IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 December 2023, accepted 11 January 2024, date of publication 19 January 2024, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3356187

RESEARCH ARTICLE

ViTFSL-Baseline: A Simple Baseline of Vision Transformer Network for Few-Shot Image Classification

GUANGPENG WANG, YONGXIONG WANG, (Member, IEEE), ZHIQUN PAN, XIAOMING WANG, JIAPENG ZHANG, (Graduate Student Member, IEEE), AND JIAYUN PAN

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Corresponding author: Yongxiong Wang (wyxiong@usst.edu.cn)

This work was supported by the Natural Science Foundation of Shanghai under Grant 22ZR1443700.

ABSTRACT Few-shot image classification, whose goal is to generalize to unseen tasks with scarce labeled data, has developed rapidly over the years. However, in traditional few-shot learning methods with CNNs, non-local features and long-rang dependencies of the image may be lost, and this leads to a poor generalization of the trained model. With the advantage of the self-attention mechanism of Transformer, researchers have tried to use vision transformer to improve few-shot learning recently. However, these methods are more complicated and take up a lot of computing resources, and there is no baseline to measure their effectiveness. We propose a new method called ViTFSL-baseline. We take advantage of vision transformer and train our model on all train set without episodic training. Meanwhile, we design a new nearest-neighbor classifier to used for few-shot image classification. Furthermore, in order to narrow the gap between difference of same class, we introduce centroid calibration in classifier after the feature extraction of backbone. We run the experiments on popular benchmarks to show that our method is a simple and effective for few-shot image classification. Our approach could be taken as the baseline upon vision transformer for few-shot learning.

INDEX TERMS Deep learning, few-shot learning, feature processing, image classification, vision transformer.

I. INTRODUCTION

Deep learning has achieved significant improvement on machine vision tasks such as image classification. The excellent performance, however, is based on training models with large number of annotated examples. Sometimes we need lots of labeled images for each novel category to accomplish recognition tasks even under the circumstances of pre-training on abundant dataset. The cost of labeled images is generally expensive and the scarcity of samples, such as rare animals species. The situations severely restrict the application of machine vision to recognize novel visual concepts. Regardless of the limitation, the visual recognition systems of human can learn novel categories with extremely few labeled instances. It is challenging and interesting to recognize novel categories with a few labeled samples for each new category. There needs to devise learning strategies for novel classes with limited labeled data. The common practice is taken as few-shot learning(FSL), which has attracted much more attention.

Various methods for the few-shot learning tasks contains meta-learning and whole-classification. During training on source domain dataset, some researchers employ meta-learning methods [1], [2], [3] to construct episodes that are same to the target domain, and learn a model that can generate to new target tasks quickly. Others utilize whole-classification methods [4], [5] argue that leveraging feature information from numerous images across all classes within the source domain dataset enhances performance. Subsequently, they fine-tune the

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

well-trained models to the target task using adaptation techniques.

While few-shot learning(FSL) approaches upon convolutional neural networks are extensively applied in machine vision field, there are some shortcomings. Firstly, a shallow network is usually employed for embedding with meta learning approaches, leading to the loss of semantic and spatial information in the feature map. Secondly, people usually fail to capture information of non-local and obtain long-range dependencies with the local receptive field of convolutional neural network.

Some researchers utilize long-range relationships of local image patches and transformer networks with the self-attention mechanism to achieve better image recognition performance. Recent works [6], [7], [8] have proposed either combining or replacing CNNs with transformer networks. These efforts often entail employing long token sequence learning within transformers to mitigate catastrophic forgetting. The transformer networks aim to capture more enhanced semantic features by learning from dataset in parallel. Transformer [9] and its variants have recently been taken as a prospective alternative to convolutional neural networks for machine vision. However, researchers always rely on large volume of images to train transformer. It is challenging for transformer to perform the target task with few annotated data.

When employing transformer networks directly for few-shot learning within conventional episodic paradigm [10], [11], the performance tends to be inferior. The common practice can only learn few features of images successively and cannot obtain overall cognition on the distribution of training set. Consequently, transformer networks have rarely been taken as the backbone to Few-Shot Learning(FSL) tasks. Transformer have comparable performance than ResNet [12] only after pretraining on large dataset [9]. However, the data-rich models might lack inherent inductive biases that are crucial for smaller datasets. Thus, it leads to unsatisfying performance because of the limited annotated data.

In this paper, we enable the inheritance of parameters from pre-trained vision transformer(ViT) [9] without incurring additional costs. We adopt transformer as the backbone to train on source domain and use the whole-classification method to learn a model. We capture a general sense of semantic feature rather than only episodic instances. We finetune the top layer of transformer during testing on target domain. Now some works [13] get down to apply ViT in image recognition, but there is no simple and effective method for few-shot image classification. Our method offers a baseline to measure the effectiveness of approaches with ViT in the future. We show that simple training on whole class of training set with pre-trained model is sufficient for few-shot learning. During testing, we simply union the backbone and the last few layers to train and achieve promising performances [14], [15], [16].Our method is simple, scalable and can be used as a baseline for transformer on few-shot setting(FSL).

Moreover, we propose a simple Nearest-neighbor Centroid classifier with L2-normalization(NCL) for few-shot learning,which leads to further improvements with the ViT network. Although the similar method has been previously considered, it has been overlooked that it outperforms many sophisticated few-shot learning methods.To address this aspect, we introduce a transformer-based model along with a novel nearest-neighbor classifier NCL for few-shot learning(FSL). Our method is called ViTFSL-baseline for few-shot learning. Consequently, ViTFSL-baseline can also be a significant baseline that has been overlooked.

To summarize, the main contributions are as follows:

1. A simple and effective transformer-based method is proposed for few-shot image classification. To the best our knowledge, we are among the first to take simple transformer as backbone network with whole-classification to achieve the few-shot learning tasks. Our approach could be acted as the simple baseline for transformer-based few-shot image recognition methods in the future.

2. A classifier NCL based on nearest-neighbor algorithm is proposed to suppress the impact of feature inconsistencies and employed to promote accuracy of few-shot image classification.

3. Extensive experiments demonstrate that our few-shot learning method achieves a high performance on benchmarks and outperform its counterparts in terms of accuracy.

The remainder of this paper is arranged as follows. Section II discusses related work, include meta-learning and whole-classification methods. Section III illustrates the content of the proposed model for few-shot learning in details. Section IV evaluates the performance of the proposed model with some comparable few-shot learning methods. Finally, the conclusion is given in Section V.

II. RELATED WORK

Few-shot learning is a challenging work in computer vision field where the goal is rapid generalization to unseen tasks. In recent years, researchers have proposed many novel approaches to address the few-shot learning problems, resulting in significant advancements. Most of these approaches fall under the umbrella of meta-learning [10], [17], which involves training a model on the meta-train dataset by mimicking few-shot image classification tasks.The well-trained learner is then applied to the meta-test dataset.These various meta-learning architectures for few-shot learning can be roughly categorized into two groups,optimization-based approaches and metric-based approaches.

Optimization-based approaches involve the application of an optimization process on the support set in an episode with the meta-learning method. Finn et al. [1] proposes MAML to learn a model agnostic initialization that can effectively generalize to new tasks with just a few gradient steps. In MetaOptNet [18], the feature map is extracted to be adapted well for a linear support vector machine(SVM) classifier. Another study [19] employs LSTM-based meta-learner to replace the stochastic gradient decent optimizer.

While optimization-based approaches provide clarity in the learning process, metric-based approaches can be utilized to learn a deep feature representation with a distance metric learning in the feature space. For instance, prototypical network [10] calculates the euclidean average feature of each category in support set and classifies new samples from the query set using the nearest-centroid algorithm. Sung et al. [11] proposes a relation module that compares the feature of support and query set samples, which is trained jointly with the backbone network. TADAM [20] is used to employ a task conditioned metric distance leading to a task-dependent metric space.

Our approach is similar to metric learning. We take vision transformer [9] as the backbone network and evaluate few-shot learning tasks with meta-learning method directly, however the performance of the few-shot image classification is poor.

While meta-learning approaches have shown significant improvements in few-shot learning, their effectiveness is challenged by a recent line of literature with simple whole-classification. The whole-classification is a classification model trained on the whole training label-set. TheÂ authors of Cosine classifier [21]and Baseline++ [22] replace the last linear layer with cosine classifier and train the whole-classification model on train dataset. They then adapt the well-trained model to few-shot image classification tasks of novel classes by fine-tuning a new classifier. These methods have demonstrated competitive performance in few-shot learning using non-episodic training paradigm. Many recent works have also focused on transformer-based methods. However, the effective adaptation of transformer-based methods into the few-shot image classification tasks with whole-classification is still underexplored.

This work aims to integrate the vision transformer(ViT) into the whole-classification models and explore the effectiveness of the solution. During training, we take ViT as the backbone of the whole-classification model and then process the feature with normalization. We adopt the nearest-neighbor classifier NCL to replace the linear layer on top. During testing, we adapt our well-trained model for few-shot classification on novel classes. We regard our method as ViTFSL-baseline for few-shot image classification tasks. Our results show that ViTFSL-baseline is a simple and effective few-shot learning baseline that has been overlooked. The solution avoids the need to train over large number of episodes as in meta-learning and achieves competitive performance compared to other sophisticated algorithms. Although each component in ViTFSL-baseline is not novel, to the best of our knowledge, previous works have not explored them as a whole.

III. METHOD

A. PROBLEM DEFINE

We train a model on a labeled dataset D_{base}, which consists of abundant images from each base class in C_{base}. With the well-trained model, our goal is to learn concepts on novel classes D_{novel} with a few labeled samples per class. The dataset D_{novel} also has many unlabeled images in each class C_{novel}, where $C_{novel} \cap C_{base} = \varnothing$. The few-shot learning task means the N-way K-shot image classification task in this work and includes a support set D_s and a query set D_q. Usually in a conventional few-shot image classification task, a small support set contains N classes with K images per class, which is sampled from D_{novel}, and the corresponding query set consists of images from the same N classes with Q images each class. The goal of few-shot learning task is to classify the $N \times Q$ query samples into N classes.

B. VISION TRANSFORMER AS FEATURE EXTRACTOR

1) IMAGE SEQUENTIALIZATION

The images are fed into ViT and a sequence of flattened patches $X_p \in R^{ND}$ are generated. where $N = HW/P^2$ is the number of patches in an image, D is the dimension of images, (H, W) and (P, P) are the resolutions of the image and patch, respectively. We denote the input image with resolution $H \times W$, the size of image patch as P.

2) PATCH EMBEDDING

To obtain the position information of patches in image, position embedding is also added to inputs. Therefore, the input sequence z_0 of ViT is denoted as:

$$z_0 = \left[x_p^1 E, x_p^2 E, \ldots, x_p^N E\right] + E_{pos}. \tag{1}$$

where N is the number of image patches, $E \in R^{(P^2C)\times D}$ is the patch embedding projection, and $E_{pos} \in R^{ND}$ denotes the position embedding.

The encoder layer of transformers contains L layers of multi-layer perceptron(MLP) and multi-head self-attention(MSA) blocks. Residual connections and LayerNorm are applied before and after every block, respectively. Therefore, the output of the l-th layer can be shown as follows:

$$Z_l' = MSA\left(LN\left(Z_{l-1}\right)\right) + Z_{l-1} \quad l \in 1, 2, \ldots, L. \tag{2}$$

$$Z_l = MLP\left(LN\left(Z_l'\right)\right) + Z_l' \quad l \in 1, 2, \ldots, L. \tag{3}$$

where Z_l is the image encoding operation and $LN(\cdot)$ is the layer normalization. The first token of the last encoder layer Z_L^0 is employed as the representation of global feature and feed into a classifier to get the final classification results and the potential information retained in the rest of the tokens is discarded.

C. OUR APPROACH

The goal of training stage is to learn a feature extractor f_θ and classifier C_b by training the model on the large labeled base dataset D_{base}. Then given a N-way K-shot few-shot task sampled from the novel dataset D_{novel}, we utilize the feature extractor f_θ with a new classifier C_n to classify novel classes with few labeled samples. We show that a simple method of training a base model and then adapting it to novel classes

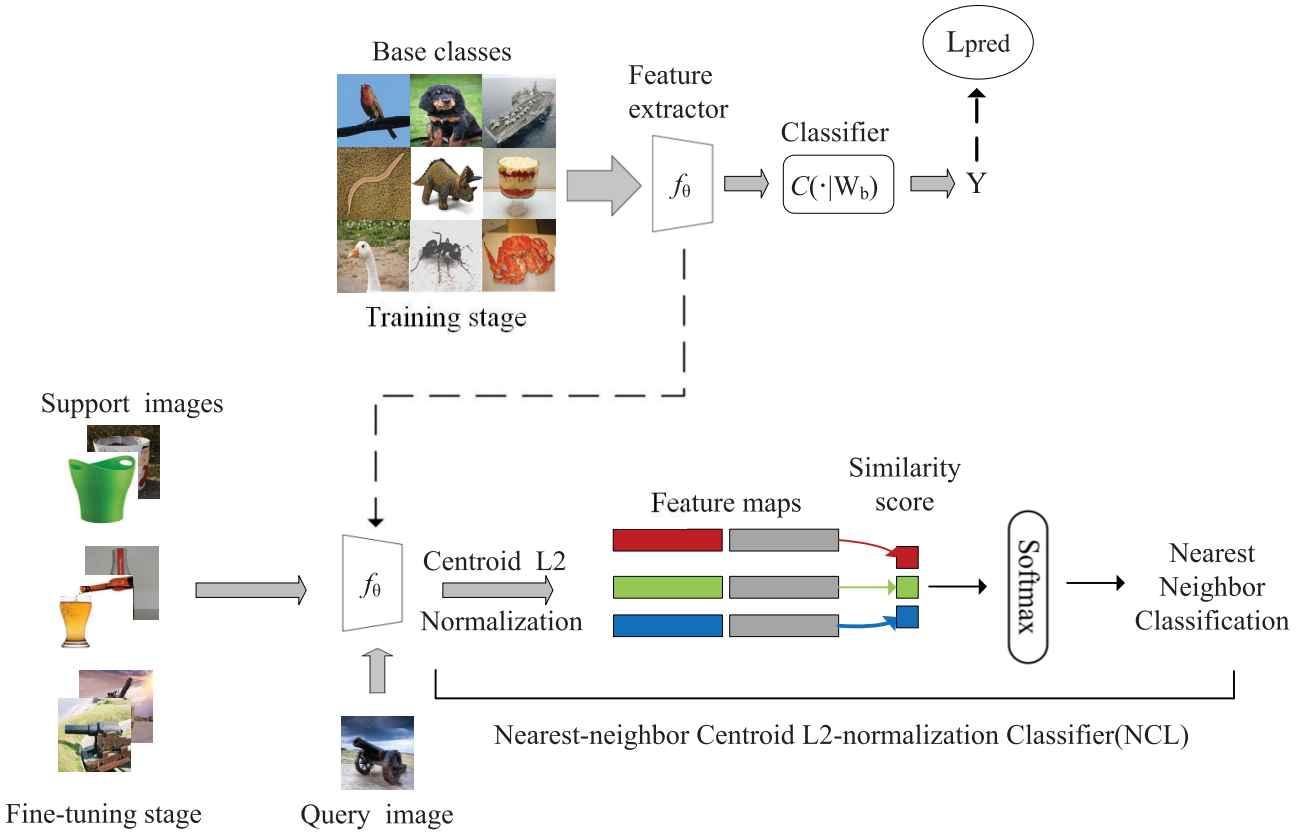

FIGURE 1. Network architecture for a 3-way 2-shot problem with one query example.

can obtain competitive performance. The overall framework is illustrated as Fig. 1.

In the training stage of Fig. 1, the samples of base classes are fed into feature extractor f_θ and classifier $C(\cdot \mid W_b)$ with the conventional cross-entropy loss. At the fine-tuning stage, the support samples are fed into backbone to get feature maps and then feature maps are processed by feature transformation. Then we obtain similarity scores by a distance metric between support and query feature maps and predict the class label. The cinerus rectangles and the colorful rectangles are feature maps of query set samples and support set samples respectively. The colourful squares before softmax are similarity scores between support and query images. Similarity scores are measured by Euclidean distance in feature space. The process of softmax are shown in Eq. (6).

To build a simple baseline, we introduce a whole-classification model which is trained on the whole class in base dataset D_{base}. We train the feature extraction network and a nearest neighbor classifier with conventional cross entropy loss on all base categories of train set. Then we utilize well-trained feature extraction network and a new classifier to deploy few-shot tasks with nearest neighbor algorithm in the novel dataset. Concretely, we train a model on all base categories with cross-entropy loss, then and retain the encoder f_θ to map data points to an embedding feature space.

A few-shot task consists of a support set S and a query set Q. S_c is defined as the few-shot examples in class c on the support set. We calculate the mean embedding W_c as the centroid of class c. We construct class centroids dynamically for each episode and then performs nearest centroid classification.

$$w_c = \frac{1}{\mid S_C \mid} \sum\nolimits_{x \in S_C} f_\theta(x). \tag{4}$$

A feature vector w_c is subtracted by $\widetilde{w_c}$ and then the result is normalized with L2 norm. $\widetilde{w_c}$ denotes feature vector of images. The formula is as follows:

$$\widetilde{w_c} \leftarrow \frac{\widetilde{w_c} - w_c}{\|\widetilde{w_c} - w_c\|_2}. \tag{5}$$

Mean subtraction operation does not change Euclidean distances between feature vectors, but it can achieve good performance in combination with L2-normalization. We find the centers to each category novel training sample and then average the samples along with L2 normalization to obtain the new centroid for the novel classes, similar to simpleshot [23]. In setting of K-shot, we get a centroid for each class by calculating the mean of the samples in same

class. At inference time, we simply get the nearest center to the query image and assign its label. The Nearest-neighbor Centroid classifier with L2-normalization(NCL) requires less memory and computation.

Interestingly, we obtain much closer performance to state of the art with this simple method. For query example x_i in few-shot tasks, we predict the label of sample x_i according to the distance between the feature of sample x_i and the mean vector of class c:

$$\hat{p}(y = c \mid x_i) = \frac{\exp\left(-d\left(f_\theta(x_i), \widetilde{w_c}\right)\right)}{\sum_{c'=1}^{C} \exp\left(-d\left(f_\theta(x_i), \widetilde{w_c'}\right)\right)}. \quad (6)$$

where $(\cdot,\cdot)$ denotes the distance of two vectors.

We are devoted to propose a simple baseline. There is no additional complex techniques and hyper-parameter optimization for the whole-classification training process. Each component in our ViTFSL-baseline has been proposed before, but we observe that none of the prior works exploits them as a whole. Consequently, ViTFSL-baseline can also be an significant baseline that has been overlooked.

D. LOSS FUNCTION

We exploit a few-shot classifier based on nearest-neighbor algorithm. The feature map $f_\theta(x_i)$ extracted from image x_i is processed by the nearest-neighbor learner.

$$\mathcal{L}_1 = -\min_\theta \frac{1}{\mid B \mid} \sum_{(x_i, y_i) \in D_{base}} \left(y_i \log\left(\hat{p}_i\right)\right) \quad (7)$$

where the cross-entropy loss is loss function $L1$, B is the number of training samples in D_{base}. (x_i,y_i) means refers to the samples and corresponding labels. $\hat{p}_i$ denotes the label that which needs to be predicted. The parameter θ is embedded in the $\hat{p}_i$ by the $f_\theta(x_i)$ in the Eq. (6). All the learnable weights involved in our method are finetuned by minimizing L_1.

IV. RESULTS

Experimental details and results are presented in the section. We run experiments on three benchmark datasets: miniImagenet [24], tieredImagenet [2], and CUB [25]. We also present analysis of experimentation to evaluate the effectiveness of our approach. Conventionally, all results are averaged over 1000 episodes for training and 600 tasks for testing.

A. DATASETS

- **MiniImageNet.** The miniImageNet [24] benchmark dataset is a subset of the ILSVRC-12 ImageNet dataset [26], which has 100 classes and each category has 600 images as the setup [24]. We divide the classes of miniImageNet into 64 base classes for training, 16 classes for validation, and 20 novel classes for the test [27].
- **TieredImageNet.** A subset of ILSVRC-12 ImageNet dataset [26], which consists a total of 34 categories and 608 classes. It is another popular benchmark with much larger scale. All the images are divided into 20 base categories (351 classes) for training, 6 validation categories (97 classes) and 8 novel categories (160 classes) for the test, following the [1] and [2]. The setup is more challenging since base classes and novel classes are disjoint classes and come from different super-categories.
- **CUB-200-2011(CUB).** Caltech-UCSD Birds-200-2011 [25] is an extended dataset of CUB-200. We call it CUB for short hereafter. The CUB benchmark dataset is originally presented for fine-grained bird classification task. CUB contains a total of 11,788 images, which belong to 200 classes. As the paper [22] suggested, these are grouped into 100 classes for training, 50 classes for validation and 50 classes for testing.
- **MiniImageNet→CUB.** Following [22] and [28], we use miniImageNet [24] as our base class and train a model on it. We random sample 50 classes for validation and 50 novel class for the test, all of the classes are from CUB. We use CUB [25] dataset to verify the domain transfer ability of model. In the experimental setting, we use all 100 classes of miniImageNet for training and utilize the testing set (50 classes) of CUB dataset for the test.

B. IMPLEMENTATION DETAILS

- **Experimental setting.** Following prior work [23], we train our model by randomly sampling 1000 N-way K-shot episodes from the training set. We obtain the average accuracy over all the tasks, and the average accuracy with 95% confidence interval. For a N-way K-shot task in testing set, the episode [11] contains N classes that each class includes K labeled samples for support set, and 15 query samples for query set. We verify our method on three benchmarks under the setting of 5-way 1-shot and 5-way 5-shot tasks. During testing, 600 tasks are randomly sampled to evaluate our model unless noted otherwise. For ease of comparison, we report the mean accuracy and the 95% confidence interval of the estimate of the mean accuracy.
- **Implementation details.** We are firstly to utilize pre-trained model on the base dataset with the ViT [9]. Then, we fine-tune our model on the few-shot tasks in the novel dataset. In this paper, we use ViT as backbone to extract feature. We do not train the model from scratch, but load the weights of pre-training to initialize our model. We add a softmax layer on the top of ViT to classify the classes. In addition, our method could be taken as the baseline model of ViT for few-shot learning. We run all our experiments on one NVIDIA TITAN GPU (12G) by Pytorch [29]. During training, we select horizontal flip, random crop as data augmentation. All networks are trained for 30 epochs and stochastic gradient descent is used as the optimizer to minimize

the cross-entropy loss for few-shot classification tasks. The initial learning rate is set to 0.01. We utilize the validation set to decide when to early stop, and do not use regularization techniques for simplicity.

- **Vision Transformer.** We employ the base vision of Version Transformer(ViT) [9] as the backbone to extract feature. All images are resized to 224 × 224 for ViT-base. We set the patch size as 16 for ViT-base. We utilize standard ViT-base with 12 layers, 12 attention heads, feature dimension as 768. MLP is used for the projection head with ReLU activation function in the hidden layer. The projection head is a MLP and it has two layers with GELU applied to the first fully-connected layer and LayerNorm applied to the second fully-connected layer. The MLP used to aggregate the dense score matrix has two layers with GELU applied to the first fully-connected layer and its output fullyconnected layer is 1-d.

C. EXPERIMENTAL RESULTS

We verify the effectiveness of our method on popular FSL datasets. Our approach is competitive with these state-of-art models based on inductive method released on recent literature. The results of 5-way 1-shot and 5-way 5-shot learning are shown in Table 1 as they are the most common practice. We didn't run experiments for other algorithms and get results from the original paper. When we choose methods to compare against, we consider three aspects about algorithms:1) the methods are published recently, 2) the methods use architectures most commonly, 3) the algorithms are not significantly more complicated than ours. Note that our goal is not to boost the state of the art, but rather to evaluate the performance of the ViTFSL-baseline method about few-shot learning which is severely underestimated in past years. With the nearest-neighbor classifier NCL, our vanilla ViT with whole-classification method is competitive with other methods in spite of simplicity. With the self-attention module in ViT, the network extracts the non-local information and generates more discriminative features. Generally, few-shot learning is implemented by meta-learning method, learns feature information in episodes of train set, which could extract feature only from several classes in episodes each time. Different from the common practice,our network is simple trained on whole class of training set. The train method makes the net learn more semantic feature and generate to the novel classes of test set better. We process the feature extracted by backbone network with centriod L2 normalization in NCL classifier. As shown in Table 1, The results are surprisingly well compared to approaches that utilize meta-learning, and also compared against the highlight simple baselines based upon pre-training with standard cross-entropy loss. Moreover, the centriod with L2 normalization of feature is a useful tool to reduce intra-class variation in the few-shot image classification setting. Here, the results demonstrate that ViTFSL-baseline boost the Baseline by a large margin and becomes competitive even when compared against other meta-learning approaches. Discussion and analysis in the next section show the effectiveness of our approach. The performance of QSFormer is slightly better than ours.The QSFormer involves global query-support sample Transformer branch and local patch Transformer learning branch.They pay attention to the cross-attention mechanism of support and query sets for image representation.Besides, they adopt a local patch Transformer to extract structural representation for each image sample by capturing the long-range dependence of local image patches. The proposed method outperforms most ResNet backbone networks in Table 1. The advantage of transformer is fully exploited based on whole-classification in our method. Transformer involves a mixture of local and global feature information, no matter for low level layers or high level layers. ResNet follows the process of extracting global features from local features more strictly. The Transformer network has a better ability to integrate global information. The skip connection structure in Transformer protects the transfer of representation from the bottom level to the higher level, whilst ResNet skip connections transmit less information at higher levels, which further significantly reduces the precision of local information in higher levels. The performance of HT is inferior than others. HT has only 3 basic Transformer layers. The parameter capacity is small, which makes it difficult to extract complex image feature well. Compared to other methods, our approach is very simple and easy to understand.

To further show the superiority of our model,as illustrated in Table 2, we also compare the results with other methods on another animals dataset CUB. For experiments on CUB dataset, we use the same training parameters for miniImageNet. The CUB is a fine-grained benchmark for few-shot image classification on birds dataset. The similarity between the base and novel classes is greater than previous datasets. It deserves our attention to understand how the feature generalization transfers from base to novel dataset. Our method shows reasonably well compared to previous approach indicating that our simple approach is alternative solution. We adopt transformer as the backbone to train on source domain and use the whole-classification method to learn a model. We capture a general sense of semantic feature rather than only episodic instances. The feature information extracted by our method is better than the method who is based on Resnet with episodic paradigm. The other methods can only be used to learn few features of images successively and cannot obtain overall cognition on the distribution of training set. In addition, our feature transformation in NCL classifier has positive effect on the performance. The current works tend to sophisticated algorithm and architectures for performance gains. Different from them, we set out to establish a baseline of ViT for few-shot image classification. Our results are comparable or better than current literature.

To validate the generalization ability of our model, we perform a cross-domain experiment by following setup in [22] and [35]. A cross-domain scenario is more challenging

TABLE 1. Average accuracies (%) of 1-shot and 5-shot classifiers for 5-way classification on miniImageNet and tieredImageNet. Our results are in bold. All results of competitors are from the original papers. '–': not reported. * means results reported in QSFormer [28]. Our results are averaged over 600 episodes. Higher is better. All the results are reported with 95% confidence intervals.

Model	Backbone	miniImageNet		tieredImageNet	
		1-shot	5-shot	1-shot	5-shot
CTX [30]	Conv-4	52.38 ± 0.20	68.34 ± 0.16	55.32 ± 0.22	73.12 ± 0.19
SSFormers [31]	Conv-4	55.00 ± 0.22	70.55 ± 0.17	55.54 ± 0.19	73.72 ± 0.21
ProtoNet [22]	ResNet10	51.98 ± 0.84	72.64 ± 0.64	-	-
RelationNet [22]	ResNet10	52.19 ± 0.83	70.20 ± 0.66	-	-
MAML [22]	ResNet10	51.98 ± 0.84	66.62 ± 0.83	-	-
TapNet [32]	ResNet12	61.65 ± 0.15	76.36 ± 0.10	-	-
MetaOptNet [18]	ResNet12	62.64 ± 0.61	78.63 ± 0.46	65.99 ± 0.72	81.56 ± 0.53
CAN [17]	ResNet12	63.85 ± 0.48	79.44 ± 0.34	69.89 ± 0.51	84.23 ± 0.37
DC [33]	ResNet12	62.53 ± 0.19	79.77 ± 0.19	-	-
Meta-Baseline [34]	ResNet12	63.17 ± 0.23	79.26 ± 0.17	68.62 ± 0.27	83.74 ± 0.18
DeepEMD [35]*	ResNet12	65.43 ± 0.28	79.28 ± 0.20	69.84 ± 0.32	84.06 ± 0.23
DeepBDC [36]*	ResNet12	60.76 ± 0.28	78.25 ± 0.20	63.03 ± 0.31	81.57 ± 0.22
QSFormer [28]	ResNet12	65.24 ± 0.28	79.96 ± 0.20	72.47 ± 0.31	85.43 ± 0.22
RFIC-simple [37]	ResNet12	62.02 ± 0.63	79.64 ± 0.44	69.74 ± 0.72	84.41 ± 0.55
NCA assignment [38]	ResNet12	62.55 ± 0.12	76.93 ± 0.11	68.35 ± 0.13	81.04 ± 0.09
MSKPRN [39]	ResNet12	59.20 ± 0.84	75.03 ± 0.68	-	-
Chen [22]	ResNet18	51.87 ± 0.77	75.68 ± 0.63	-	-
MTL [40]	ResNet18	62.10 ± 1.80	75.50 ± 0.80	-	-
Simpleshot [23]	ResNet18	62.85 ± 0.20	80.02 ± 0.14	69.09 ± 0.22	84.58 ± 0.16
wDAE-GNN [41]	WRN	61.07 ± 0.15	76.75 ± 0.11	68.18 ± 0.16	83.09 ± 0.12
LEO [42]	WRN	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09
HT [8]	Transformer	$54.10 \pm -$	$68.50 \pm -$	$56.10 \pm -$	$73.30 \pm -$
SUN [7]	Transformer	66.54 ± 0.45	82.09 ± 0.30	72.93 ± 0.50	86.70 ± 0.33
FewTURE [43]	Transformer	68.02 ± 0.88	84.51 ± 0.53	72.96 ± 0.92	86.43 ± 0.67
Ours	**Transformer**	$\mathbf{63.51 \pm 0.20}$	$\mathbf{80.30 \pm 0.14}$	$\mathbf{70.29 \pm 0.15}$	$\mathbf{84.12 \pm 0.27}$

TABLE 2. Average accuracies(%) of 1-shot and 5-shot classifiers for 5-way classification on CUB. +: Results reported in [44]. * means results reported in QSFormer [28]. Our results are in bold. All the results are reported with 95% confidence intervals. Our results are averaged over 600 episodes.All results of competitors are from the original papers. Higher is better.

Model	Backbone	CUB	
		1-shot	5-shot
MAML [1]+	ResNet18	68.42 ± 0.12	83.47 ± 0.19
RelationNet [11]+	ResNet18	68.58 ± 0.41	84.05 ± 0.15
Chen [22]+	ResNet18	67.02 ± 0.16	83.58 ± 0.12
SimpleShot [23]	ResNet18	65 ± 0.14	81.41 ± 0.20
Light transformer [45]	ResNet12	$70.25 \pm -$	$86.60 \pm -$
RelationNet [11]*	ResNet12	66.20 ± 0.99	82.30 ± 0.58
ProtoNet [10]*	ResNet12	70.93 ± 0.30	85.55 ± 0.19
MatchNet [24]*	ResNet12	70.21 ± 0.30	82.69 ± 0.22
DeepEMD [35]*	ResNet12	70.71 ± 0.30	86.13 ± 0.19
DeepBDC [36]*	ResNet12	65.45 ± 0.29	85.01 ± 0.19
QSFormer [28]*	ResNet12	75.44 ± 0.29	86.30 ± 0.19
Ours	**Transformer**	$\mathbf{76.41 \pm 0.20}$	$\mathbf{89.86 \pm 0.11}$

TABLE 3. Average accuracies(%) of 1-shot and 5-shot classifiers for 5-way classification on Cross-domain experiments(miniImagenet →CUB). * means results reported in QSFormer [28]. Our results are in bold. All the results are reported with 95% confidence intervals. Our results are averaged over 600 episodes.All results of competitors are from the original papers. Higher is better.

methods	1-shot	5-shot
ProtoNet [10]	50.01 ± 0.82	72.02 ± 0.67
MatchNet [24]	51.65 ± 0.84	69.14 ± 0.72
FEAT [46]*	52.67 ± 0.29	72.65 ± 0.25
DeepEMD [35]	54.24 ± 0.86	78.86 ± 0.65
DeepBDC [36]*	50.28 ± 0.27	76.49 ± 0.23
QSFormer [28]	55.04 ± 0.29	77.12 ± 0.24
Ours	$\mathbf{56.90 \pm 0.21}$	$\mathbf{76.42 \pm 0.17}$

due to the larger divergence between two datasets. We train our model on the miniImagenet dataset, and then evaluate the model on CUB few-shot tasks with test classes. As illustrated in Table 3, our ViTFSL-baseline achieves competitive performance on the 1-shot setting. For the QSFormer, our model also outperforms it on the 1-shot task by +1.86%. The proposed method is also better than DeepEMD on the 1-shot task. On the 5-shot task, our method is inferior than QSFormer and DeepEMD, however, is better than other approaches. The results demonstrate that our model extracts the discriminative feature across domains. Our model has a good generalization capability in the cross-domain setting.

D. VISUALIZATIONS

We verify the effectiveness of our method in this section. For the sake of illustration, we run our experiments on the birds dataset CUB.Specifically, we sample samples from target dataset, and employ our method to extract features. In order to compare with other approaches, the instances are also fed into baseline model [47]. Heatmaps are illustrated in Fig. 2. The goal of our model is to extract the non-local feature information in the images. Compared to baseline method, Our attention maps could cover more key parts and the cover area is bigger. The spatial correspondence can be recognized in the instances by our structure. Our method is better than the baseline approach in term of the cover area of object in the foreground. Fig. 3 is the confusion matrix for the birds images classification on

TABLE 4. The precision, specificity, sensitivity and f-score for CUB.

classes	Precision	Recall	F1-score
Sooty_Albatross	0.778	0.683	0.727
Least_Auklet	1.000	0.875	0.933
Rusty_Blackbird	0.833	0.333	0.476
Bobolink	0.821	0.939	0.876
Cardinal	0.826	0.905	0.864
Eastern_Towhee	0.812	0.886	0.847
Fish_Crow	0.685	0.860	0.763
Purple_Finch	0.842	0.780	0.810
Gadwall	0.782	0.935	0.852
Eared_Grebe	0.889	1.000	0.941
Pine_Grosbeak	0.769	0.667	0.714
Ivory_Gull	0.892	0.786	0.836
Blue_Jay	0.887	1.000	0.940
Mallard	0.864	0.809	0.836
Mockingbird	0.862	0.543	0.666
Nighthawk	0.902	0.860	0.880
Ovenbird	0.689	0.894	0.778
Sayornis	0.635	0.851	0.727
Geococcyx	0.933	1.000	0.965
Fox_Sparrow	0.885	0.548	0.677
Artic_Tern	0.808	0.955	0.875

TABLE 5. The precision, specificity, sensitivity and f-score for miniImageNet.

classes	Precision	Recall	F1-score
nematode	0.507	0.677	0.580
king_crab	0.743	0.663	0.701
gold_retriever	0.708	0.500	0.586
malamute	0.560	0.577	0.568
dalmatian	0.759	0.683	0.719
Afric_huntdog	0.605	0.837	0.702
lion	0.624	0.670	0.646
ant	0.576	0.470	0.518
black_ferret	0.467	0.593	0.523
bookshop	0.811	0.63	0.709
crate	0.583	0.623	0.602
cuirass	0.592	0.517	0.552
elec_guitar	0.484	0.637	0.550
hourglass	0.581	0.440	0.501
mixing_bowl	0.672	0.417	0.515
school_bus	0.826	0.917	0.869
scoreboard	0.869	0.820	0.844
thea_curtain	0.661	0.820	0.732
vase	0.480	0.367	0.416
trifle	0.659	0.740	0.697

CUB. The results are good for ''Eared_Grebe'', ''Blue_Jay'' and ''Geococcyx''. Our module has a better ability to distinguish between different images in the few-shot learning tasks. The performances are worse for ''Rusty_Blackbird'', ''Mockingbird'' and ''Fox_Sparrow''. Due to the more similarity with several classes, they are easily recognized other categories wrongly. Other corresponding metrics are

TABLE 6. The precision, specificity, sensitivity and f-score for tieredImageNet.

classes	Precision	Recall	F1-score
Siberianhusky	0.706	0.751	0.728
Goldfish	0.835	0.840	0.837
Schipperke	0.669	0.729	0.698
Tibetanmastiff	0.750	0.798	0.773
Orange	0.608	0.689	0.646
Bee	0.644	0.851	0.733
Dragonfly	0.834	0.694	0.758
Fence	0.787	0.692	0.736
Coffee	0.785	0.645	0.708
Winebottle	0.796	0.535	0.640
Sandbeach	0.874	0.820	0.846
bathtub	0.660	0.642	0.651
Teapot	0.478	0.611	0.536
Strawberry	0.831	0.763	0.796
Butterfly	0.913	0.915	0.914
Pizza	0.895	0.835	0.864
Electricray	0.655	0.734	0.692
Pomegranate	0.782	0.662	0.717
Fugu	0.666	0.688	0.677
Wirenetting	0.671	0.734	0.701
Teacup	0.593	0.512	0.550
Hamburger	0.898	0.811	0.852
Lemon	0.662	0.803	0.726

shown in Table 4. It can be seen from Fig. 4 that the classification performance is very good for ''Afric_huntdog'', ''school_bus'' and ''scoreboard''. Our module has a better distinguishing ability for these classes. The performance metrics for ''hourglass,'' ''mixing_bowl,'' and ''vase'' are comparatively lower. This is attributed to their higher resemblance to several other classes, leading to a higher likelihood of misclassification into different categories, potentially resulting in lower evaluation scores for these three classes. Other corresponding metrics are shown in Table 5. Fig. 5 illustrates the notably strong classification performance achieved for 'Goldfish,' 'Bee,' and 'Butterfly' by our model. These categories exhibit distinct recognition due to the robust capabilities of our model. However, comparatively lower performance metrics are observed for 'Teacup,' 'Teapot,' and 'Winebottle.' This is attributed to their greater similarity with several other classes, leading to a higher probability of misclassification into different categories and subsequently resulting in lower evaluation scores for these specific classes. Further corresponding metrics are detailed in Table 6. Fig. 6 is the t-SNE plots of target images based on our method and baseline. We could observe that the discrimination of the features learned by our approach is higher than that of baseline. The features obtained by baseline are mixed, while the clusters of the features generated by our model are separable, the distances between different clusters are relatively larger. These positive performances verify the effectiveness of our method.

FIGURE 2. Comparison of heatmap visualization between Baseline and our method. The attention maps cover on the images in form of heatmap. We can observe that our method is better than baseline on the fraction of coverage.

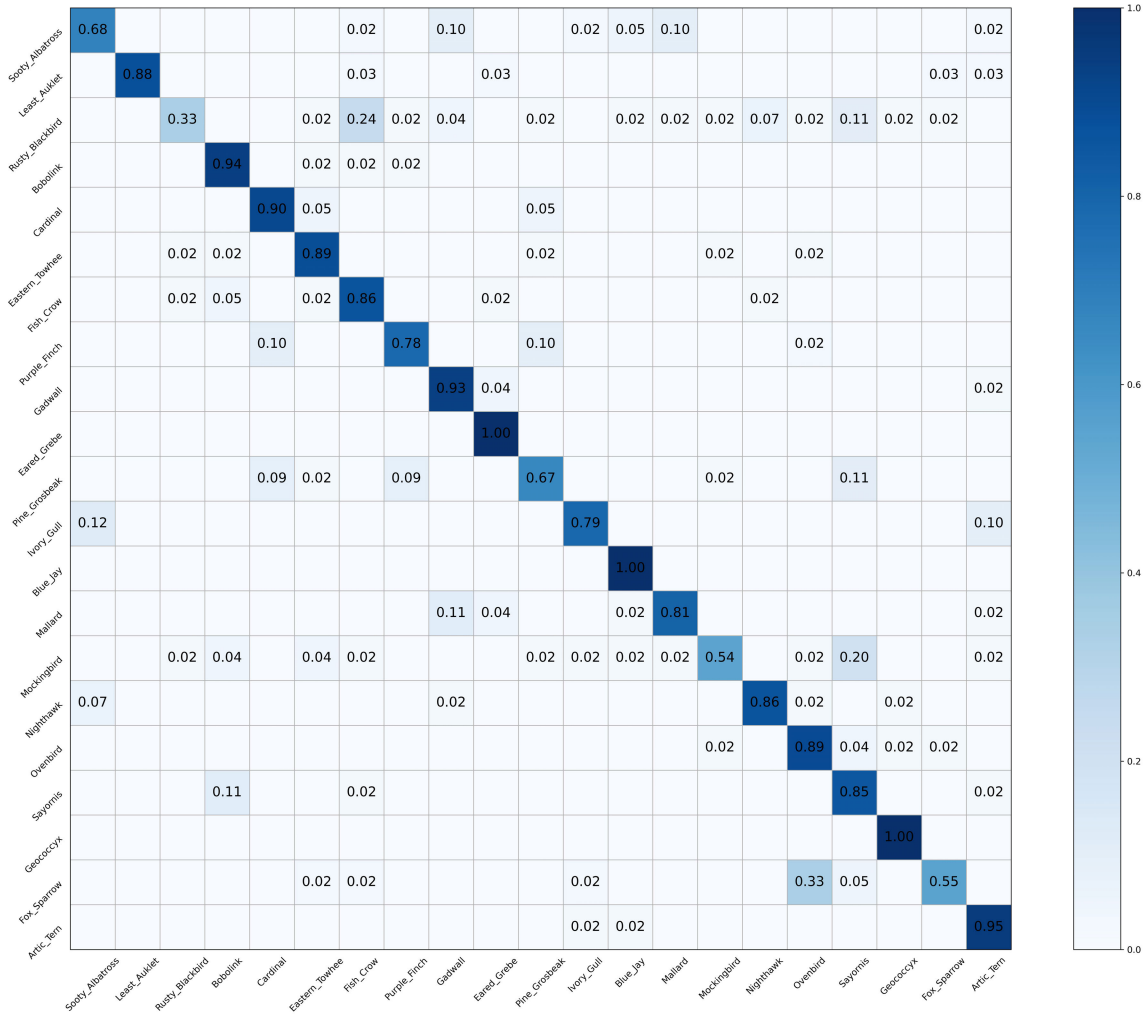

FIGURE 3. Confusion matrix for the birds images classification on CUB.

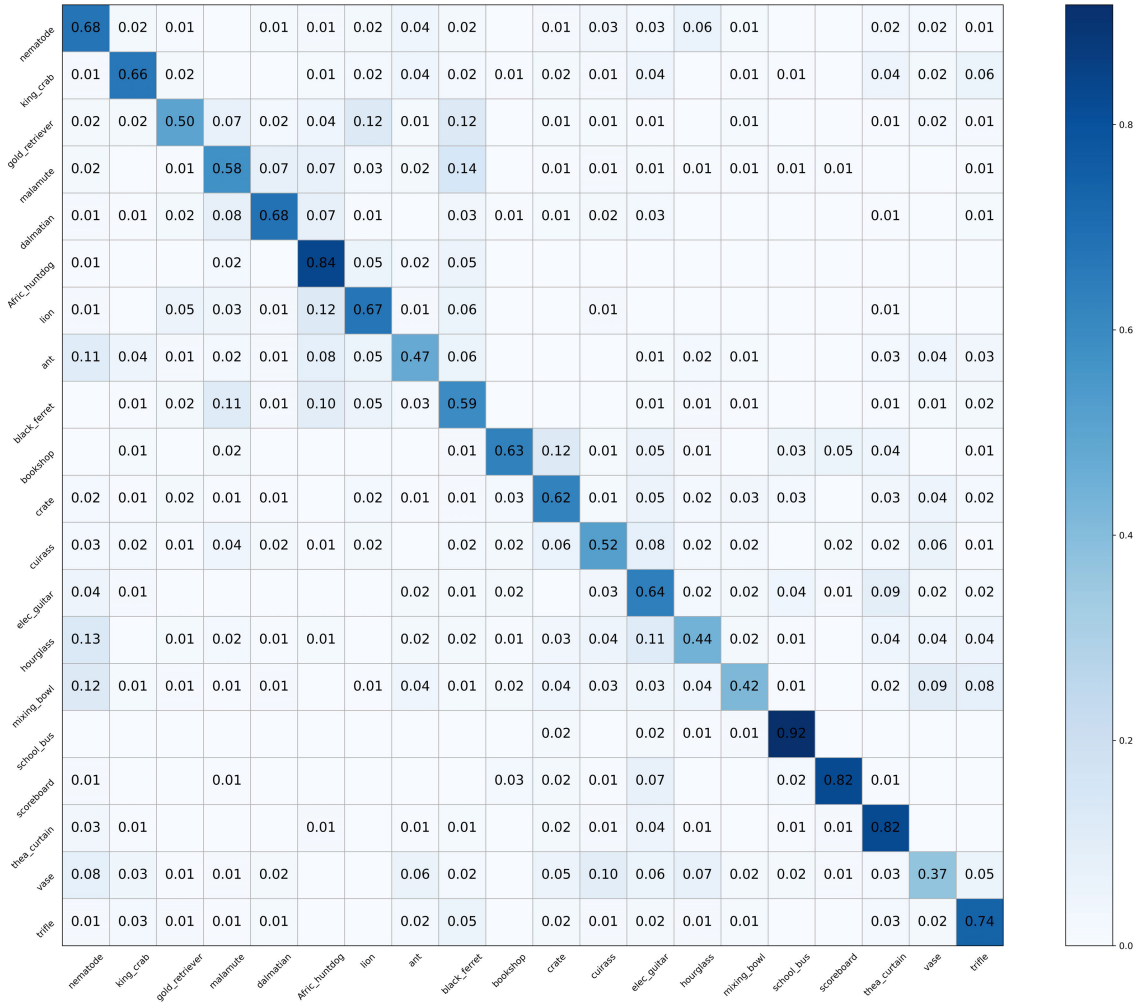

FIGURE 4. Confusion matrix for image classification on miniImageNet.

E. ANALYSIS OF EXPERIMENTS

We run simple ablation experiments on miniImageNet for 5-way 1-shot and the backbone is ViT. We select different kinds of measure distance and feature process to verify the effectiveness of our NCL classifier. Table 7 shows our results. Feature processing techniques are helpful for the improvement. In Table 7, the feature processing techniques utilized are U, L and C, which are like the method simpleshot. U is stand for unprocessed feature, L is L2 normalization for feature. C represents we subtract mean of feature first and then do L2 normalization. That is to say, C denotes equation (5). The experiments show that classifiers with C outperform their U and L counterparts. Euclidean NCL classifier works a bit better than NCL classifier with squared cosine distance. The NCL classifier with L1 distance achieves worst of all. The L1 distance cannot fit the distance of features effectively. We do not claim novelty in this literature. ViTFSL-baseline could be taken as a strong baseline for few-shot learning with transformer networks. It is a combination of different techniques. Our method is sometimes better than or comparable to existing algorithms on few-shot learning. We also conduct experiments to evaluate the whole-classification method with ViT. The results are shown in Table 8. The meta-learning methods, ProtoNet and MatchNet, whose backbone network are ConvNet-4, are classical algorithms in few-shot learning. Although the ViT backbone is more sophisticated than the ConvNet-4, we implementation the two methods with ViT and obtain much worse performance than the original paper. It can be observed that the transformer is not suitable for meta-learning methods. We can try to use transformer with whole-classification for few-shot learning in the future.

The parameters and FLOPs are shown in Table 9. Although the parameters of our method is highest of all, our method is simplicity itself. The FLOPs of our method is in the middle level, even is lower than the method QSFormer and Chen, which are based on ResNet12 or ResNet18. The results show

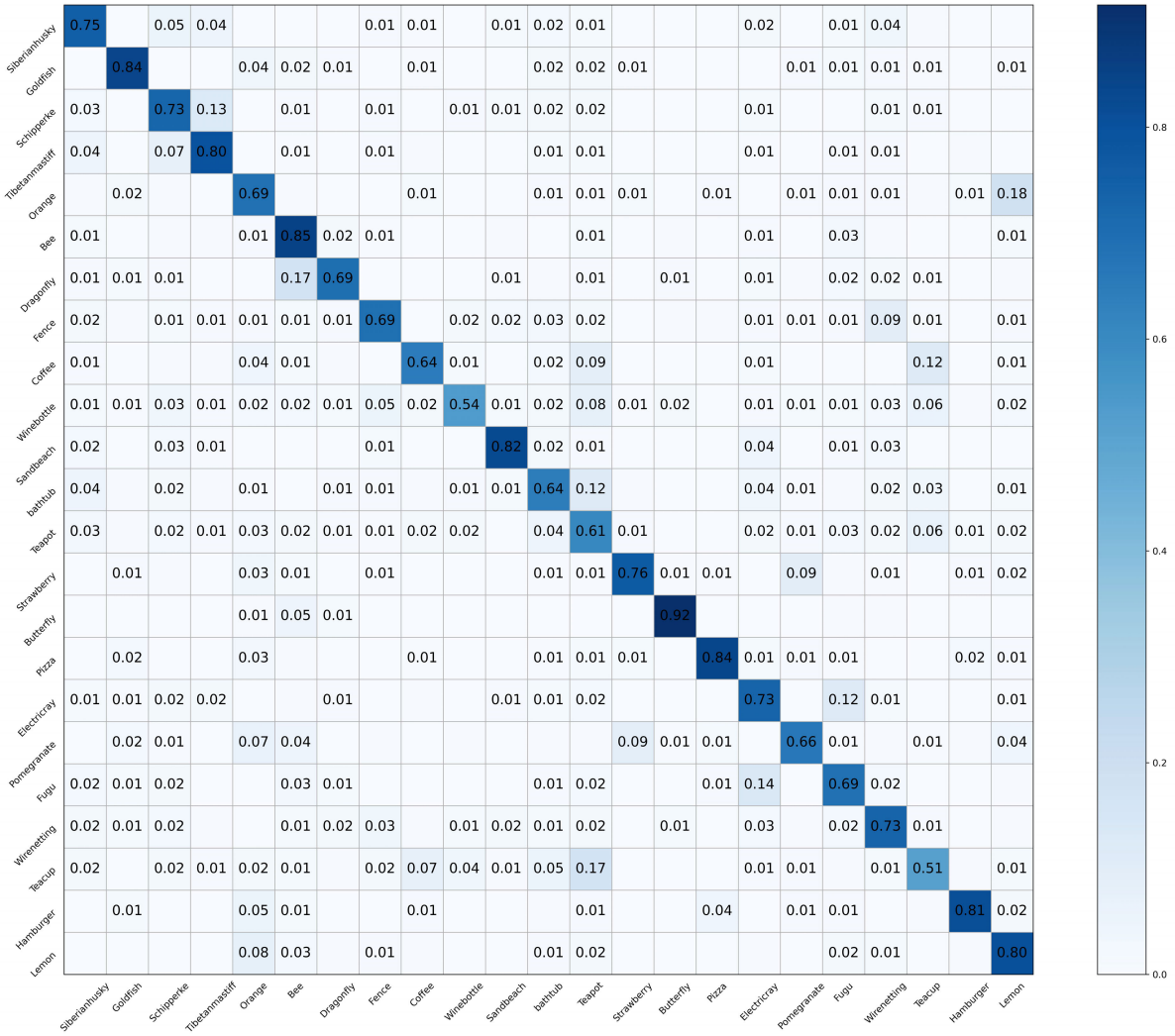

FIGURE 5. Confusion matrix for image classification on tieredImageNet.

TABLE 7. The experiments for 5-way 1-shot classification on miniImageNet.The methods utilize ViT as the backbone. All the results(%) are reported with 95% confidence intervals. Higher is better. Our results are averaged over 600 episodes.

Methods	Feature process		
	U	L	C
cosine	60.52 ± 0.21	61.58 ± 0.20	62.19 ± 0.20
Euclidean	61.59 ± 0.20	62.60 ± 0.21	63.51 ± 0.20
L1	20.17 ± 0.01	20.53 ± 0.02	20.78 ± 0.06
L2	60.08 ± 0.20	60.92 ± 0.14	61.80 ± 0.20

TABLE 8. The experiments for 5-way 1-shot classification on miniImageNet. * means the meta-learning methods are reimplemented with ViT.

methods	backbone	accuracy
ProtoNet [10]	ConvNet4	49.42 ± 0.78
ProtoNet [10]*	Transformer	39.76 ± 0.20
MatchNet [24]	ConvNet4	43.56 ± 0.84
MatchNet [24]*	Transformer	39.58 ± 0.20
Ours	**Transformer**	$\mathbf{63.51 \pm 0.20}$

that the proposed method is relatively simple and takes up fewer computing resources.

We've provided a description of the applicability of our proposed methods. Our approach utilizes a transformer model based on a self-attention mechanism, primarily suited for natural data within the ImageNet series. These datasets exhibit even feature distribution, allowing for easy extraction of global features. For instance, miniImageNet and tieredImageNet showcase compatibility with our method. However, our method's performance diminishes when applied to datasets characterized by uneven feature distribution, significant scale differences, and an abundance of local details, such as remote sensing data set.

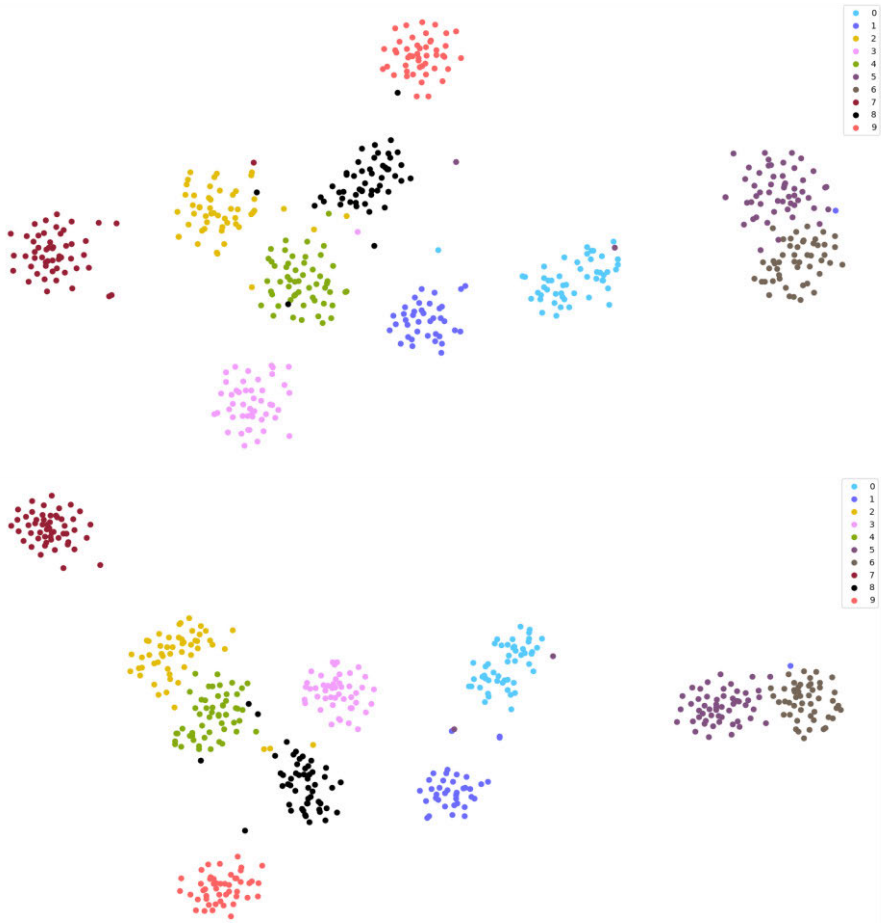

FIGURE 6. t-SNE plots of tested embedding of birds samples on CUB. The top images is performance of the baseline, the bottom is results for our approach.

TABLE 9. Comparisons of parameters and FLOPs.

Methods	Backbone	Params(M)	FLOPs(G)
QSFormer [28]	ResNet12	12.42	25.74
Chen [22]	ResNet18	11.68	27.34
LEO [42]	WRN	36.49	292.13
HT [8]	Transformer	21.91	4.31
SUN [7]	Transformer	39.48	23.43
FewTURE [43]	Transformer	21.59	4.25
Ours	**Transformer**	**85.65**	**16.88**

V. CONCLUSION

While almost all present few-shot image classification approaches are based upon the meta-learning algorithm, we propose a new whole-classification model for few-shot image recognition with ViT to effectively extract the feature from meta-train dataset and adapt well to the novel class data. We propose a new simple framework ViTFSL-baseline by incorporating the whole-classification model with ViT and a novel NCL classifier for few-shot image recognition learning. We take the advantage of ViT to train a large number of data, and then process the features in the classifier, so that similar features are easy to be aggregated and classified. We run extensive experiments on the few-shot recognition tasks to demonstrate that our proposed ViTFSL-baseline achieves appealing performance. Meanwhile, we also illustrates the effectiveness and simple of our solution, which only employs the whole-classification method and NCL classifier. In our future work, we aim to explore the attention mechanism between support set and query set based on the current work. At the same time, the model should be simplified and have good performance in the semi-supervised few-shot learning scenario.

REFERENCES

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

[2] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, "Meta-learning for semi-supervised few-shot classification," 2018, *arXiv:1803.00676*.

[3] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9924–9933.

[4] C.-N. Yu and Y. Xie, "A study on representation transfer for few-shot learning," 2022, *arXiv:2209.02073*.

[5] A. Subramanya and H. Pirsiavash, "A simple approach to adversarial robustness in few-shot image classification," 2022, *arXiv:2204.05432*.

[6] Y. He, W. Liang, D. Zhao, H.-Y. Zhou, W. Ge, Y. Yu, and W. Zhang, "Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9109–9119.
[7] B. Dong, P. Zhou, S. Yan, and W. Zuo, "Self-promoted supervision for few-shot transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–6.
[8] A. Zhmoginov, M. Sandler, and M. Vladymyrov, "Hypertransformer: Model generation for supervised and semi-supervised few-shot learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 27075–27098.
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–12.
[10] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017.
[11] 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.
[12] 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.
[13] Z. Pan, Y. Wang, and B. Tan, "Regressive pseudo label for weakly supervised facial landmark detection," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17979–17988, Sep. 2022.
[14] A. Nakamura and T. Harada, "Revisiting fine-tuning for few-shot learning," 2019, *arXiv:1910.00216*.
[15] X. Wang, T. E. Huang, T. Darrell, J. E. Gonzalez, and F. Yu, "Frustratingly simple few-shot object detection," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9919–9928.
[16] J. Cai and S. Mei Shen, "Cross-domain few-shot learning with meta fine-tuning," 2020, *arXiv:2005.10544*.
[17] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, 2019, pp. 1–9.
[18] 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.
[19] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–10.
[20] B. Oreshkin, P. R. López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 31, 2018, pp. 1–11.
[21] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4367–4375.
[22] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–6.
[23] 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*.
[24] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.
[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Tech. Rep., 2011.
[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[27] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-To-class measure for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7253–7260.
[28] X. Wang, X. Wang, B. Jiang, and B. Luo, "Few-shot learning meets transformer: Unified query-support transformers for few-shot classification," 2022, *arXiv:2208.12398*.
[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Workshop Autodiff Submission, 2017, pp. 1–4.
[30] C. Doersch, A. Gupta, and A. Zisserman, "CrossTransformers: Spatially-aware few-shot transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21981–21993.
[31] H. Chen, H. Li, Y. Li, and C. Chen, "Sparse spatial transformers for few-shot learning," *Sci. China Inf. Sci.*, vol. 66, no. 11, Nov. 2023, Art. no. 210102.
[32] S. W. Yoon, J. Seo, and J. Moon, "Tapnet: Neural network augmented with task-adaptive projection for few-shot learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 7115–7123.
[33] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, "Dense classification and implanting for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9250–9259.
[34] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9042–9051.
[35] 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. (CVPR)*, Jun. 2020, pp. 12200–12210.
[36] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep Brownian distance covariance for few-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7962–7971.
[37] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking few-shot image classification: A good embedding is all you need?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 266–282.
[38] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24581–24592.
[39] M. Abdelaziz and Z. Zhang, "Multi-scale kronecker-product relation networks for few-shot learning," *Multimedia Tools Appl.*, vol. 81, no. 5, pp. 6703–6722, Feb. 2022.
[40] Q. Sun, Y. Liu, T.-S. 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.
[41] S. Gidaris and N. Komodakis, "Generating classification weights with GNN denoising autoencoders for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 21–30.
[42] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–9.
[43] M. Hiller, R. Ma, M. Harandi, and T. Drummond, "Rethinking generalization in few-shot classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, 2022, pp. 3582–3595.
[44] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, "Laplacian regularized few-shot learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 11660–11670.
[45] H. Zhu, R. Zhao, Z. Gao, Q. Tang, and W. Jiang, "Light transformer learning embedding for few-shot classification with task-based enhancement," *Appl. Intell.*, vol. 53, no. 7, pp. 7970–7987, 2023.
[46] H.-J. Ye, H. Hu, D.-C. 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.
[47] G. Wang and Y. Wang, "Self-attention network for few-shot learning based on nearest-neighbor algorithm," *Mach. Vis. Appl.*, vol. 34, no. 2, p. 28, Mar. 2023.

GUANGPENG WANG received the B.E. degree from the Liaoning University of Technology, Jinzhou, China, and the M.E. degree from the Shandong University of Science and Technology, Qingdao, China. He is currently pursuing the Ph.D. degree in control science and engineering with the University of Shanghai for Science and Technology, Shanghai, China.

His research interests include computer vision and applications of deep learning.

YONGXIONG WANG (Member, IEEE) received the B.S. degree from Harbin Engineering University, Harbin, China, and the M.S. and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China.

He is currently a Professor with the University of Shanghai for Science and Technology. His research interests include computer vision and intelligent robot.

ZHIQUN PAN received the B.E. degree from Ludong University, Yantai, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China.

His current research interests include deep learning, computer vision, face analysis, weakly supervised learning, and domain adaptation.

XIAOMING WANG received the B.E. degree from the Shanghai University of Electric Power, Shanghai, China, and the M.E. degree from Shanghai Jiao Tong University, Shanghai. She is currently pursuing the Ph.D. degree in control engineering with the University of Shanghai for Science and Technology, Shanghai.

Her research interests include computer vision and applications of deep learning.

JIAPENG ZHANG (Graduate Student Member, IEEE) received the B.E. degree in electronic information engineering from Soochow University, China. He is currently pursuing the Ph.D. degree with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include deep learning, computer vision, and medical image analysis.

JIAYUN PAN is currently pursuing the bachelor's degree in intelligent science and technology with the University of Shanghai for Science and Technology, Shanghai, China.

Her research interests include computer vision and applications of deep learning.