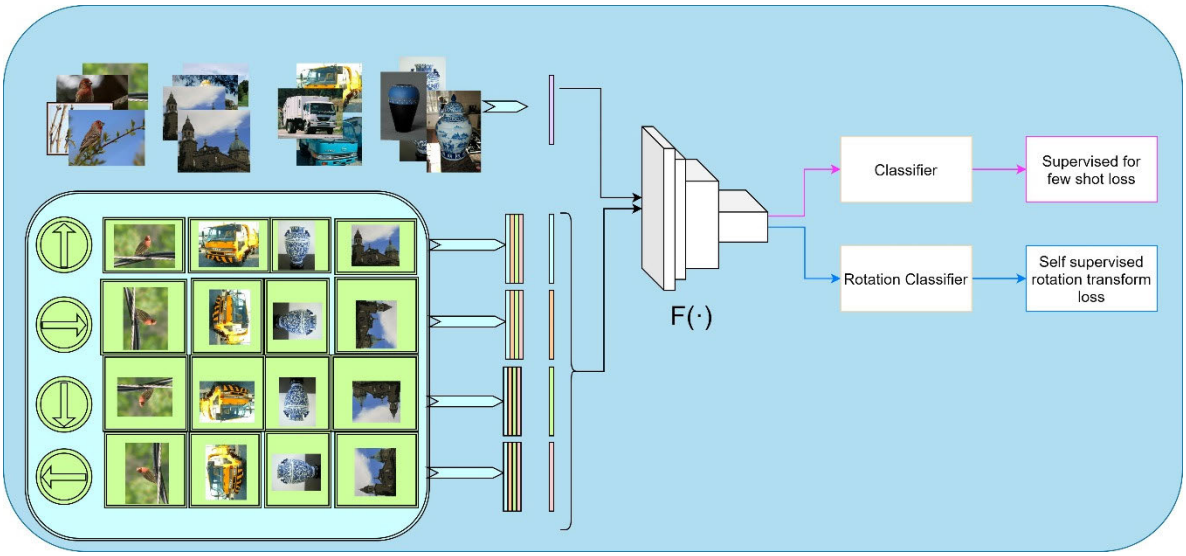

FIGURE 1. Overall framework diagram. The proposed method extracts feature from a featured network with a fused three-dimension attention mechanism. It combines self-supervised learning with rotation transformation based on few-shot learning to enrich the semantic information of features (which will be detailed in Chapter III).

The distance metric-based meta-learning, which is also the mainstream method in few-shot learning, maps the image to a metric space and employs a metric to compute the Similarity between different image samples to accomplish classification.

The above few-shot learning methods can achieve good performance. However, the classification of few-shot learning is not seen during training. Therefore, the model's generalization ability obtained by training is very high. To solve this problem, the existing Methods still have room for improvement in learning feature generalization ability and feature versatility. To obtain more general features or features with better generalization ability, it is first necessary to ensure that the features learned by the model are the most representative. The attention mechanism is used to help the feature extraction model give extra attention to different parts of the picture, increase the weight of practical features, and pay less attention to irrelevant information. Human vision quickly scans the global image to obtain the target area that needs attention and suppress other useless information. Therefore, this paper proposes integrating the attention mechanism into the feature extraction network to help the model always focus on the part of interest in image processing. Therefore, in establishing the image feature extraction model, the attention mechanism further extracts image information into practical information and the importance of learning different local features. Inspired by the SimAM proposed in [17], this paper proposes a feature extraction network that fuses three dimension attention mechanisms. However, the attention mechanism can help the model learn more representative features in few-shot image classification. Due to the small number of samples and labels in each category, it is necessary to provide the model with richer feature information to help it learn features with good generalization ability. This paper proposes to enrich the feature semantic information of the image by doing self-supervised learning on the input image based on few-shot learning, thereby improving the generalization ability of the feature. In recent years, the work of self-supervised learning by geometrically transforming the input image has [18], [19] The work of this paper is to add self-supervised learning of rotation transformation in few-shot learning to obtain richer feature information and adopt simple feature processing during training to help the model learn more general and generalization ability Characteristics.

Specifically, this paper proposes a few-shot image classification model that combines three-dimension attention and rotation transformation self-supervision, as shown in Figure 1. The feature information of the sample is enriched by adding three-dimension of attention to the feature extraction network and self-supervising the rotation transformation of the input sample. The contributions of this paper are as follows:

- Inspired by the human brain's attention, a feature extraction network based on three-dimension attention is proposed to help the network extract more representative features and optimize the existing feature extraction network.
- To address the problem of insufficient feature information and weak feature generalization learned by the model, this paper proposes to combine supervised learning with self-supervised learning to provide more general feature information.
- In this paper, a large number of experiments were carried out on four popular few-shot classification benchmark datasets mini-ImageNet, CIFAR -FS, FC100, and CUB-FS. And ablation studies were carried out, the

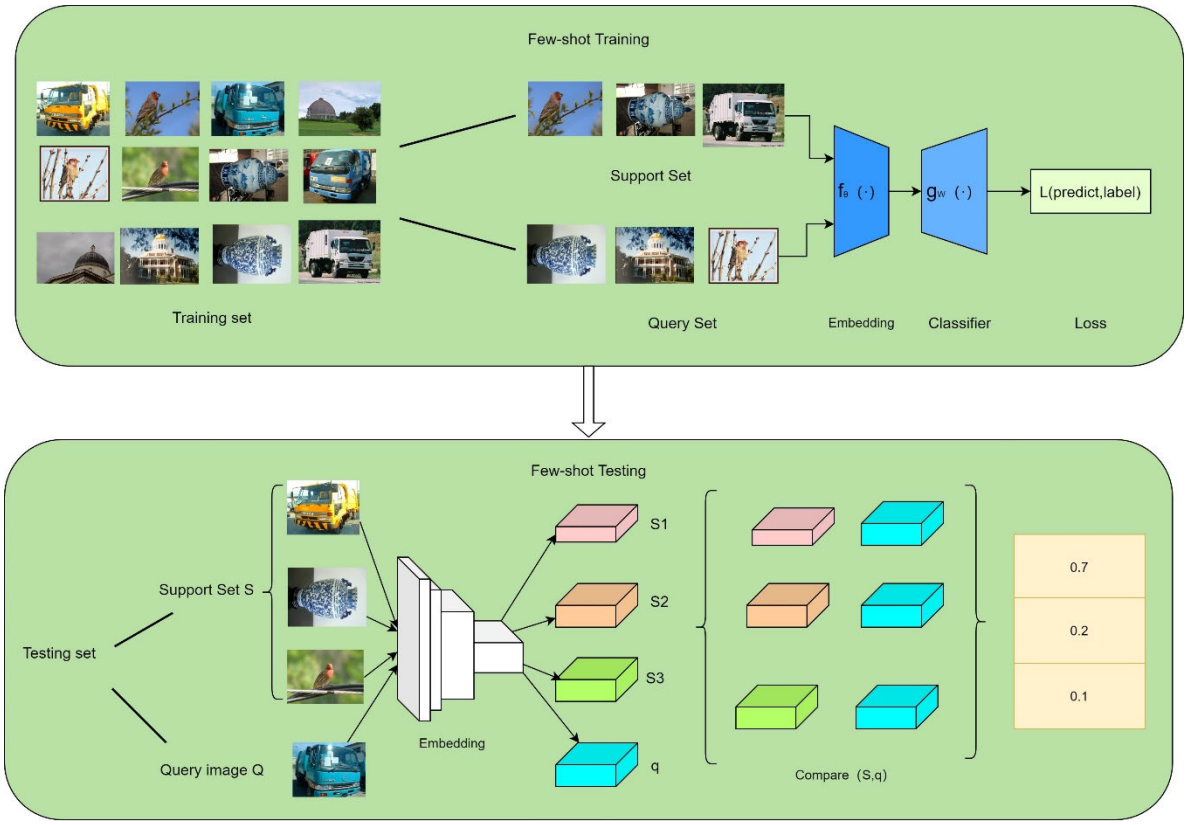

FIGURE 2. The support and query sets come from the training set during the training process. The backbone network predicts the query set by learning the features of the support set and compares the prediction result with the actual value to obtain a loss function, thereby optimizing the network. The support and query sets for the testing process come from the test set, where the categories do not overlap with those in the training set. The feature extraction network represents the support and query sets with a one-dimensional vector during the test. Then it compares them through the classifier to obtain the classification of the query set image.

experimental results showed that the algorithm in this paper reflected the robustness and generalization ability of the algorithm on both the first and fifth tasks. We believe that we can lead more researchers to use self-supervised learning for few-shot learning.

II. RELATED WORK

A. FEW-SHOT IMAGE CLASSIFICATION

The main objective of few-shot image classification is to solve the image classification problem when the sample size is limited or the labels are difficult to obtain. Its work usually learns useful feature representations from the seen training category images and then uses them to classify unseen category images. The few-shot image classification paradigm is shown in Figure 2. This visual task has prompted a lot of classic works to be proposed [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], the most similar to this work is metric-based methods, which complete the classification by measuring the distance between the support set and query set of the samples, such as [2] proposed a general network framework, the fundamental idea of the matching network is to map the image into an embedding space, which also encapsulates the label distribution, then use different architectures to project the test image into the same embedding space, and then use the cosine similarity to measure the similarity to achieve classification; [1] the difference between the prototype network and the matching network is the distance method, and a prototype representation is created for each classification, and the Euclidean distance between the prototype vector of the classification and the query point is used to determine; [20] use graph convolutional neural networks instead of simple convolutional neural networks to extract features; [21] propose that on the basis of metric learning, the method of adding fine-tuning when classifying can improve the classification effect. A simple and effective based on method, this work is based on metric-based methods. Still, this work pays more attention to improving the generalization ability of features, so it introduces attention mechanism into the feature extraction network, and fuses rotation transformation unsupervised learning in few-shot learning, which will be introduced in the following summary.

B. ATTENTION MECHANISM

The application of the attention mechanism in machine vision is mainly to imitate the unique visual mechanism of human beings, that is, to pay more attention to the parts of interest

and suppress the irrelevant parts. Therefore, there are many attention mechanisms used in the vision domain. For example, [17], [22], [23], [23] proposed weighted distribution of channel information in the convolutional network to solve the loss problem caused by the different importance of different channels of the feature map during the convolution pooling process and [22] added attention to spatial information based on [23] and added a pooling layer. Pooling is used in convolutional neural networks to extract high-level features. Therefore, different pooling means that the extracted High-level features are richer. The two methods generate 1D or 2D weights from the features, then expand the generated weights for channel and spatial attention. The SimAM module proposed by [17] directly estimates the 3-dimensional weights. This work has confirmed that it is better than [19], [20] in the convolutional network and does not bring additional calculations. Therefore, inspired by the attention mechanism in the human brain proposed in [24], this paper integrates a three-dimension based on SimAM attention module into the feature extraction network of few-shot learning to help the model extract features with more generalization ability.

C. SELF-SUPERVISED LEARNING

Word2Vec [25] has popularized the self-supervised learning approach and applied these methods to many problems, resulting in rapid development in self-supervised learning. Specifically, supervised learning requires sufficient labeled data when applied in the field of vision. Manual data labeling (images or texts) is required to obtain such information, which is time-consuming and expensive. Unsupervised learning autoencoders [26] only reduce the dimensions without containing more semantic features. That is not very helpful for few-shot image classification for the downstream classification tasks. On the contrary, self-supervised learning has attracted the attention of many researchers and proposed many representative works [27], [28], [29] due to its feature of not relying on any label values and finding the relationships between samples by mining the intrinsic features of the data. Reference [27] proposed self-supervised learning using image grayscale as the input data and corresponding color images as the label data to train the network. To solve the task of distinguishing colors, the model must understand the different objects and related parts appearing in the image to draw these parts with the same color, thus providing help for the downstream tasks. Reference [28] proposed setting proxy tasks in self-supervised learning to solve the segmentation problem. The model must learn how the segments are assembled in an object, the relative positions of different parts, and the object's shape. Therefore, these representations are helpful for downstream classification and detection tasks, but the drawback of this method is the extensive computation. Reference [29] proposed training convolutional networks to recognize the two-dimensional rotation of the input image to learn image features, which proves to be a robust supervision signal in terms of quality and quantity. The labels generated are data obtained after manually rotating the images. Because [29] has a minor computation than [27], [28] and can provide richer supervision signals, this paper adopts self-supervised learning combined with rotation transformation for few-shot learning to provide more semantic features.

D. FEATURE PROCESSING

In deep learning, the normalized feature has a mean of 0 and a standard deviation of 1 on all samples. Standardization of input data makes each feature's distribution similar, often making it easier to train an effective model [30]. Due to the small dataset of few-shot image classification, the features of the pictures need to be further emphasized in the training process to learn more generalizable features in limited pictures. In recent years, many works have been proposed, such as [31] and [32], which propose to use self-correlation and cross-correlation to emphasize the features of the picture samples, and [32] cross-self-attention to obtain more distinctive features, both of which introduce additional calculations. Reference [32] proposes to indicate which CNN uses critical areas in the image to recognize the features of this class. In the few-shot field, [33] combines self-correlation and cross-correlation to emphasize the features of the picture samples. And [3] work uses a simple feature transformation of mean subtraction and normalization to perform better than [34]. This paper adds the feature processing of mean subtraction and normalization in the training process to improve the model's performance.

III. PROPOSED METHOD

In this section, the paper first describes the definition of the few-shot classification problem, technical terms, and related symbol definitions. Then the proposed feature extraction network fused with three-dimension is introduced, and the method of combining few-shot learning with self-supervised learning of rotation transformation is introduced. Finally, we propose a few-shot learning model based on three-dimension attention and self-supervised learning. The overall framework of the method proposed in this paper is shown in Figure 1.

A. PROBLEM DEFINE

In few-shot learning, the original data set is divided into three subsets by different categories: the training set, verification set, and test set, denote as D. Among them, the training set is used in the training process. In contrast, the test set is used in the testing phase. Each subset is divided into a support set S and a query set, denote as Q when used. During the training process, the support set of the training set is used to train the model, and the query set is used to optimize the model. The validation and test sets are used for the validation and testing phases. During the test, the data set will also be divided into set S and set Q, where S is used to provide label information, and Q is the sample to be classified. When the number of categories in S is 5, and the number of samples k for each

category is 1, called the 5way-1shot paradigm for few-shot learning. When k is 5, it is called the 5way-5shot few-shot learning paradigm.

B. FEATURE EXTRACTION NETWORK

This paper selects the most commonly used Resnet12 [35] in the current few-shot learning as a feature extraction network. Inspired by the SimAM attention [16], this paper integrates the three-dimension attention module into Resnet12. Specifically, this paper uses Resnet12 The attention module is integrated into the four basic modules. Existing attention modules in computer vision focus on the channel or spatial domains. These two attention mechanisms correspond to the human brain's feature and spatial-based attention mechanisms [24]. However, these two mechanisms coexist in humans and facilitate information selection during visual processing. Therefore, this paper integrates three-dimension into the feature extraction network. In three-dimension attention, each neuron is assigned a unique weight, each of which corresponds to an energy equation:

$$e_t(w_t, b_t, y, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2. \quad (1)$$

$$\hat{t} = w_t t + b_t \quad (2)$$

$$\hat{x}_i = w_t x_i + b_t \quad (3)$$

$$M = H \times W \quad (4)$$

where w_t and b_t are the weights and biases of the neuron, t and x_i is the target and the other neurons in the single channel of the current input feature, respectively, $\hat{x}_i, \hat{t}$ are the linear transformations of t and x_i respectively. M The number of neurons in the channel comes from the second and third dimensions of input. Equation 1 shows that y_t it is equal to $\hat{t}$ and y_o is equal to $\hat{x}_i$ when the minimum of e_t is obtained. Since the calculation of minimizing e_t is complex, in [35], 1 is analytically solved as:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \quad (5)$$

$$\hat{\mu} = \frac{1}{M}\sum\nolimits_{i=1}^{M} x_i \quad (6)$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum\nolimits_{i=1}^{M}(x_i - \hat{\mu})^2 \quad (7)$$

$$X = sigmoid(\frac{1}{E}) \odot X \quad (8)$$

Among them, cross-channel and spatial dimensions are combined, and operations are added to restrict overly large values. The feature extraction network structure fused in Resnet12 is shown in Figure 3.

C. LOSS FUNCTION

Self-supervised learning can define proxy tasks, use image transformation methods to generate pseudo-labels, and use these pseudo-labels to supervise the network to complete these proxy tasks. A significant advantage of self-supervision is that it does not require additional manual labels and only constructs proxy tasks from existing data to enrich label semantic information. This feature is essential in few-shot image classification scenarios. After literature research [26], [27], [28], it is found that the effect of rotation transformation is better in the current self-supervised agent tasks. Therefore, this paper uses self-supervised rotation transformation to enhance the semantic information of pictures. Specifically, in self-supervised tasks, the input image will be rotated by different angles, and the auxiliary purpose of the model is to predict the amount of rotation applied to the image. Inspired by [28], this paper proposes that the rotation angle is [0°, 90°, 180°, 270°]. In the image classification setting, an auxiliary loss (based on the predicted rotation angle) is added to the standard classification loss function to learn a general representation suitable for image understanding tasks. This paper adds self-supervised learning of rotation transformation to few-shot learning. It forms the total loss function, the self-supervised, and the original loss function. In this way, this paper establishes a fusion of a self-supervised learning model for few-shot image classification. The overall loss function is:

$$L_{total} = 0.5 * Loss + 0.5 * Loss_{rotation} \quad (9)$$

Among them $Loss$ is obtained by the fundamental few-shot backbone network prediction value, and the actual value of the query set $Loss_{rotation}$ is obtained by self-supervised rotation transformation.

D. FEATURE NORMALIZATION

To further optimize the model performance, the feature processing adopted in the training stage of this work is mean subtraction and normalization. The specific approach taken in this paper is first to calculate the mean of the data in each dimension (using the entire data set), then subtract the mean from each dimension. The next step is to divide the data in each dimension by the standard deviation of that dimension. Numerically, this operation gives each feature a mean of 0 and a standard deviation of 1. And observationally, this method has proven effective in [1] improving the model performance. The mean subtraction and normalization formula is:

$$R \leftarrow \frac{R - \bar{R}}{\|R\|_2} \quad (10)$$

where R represents the feature, $\bar{R}$ represents the average value of R, and $\|\bullet\|_2$ represents the standard deviation calculation.

IV. RESULT AND DISCUSSION

The three-dimension-based Resnet12 network of this work uses the backbone network, and the classifier is the nearest neighbor classifier [36]. This chapter will introduce the data set used in the experiment, the experimental settings, and the results. We also did a lot of ablation experiments to prove the effectiveness of the method proposed in this work.

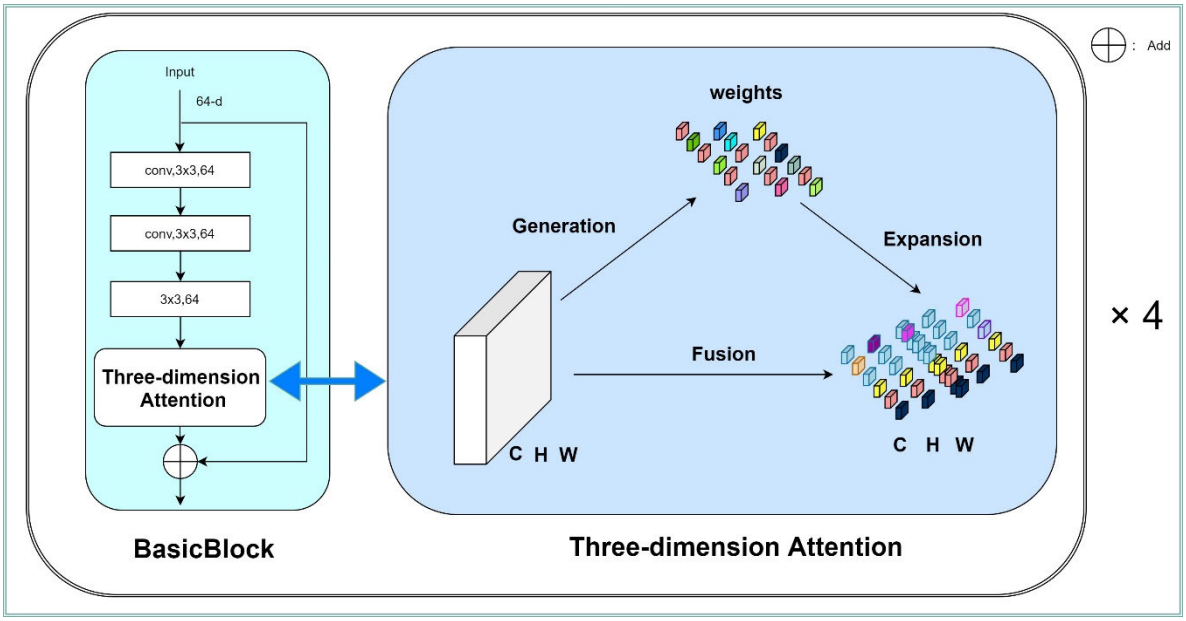

FIGURE 3. Each basic module in this paper's four-layer network of Resnet12 incorporates the three-dimension attention module.

A. DATASET

Mini-Imagenet: A subset of ImageNet, often used to study few-shot learning, contains 100 classes with 600 examples per class. The dataset is split to have 64 base classes, 16 validation classes, and 20 new classes. This paper rescaling and center cropping resizes the image to 84 × 84 pixels.

CIFAR-FS: The full name of the CIFAR-FS dataset is the CIFAR100 Few-Shots dataset, which comes from the CIFAR 100 dataset, which contains 100 categories, 600 images for each category, and a total of 60,000 images. In use, it is usually divided into a training set (64 types), a verification set (16 types), and a test set (20 types), and the image size is unified to 32 pixels.

Fc-100: The full name is the Few-shot CIFAR100 dataset, similar to the CIFAR-FS dataset above. It also comes from the CIFAR100 dataset. It contains 100 categories, 600 images for each category, and 60,000 images. But the difference is that FC100 is not divided according to the category but according to the superclass (Superclass). It contains a total of 20 super-classes (60 categories), including 12 super-classes in the training set, four super-classes (20 categories) in the verification set, and four super-classes (20 categories) in the test set.

CUB200: This dataset is particularly fine-grained and challenging because it only consists of pictures of birds. There are 100 base classes, 50 validation classes, and 50 new classes. The image size is unified to 50 pixels.

B. EXPERIMENT SETTINGS

This experiment is implemented based on the deep learning tool named Pytorch. In this experiment, the cosine annealing method [1] is adopted for training, in which the learning rate varies between the initial given value and 0 during the cosine cycle. This paper sets the cycle to 60, the initial learning rate to 0.1, and reduces the learning rate by 10% in each cycle. This paper sets the total rounds to 240. At the test time, this paper compares the accuracy of this paper's method and other few-shot learning algorithms by dividing 10,000 k-shot C-way tasks from the new class. Each task has C new classes, K labeled samples (support set) images, and 15 test samples (query set) per class. This paper averages the test accuracy of all test images and all tasks and reports the average accuracy and 95% confidence interval. In the training stage, this paper adopts feature de-mean and normalization.

C. EXPERIMENTAL RESULTS

Tables 1, 2, 3, and 4 are the results of the method in this paper on the dataset Mini-ImageNet, CIFAR-FS, FC-100, and CUB200, respectively.

It can be seen from Table 1 to Table 4 that the method in this paper performs well on commonly used few-shot datasets, especially the best performance on the data set CUB200. And CUB200 is a fine-grained picture, so it can be seen that the method in this paper is not only suitable for normal image sets but also fine-grained images. All datasets have results, showing that the method proposed in this paper can help the model learn more representative features, so it can still perform better than other algorithms in fine-grained pictures.

From Table 1 to Table 4, we can see that our model achieves good accuracy in both 5-way 1-shot and 5-way 5-shot scenarios. Compared with other algorithms, our model adds attention mechanism and self-supervised learning proxy

TABLE 1. Average accuracy (%) for 5way 1shot and 5way 5shot on mini-Imagenet.

Method	5-way 1-shot	5-way 5-shot
SimpleShot [34]	62.85 ± 0.20	80.02 ± 0.14
TADAM[37]	58.50 ± 0.30	76.70 ± 0.30
ProtoNet[1]	60.37 ± 0.83	78.02 ± 0.57
R2-D2[38]	64.79 ± 0.45	81.08 ± 0.32
DeepEMD[6]	65.91 ± 0.82	82.41 ± 0.56
Baseline++[21]	53.97 ± 0.79	75.90 ± 0.61
MELR[39]	67.40 ± 0.43	83.40 ± 0.28
S2M2R[29]	64.93 ± 0.18	83.18 ± 0.11
CAN[32]	63.85 ± 0.48	79.44 ± 0.34
FEAT[40]	65.10 ± 0.20	81.11 ± 0.14
Ours	**66.02±0.19**	**84.7±0.13**

TABLE 2. Average accuracy (%) for 5way 1shot and 5way 5shot on CIFAR-FS.

Method	5-way 1-shot	5-way 5-shot
MetaOptNet[41]	72.0 ± 0.7	84.2 ± 0.50
RENet[3]	74.51±0.46	86.60±0.32
NCA nearest centroid[42]	72.49±0.12	85.15±0.09
R2-D2[38]	76.51 ± 0.47	87.63 ± 0.34
S2M2R[29]	74.81 ± 0.19	87.47 ± 0.13
ConstellationNet[43]	69.3 ± 0.3	82.7 ± 0.20
PLCM[44]	77.62±1.15	86.13±0.67
Ours	**72.76±0.21**	**86.71±0.15**

TABLE 3. Average accuracy (in %;) for 5way 1shot and 5way 5shot on FC-100.

Method	5-way 1-shot	5-way 5-shot
MetaOptNet[42]	41.1 ± 0.6	55.5 ± 0.60
TADAM[37]	40.10 ± 0.40	56.10 ± 0.40
ProtoNet[1]	41.54 ± 0.76	57.08 ± 0.76
MTL[9]	45.1 ± 0.18	57.6 ± 0.90
MAML[8]	38.1 ± 1.7	50.4 ± 1.0
J. Kim et al.[46]	42.31 ± 0.75	58.16 ± 0.78
Dhillon et al.[47]	43.16 ± 0.59	57.57 ± 0.55
Ours	**43.6±0.19**	**59.56±0.19**

tasks to the basic model, which enables the model to extract representative features well from both local and semantic information. This is a key factor in ensuring model robustness and generalization ability.

In the field of image recognition, image features are crucial information for models. Our method starts from this aspect and uses an attention mechanism to address the trade-off between local information and semantic information in images, placing more weight on more representative features. On the other hand, semantic information also plays an irreplaceable role in the recognition process. The model must remember the semantic information of each category to achieve generalization and robustness during inference. Therefore, the experimental results confirm the effectiveness of our method, i.e., attention helps the model extract more representative features from images, and self-supervised learning proxy tasks enrich the semantic information of the dataset. Combining the two can significantly improve the model's generalization ability and robustness in few-shot learning.

TABLE 4. Average accuracy (in %;) for 5way 1shot and 5way 5shot on CUB200.

Method	5-way 1-shot	5-way 5-shot
ProtoNet[1]	66.09 ± 0.92	82.50 ± 0.58
DeepEMD v2[6]	79.27 ± 0.29	89.80 ± 0.51
FEAT[40]	68.87 ± 0.22	82.90 ± 0.10
S2M2R[29]	80.68 ± 0.81	90.85 ± 0.44
RelationNet[3]	66.20 ± 0.99	82.30 ± 0.58
DEML[15]	67.28 ± 1.08	83.47 ± 0.59
MatchNet[2]	71.87 ± 0.85	85.08 ± 0.57
MAML[8]	67.28% ± 1.08	83.47 ± 0.59
Ours	**77.3±0.19**	**90.88±0.10**

D. ABLATION STUDIES

To further prove the effectiveness of the proposed method, this paper tests the effectiveness of the three-dimension module, the self-supervised rotation transformation proxy task, one by one in the experiments in this section. The results are shown in Table 5. A tick mark ✔ means the module is used, while a blank means the module or operation is not used.

It can be seen that the proposed method can achieve good results. Table 5 of the ablation experiment shows that whether it is only the attention mechanism or only the self-supervised learning module, the model can achieve higher accuracy. When the attention mechanism and self-supervised learning modules are fused at the same time, the model can achieve the best performance. For such experimental results, it can be fully demonstrated that the attention mechanism proposed in this paper can better help the model to obtain more representative features, thereby improving the robustness of the model. The improvement effect of self-supervised learning is higher than that of the attention mechanism, which shows that self-supervised learning plays an irreplaceable role in the field of few-shot learning because it can reduce manual labeling while obtaining more semantic information of samples.

TABLE 5. Average accuracy (in %) of ablation experiments of our method.

Dataset	Three dimension attention	self-supervised rotation transformation	5-way 1-shot	5-way 5-shot
Mini-Imagenet			64.59±0.19	82.60±0.13
	✔		65.40±0.20	82.64±0.13
		✔	66.95±0.19	84.55±0.13
	✔	✔	67.95±0.19	84.70±0.12
CIFAR-FS			64.16±0.22	85.23±0.15
	✔		64.22±0.22	85.39±0.15
		✔	67.00±0.22	85.86±0.16
	✔	✔	72.81±0.21	86.73±0.15
Fc-100			42.00±0.18	61.74±0.18
	✔		42.12±0.18	61.89±0.18
		✔	43.14±0.18	62.59±0.19
	✔	✔	46.83±0.19	63.49±0.19
CUB200			67.91±0.22	88.34±0.12
	✔		73.55±0.20	89.00±0.11
		✔	76.95±0.20	90.75±0.11
	✔	✔	77.3±0.19	90.88±0.10

The ablation experiment results in Table 5 show that self-supervised learning improves the basic model more significantly than the attention module. The model's progress can be further improved when the two are integrated. On mini-Imagenet, using the attention module alone can improve the model accuracy very well, which shows that three-dimension attention can better help the feature extraction network to notice more representative features. When the model only uses the rotation transformation proxy task, the model's performance can be well improved in the four data sets, which shows that the self-supervised rotation transformation proxy task can effectively provide the model with richer semantic information. Thereby enhancing the performance of the model—the robustness of extracted features.

When attention mechanisms are used alone, all datasets can achieve some improvement in accuracy, but compared to using self-supervised learning proxy tasks alone, the improvement is not very significant. This may be due to the characteristics of small sample datasets, where there are many categories but few samples per category. We hypothesize that the attention module is used to help the model better focus on representative features. Experimental results show that the attention mechanism can only play a better role in the fusion of self-supervised proxy tasks. Our idea is to enrich the semantic information of images through self-supervision, which can help the model obtain better robustness and generalization ability. As shown in Table 5, whether using self-supervised learning proxy tasks alone or in conjunction with attention mechanisms, the model can achieve high accuracy on all four datasets. Therefore, it can be considered that the role of self-supervised learning is more significant than that of attention mechanisms in the small sample domain. This also inspires us that future research should perhaps focus on investigating whether other forms of proxy tasks can enrich the semantic information of images, to verify the irreplaceable role of self-supervised learning in other few-shot domains.

V. CONCLUSION

We propose an end-to-end few-shot learning algorithm framework based on a three-dimension attention mechanism and a self-supervised rotation transformation pretext task. In this paper, we apply self-supervised learning to few-shot learning and develop an integrated framework that combines a three-dimension attention mechanism and self-supervised learning with few-shot learning. The feature extractor was trained using rotational proxy tasks and fused the attention mechanism to improve the model's ability to extract features and obtain higher generalization ability and robustness. Our experiment shows that our method exhibits better robustness. The model has better generalization ability and achieved new state-of-the-art results on Mini-ImageNet, FC100, Cifar-FS, and CUB datasets.

A. DECLARATIONS

1) CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. NIPS*, 2017, pp. 1–15.
[2] O. Vinyals, "Matching networks for one shot learning," in *Proc. NIPS*, 2016, pp. 1–22.

[3] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.
[4] S. Bartunov and D. P. Vetrov, "Few-shot generative modelling with generative matching networks," in *Proc. AISTATS*, 2018, pp. 670–678.
[5] G. R. Koch, "Siamese neural networks for one-shot image recognition,"in *Proc. ICML Deep Learn. Workshop*, 2015, pp. 1–30.
[6] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12203–12213.
[7] J. Y. Lim, K. M. Lim, S. Y. Ooi, and C. P. Lee, "Efficient-PrototypicalNet with self knowledge distillation for few-shot learning," *Neurocomputing*, vol. 459, pp. 327–337, Oct. 2021.
[8] C. Finn and P. S. A. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017, pp. 1126–1135.
[9] Q. Sun, Y. Liu, T. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 403–412.
[10] A. Nichol and J. Schulman, "Reptile: A scalable metalearning algorithm," 2018, *arXiv:1803.02999*.
[11] A. Antoniou, H. Edwards, and A. Storkey, "How to train your MAML," 2019, *arXiv:1810.09502*.
[12] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017, pp. 1–11.
[13] Y. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7278–7286.
[14] Y.-X. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small sample learning," in *ECCV*, 2016, pp. 1–19.
[15] F. Zhou and B. Z. W. Li, "Deep meta-learning: Learning to learn in the concept space," 2018, *arXiv:1802.03596*.
[16] T. Munkhdalai and H. Yu, "Meta networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 2554–2563.
[17] L. Yang, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *ICML*, 2021, pp. 11863–11874.
[18] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.
[19] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[20] X. Zhou, Y. Zhang, and Q. Wei, "Few-shot fine-grained image classification via GNN," *Sensors*, vol. 22, no. 19, p. 7640, Oct. 2022.
[21] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," 2019, *arXiv:1904.04232*.
[22] S. Woo, "CBAM: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.
[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[24] M. Carrasco, "Visual attention: The past 25 years," *Vis. Res.*, vol. 51, no. 13, pp. 1484–1525, Jul. 2011.
[25] T. Mikolov, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013, pp. 1–12.
[26] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2014, *arXiv:1312.6114*.
[27] T. N. Mundhenk, D. Ho, and B. Y. Chen, "Improvements to context based self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9339–9348.
[28] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. ECCV*, 2018, pp. 1–19.
[29] P. Mangla, M. Singh, A. Sinha, N. Kumari, V. N. Balasubramanian, and B. Krishnamurthy, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2207–2216.
[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[31] Y. Ding and Y. Liu, "A novel few-shot action recognition method: Temporal relational CrossTransformers based on image difference pyramid," *IEEE Access*, vol. 10, pp. 94536–94544, 2022.
[32] R. Hou, "Cross attention network for few-shot classification," in *Proc. NeurIPS*, 2019, pp. 1–12.
[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[34] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. van der Maaten, "SimpleShot: Revisiting nearest-neighbor classification for few-shot learning," 2019, *arXiv:1911.04623*.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[37] B. N. Oreshkin, P. Rodriguez, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," 2018, *arXiv:1805.10123*.
[38] J. Liu and F. C.-M. C. Lin, "Task augmentation by rotating for meta-learning," 2020, *arXiv:2003.00804*.
[39] N. Fei, Z. Lu, T. Xiang, and S. Huang, "MELR: Meta-learning via modeling episode-level relationships for few-shot learning," in *Proc. ICLR*, 2021, pp. 1–20.
[40] H. Ye, H. Hu, D. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8805–8814.
[41] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10649–10657.
[42] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," in *Proc. NeurIPS*, 2021, pp. 1–18.
[43] W. Xu, "Attentional constellation nets for few-shot learning," in *Proc. ICLR*, 2021, pp. 1–16.
[44] K. Huang, J. Geng, W. Jiang, X. Deng, and Z. Xu, "Pseudo-loss confidence metric for semi-supervised few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8651–8660.
[45] J. Kim and H. G. Kim Kim, "Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning," in *Proc. ECCV*, 2020, pp. 599–617.
[46] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," 2019, *arXiv:1909.02729*.

YONG LIANG received the Ph.D. degree from the College of Mechanical and Electronic Engineering, Northwest A&F University, China, in 2016. He is currently an Associate Professor with the Guilin University of Technology. His research interests include intelligent robot, machine vision, few-shot learning, FPGA, and edge computing.

ZETAO CHEN (Member, IEEE) is currently pursuing the degree with the School of Mechanical and Control Engineering, Guilin University of Technology, China. His main research interests include few-shot learning and object detection.

DAOQIAN LIN is currently pursuing the degree with the School of Mechanical and Control Engineering, Guilin University of Technology, China. His main research interests include intelligent robot and FPGA.

JIE LI is currently pursuing the degree with the School of Mechanical and Control Engineering, Guilin University of Technology, China. His research interests include FPGA and deep learning.

JUNWEN TAN is currently pursuing the degree with the School of Mechanical and Control Engineering, Guilin University of Technology, China. His main research interests include FPGA, deep learning, and intelligent robot.

ZHENHAO YANG is currently pursuing the degree with the School of Mechanical and Control Engineering, Guilin University of Technology, China. His research interests include FPGA and robot operation systems (ROS).

XINHAI LI is currently pursuing the degree with the School of Mechanical and Control Engineering, Guilin University of Technology, China. His research interests include industrial vision detection and few-shot learning.