Boosting Few-shot Learning by Self-calibration in Feature Space

Kaipeng Zheng[*]
University of Electronic Science and Technology of China
Chengdu, China
kaipengm2@gmail.com

Liu Cheng[*]
Southwest University
Chongqing, China
cliuarsa@gmail.com

Jie Shen[†]
University of Electronic Science and Technology of China
Chengdu, China
sjie@uestc.edu.com

ABSTRACT

Few-shot learning aims at adapting models to a novel task with extremely few labeled samples. Fine-tuning the models pre-trained on a base dataset has been recently demonstrated to be an effective approach. However, a dilemma emerges as whether to modify the parameters of the feature extractor. This is because tuning a vast number of parameters based on only a handful of samples tends to induce overfitting, while fixing the parameters leads to inherent bias in the extracted features since the novel classes are unseen for the pre-trained feature extractor. To alleviate this issue, we novelly reformulate fine-tuning as calibrating the biased features of novel samples conditioned on a fixed feature extractor through an auxiliary network. Technically, a self-calibration framework is proposed to construct improved image-level features by progressively performing local alignment based on a self-supervised Transformer. Extensive experiments demonstrate that the proposed method vastly outperforms the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Image representations**.

KEYWORDS

few-shot learning, image recognition, self calibration

ACM Reference Format:
Kaipeng Zheng, Liu Cheng, and Jie Shen. 2022. Boosting Few-shot Learning by Self-calibration in Feature Space. In *Proceedings of the 1st International Workshop on Methodologies for Multimedia (M4MM '22), October 14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3552487.3556437

1 INTRODUCTION

Deep learning has achieved great success in multiple tasks of computer vision and natural language processing. However, it requires sufficient training data to guarantee the generalization performance, which is usually difficult to meet in practical scenarios. Few-shot learning has thus been extensively studied to address the problem of model adaptation on novel tasks with extremely few labeled samples.

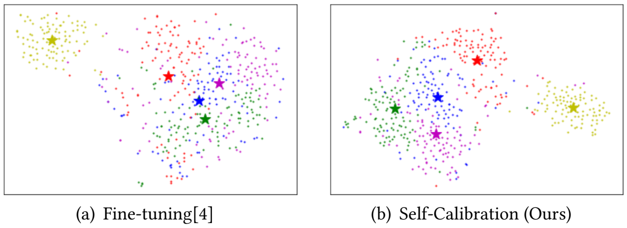
(a) Fine-tuning[4]
(b) Self-Calibration (Ours)

Figure 1: t-SNE visualization of the query sample features when adapting a feature extractor pre-trained on the base dataset to a novel 5-way 1-shot task: (a) fine-tuning the final fully connected layer[4] (b) self-calibration (ours)

It is very difficult to train a completely new feature extractor from scratch using only those few samples on the novel task, as this can easily lead to overfitting. A practical solution is to first pre-train the feature extractor using a base dataset with sufficiently annotated samples, and then adapt the model to the novel task. Meta-learning based approaches have dominated previous studies in few-shot learning . The core idea is to mimic the construction of the novel task by continuously sampling tasks from the base dataset to train the feature extractor. However, two major limitations of meta-learning have been observed by recent researches[5, 9], including (1) Poor scalability. Knowing in advance how the target task is constructed is crucial for meta-learning[9], as it requires sampling the base task in the same way for training to produce the best performance. This will result in poor scalability when applied to real-world scenarios with varying numbers of classes and samples. (2) Difficult to train. Researches[5] have demonstrated that during the training process of meta-learning, as the performance on base tasks improves, the performance on novel tasks instead is likely to decrease due to objective discrepancy. Optimal performance always requires a fine-grained training strategy to be obtained.

Recent researches[1, 4, 7, 9, 24] have shown that competitive performances can be achieved through fine-tuning[4] without the need for sophisticated meta-learning scheme. This is achieved by pre-training the feature extractor on the base dataset using a cross-entropy loss and fine-tuning the parameters of the final fully-connected layer on the novel task. The success of this approach can be attributed to the partially shared low-level semantic features between different classes. Compared with meta learning, fine-tuning is more robust in dealing with varying numbers of categories and samples[9], and is also easier to train. The effectiveness of this alternative approach is highlighted by previous studies, and

[*] Equal contribution
[†] Corresponding author

M4MM '22, October 14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9505-2/22/10…$15.00
https://doi.org/10.1145/3552487.3556437

further improvements are made by leveraging base class samples to aid transferring[1, 7, 24]. Afrasiyabi et al.[1] propose to fine-tune the feature extractor with the help of base class samples that are similar to the novel samples in the feature space. ConFT[7] and POODLE[24] demonstrate it is effective to treat base class samples as negative samples and train with contrastive loss. However, for a pre-trained model, the samples used for training are likely to be unavailable in practical scenarios, which will become a bottleneck of these methods.

In this paper, we also investigate on improving the performance of fine-tuning based methods. Moreover, we target a more challenging setting of not using base class datasets as well as any external datasets. However, in such a scenario, a dilemma of fine-tunning emerges. On the one hand, adapting an entire feature extractor solely using a handful of novel samples is a daunting task. On the other hand, based on a fixed feature extractor pre-trained on the base dataset, bias and misalignment are inherent in the feature of a novel sample since it belongs to an unseen class for the feature extractor (fig. 1-(a)). Only limited gains can be achieved by merely fine-tuning the final fully connected layer. Based on these observations, we discard the routine method of directly tuning the parameters of the network to adapt to novel tasks. Instead, we fix the pre-trained feature extractor to retain the prior learned from base classes and reformulate fine-tuning as calibrating biased features via an auxiliary network. The hope is to seek a transformation that generates more discriminative representations of novel samples. Inspired by the unsupervised methods in representation learning[3, 16, 18], we propose to calibrate the features in a self-aligned manner, aiming at leveraging the structural information in feature space to learn an improved embedding. A conventional step in converting the input image into a feature vector is to flatten the feature map by global average pooling. However, it leads to loss of detailed local information. Exploiting local information has been demonstrated to be effective for feature enhancement in various computer vision tasks[11, 19, 41]. Based on these observations, we propose to introduce detailed local information of images for progressive alignment to finally construct improved image-level representations.

To this end, we propose a self-calibration framework. For a novel sample, we slice it into multiple overlapped local patches and extract the features through a fixed feature extractor. Then we employ Transformer as an auxiliary network to progressively aggregate the features of the local patches in conjunction with contextual information to finally construct improved image-level features (fig. 1-(b)), which is trained in a completely unsupervised manner. We verify that through extensive experiments the proposed method significantly improves the performance of the conventional fine-tuning.

Our contributions can be summarized as follows.

(1) We propose a self-calibration framework for few-shot learning that constructs improved image-level features of novel samples by progressively performing self-alignment on local features based on Transformer.

(2) We novelly reformulate fine-tuning in few-shot learning as feature calibration through an auxiliary network and gain improvements under a challenging setting of no exploitation of the base dataset as well as any additional data.

(3) We conduct extensive experiments and show that the proposed method exceeds the conventional fine-tuning[4] by 8%, achieving the new state-of-the-arts. Comprehensive analysis is also included to verify the effectiveness of the components.

2 RELATED WORK

2.1 Few-shot Learning

Existing researches on few-shot learning are primarily based on meta-learning. Several recent studies propose that competitive results can be achieved by simply fine-tuning the pre-trained model without resorting to sophisticated meta-learning frameworks. We then provide a brief overview of the representative works in both types.

Meta-learning: Meta-learning based approaches in few-shot learning can be categorized into optimization-based methods, augmentation-based methods and metric learning based methods. Optimization-based methods[12, 25, 34] investigate to fast optimize the model on the novel tasks, focusing on designing novel learning strategies. Augmentation based methods[17, 35, 49] is motivated by the intuition of expanding the sample size on the novel task. They are primarily concerned with generating augmented samples by synthesizing features in the latent space. Metric learning based methods[20, 36, 38, 42, 44, 45, 48, 50, 52] have predominated in recent studies. They perform nearest-neighbour classification on the extracted features based on similarities, focusing on developing advanced metrics and improved feature representations. For recent studies, efforts have been devoted to meta-train the model to learn pixel-level alignment on feature maps as well as task-specific adaptation. For instance, CAN[20] aligns feature maps by a cross-attention module. Feat[50] meta-learns a set-to-set transformation to enable features to be task-adaptive. DMF[48] incorporates channel-wise alignments to make further improvements. Despite previous works have also highlighted the use of local information, they are based on sophisticated meta-learning schemes, whereas our method is fine-tuning based and only relies on a simple auxiliary network that is much easier to train. Another relevant work to our method is proposed by Zhang et al.[51], which learns to complete prototypes of novel classes by leveraging additional attribute annotations that are cross-category and one-hot labeled. In contrast, our work does not require the use of any additional data and annotations, but instead progressively aligns local features to obtain improved image-level features in an unsupervised manner.

Fine-tuning: In addition to meta-learning, fine-tuning is considered as an alternative approach to address few-shot learning by recent studies[1, 4, 7, 9, 24]. [4, 9] propose that competitive results can be achieved by simply fine-tuning the final fully connected layer on a novel task. The idea of fine-tuning the network using samples from the base dataset which are similar to the samples on the target dataset is proposed in researches on transfer learning[13], which has been further improved in recent studies to adapt to few-shot learning[1]. On the other hand, [7, 24] propose that utilizing additional data to serve as negative samples and employing contrastive loss for training can further improve the performance of fine-tuning. Unlike previous methods, our work investigates on the setting of no use of the base dataset as well as any additional data. Instead of tuning the parameters of the pre-trained model, we

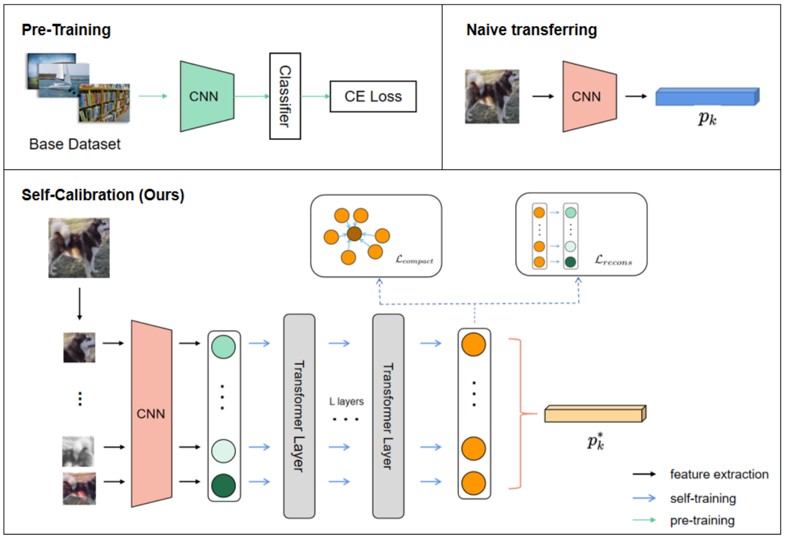

Figure 2: Given a novel 1-shot task, naive transfer learning proposed by Chen et al.[4] directly uses the support sample features extracted by a fixed feature extractor pre-trained on the base dataset as the prototypes p_k for each class, which is biased since the novel samples belong to unseen classes. Our self-calibration framework introduces an auxiliary network to calibrate the biased features. Specifically, we first produce multiple overlapped image patches to introduce detailed local information. Transformer is then used to perform progressive self-alignment on local features in an unsupervised manner and finally construct improved prototypes p_k^*.

novelly reformulate fine-tuning as calibrating the biased features of novel samples through an auxiliary network.

2.2 Unsupervised Representation Learning

Learning discriminative representations by exploiting the structural information in latent space in an unsupervised manner has been widely studied. Among them, Self-Supervised Learning based methods achieve notable progress. One common practice of these methods is to construct data augmentation based pretext tasks such as rotation prediction[15], jigsaw puzzles[30] and relative location predictions[10] to incorporate additional semantic information. Methods[3, 16, 18] using contrastive objectives recently achieve state-of-the-arts performances. They perform unsupervised alignment based on semantic similarities of features in latent space. Several prior studies[2, 6, 14, 40, 53] also explore the incorporation of self-supervised learning in few-shot scenarios. They typically follow the paradigm of constructing pretext tasks to train the feature extractor on the base dataset. The trained feature extractor are then fixed for classification on novel tasks. Specifically, [14, 40] simply construct rotation predictions on an external dataset as an additional task to train the feature extractor. IEPT[53] proposes a framework that integrates pretext task training into meta learning based episodic training. [2] makes improvements by disentangling the training process of the feature extractor into two phases for self-supervised learning and the conventional supervised learning respectively. Unlike previous works, we propose to perform unsupervised self-alignment on biased features in adaptation to novel tasks instead of using self-supervised learning to pre-train the feature extractor.

3 METHOD

3.1 Preliminaries

In few-shot learning, a task $\mathcal{T}$ is composed by support set $\mathcal{S} = \{(x_i, y_i)|y_i \in \mathcal{C}_{novel}\ , i = 1, 2, \dots, \mathcal{N} \times \mathcal{K}\}$ and query set $\mathcal{Q} = \{(x_i, y_i)|y_i \in \mathcal{C}_{novel}\ , i = 1, 2, \dots, \mathcal{M}\}$. $\mathcal{S}$ is constructed in the form of $\mathcal{N}$-way $\mathcal{K}$-shot, indicating a total of $\mathcal{N} \times \mathcal{K}$ labeled samples collected from $\mathcal{N}$ categories, each containing $\mathcal{K}$ samples. The target of few shot learning is to correctly classify the $\mathcal{M}$ samples in $\mathcal{Q}$ by the few-shot labeled $\mathcal{S}$.

The standard transfer learning has recently been proposed as an effective solution to cope with few-shot learning, which is implemented by fine-tuning the final fully-connected layer of the model. Concretely, a network consisting of a feature extractor f_θ and a classifier $c(\cdot|w_b)$ is first trained from scratch using a base dataset with a cross-entropy loss. While adapting to the novel task, the parameters of the feature extractor are fixed and the former classifier is removed. A new classifier $c(\cdot|w_n)$ is then fine-tuned on the support set instead. Typically, the new classifier $c(\cdot|w_n)$ is initialized by the prototype of each category in the support set, as follow:

$$p_k = \frac{1}{|\mathcal{S}_k|} \sum_{x \in S_k} f_\theta(x) \tag{1}$$

where $\mathcal{S}_k$ denotes the support samples of the $k - th$ class. The cosine similarity is commonly-used to calculate the classification score, which is then normalized by a softmax function to obtain the probabilities

$$P(y=k|x)=\frac{exp(cos(f_{\theta}(x),p_k)\cdot\mathcal{T})}{\sum_{j=1}^{N}exp(cos(f_{\theta}(x),p_j)\cdot\mathcal{T})} \tag{2}$$

where $\mathcal{T}$ denotes the temperature hyperparameter, and $\mathcal{N}$ denotes the total number of the classes in the support set.

3.2 Self-Calibration Framework

The overall framework of the proposed method is illustrated in Fig. 2. To alleviate the previously mentioned dilemma of fine-tuning, we reformulate fine-tuning as feature calibration conditioned on a fixed feature extractor, which is achieved by the proposed self-calibration framework. Notably, we still follow the basic setting of fine-tuning by using only cross-entropy loss to pre-train the feature extractor on the base class dataset, without any episodic training in meta-learning. Naive transferring indicates the standard fine-tuning method proposed by Chen et al.[4], which is achieved by fixing the feature extractor and merely fine-tuning the final fully connected layer. We then provide a detailed description of the proposed self-calibration framework, including the way to incorporate local features, Transformer based progressive feature calibration and the training objectives.

Incorporation of Local Features: Given an image $\mathbf{x}\in\mathcal{R}^{H\times W\times C}$, we enrich the local information by slicing it into N patches of fixed-size $\mathbf{x}=\{x_i|i=1,2,\ldots,N\}$ to introduce detailed spatial information. Similar approaches are adopted in recent researches on vision transformers[11, 29, 39], which process images by feeding a sequence of disjoint image patches to the Transformer. In our research, we make a modification on it by leveraging overlapped patches, aiming to better preserve the intrinsic correlation between different patches. To learn more robust features, we also perform color distortion based random augmentation on each patch to introduce diversity. The augmented patches are then transformed into a group of 1-D feature vectors using the feature extractor f_{θ} pre-trained on the base dataset. Thus an image is formulated into a patch-level embedding sequence $\mathbf{z}_0=\{z_i|i=1,2,\ldots,N\}\in\mathcal{R}^{N\times D}$, where $z_i=f_{\theta}(x_i)$.

Progressive Feature Calibration based on Transformer: We employ the Transformer[43] as an auxiliary network for feature calibration, which has been demonstrated to be effective in processing sequence data as well as images by previous studies[8, 11, 29]. In our study, we use it for information propagation over the enhanced local features to progressively construct improved image-level features. For a single transformer layer, it is composed of self-attention and a residual connection. For self-attention, it is first applied by projecting a sequence of features to sub-spaces which are the queries $Q\in\mathcal{R}^{N\times d}$, keys $K\in\mathcal{R}^{N\times d}$ and values $V\in\mathcal{R}^{N\times d}$ by linear transformation. Then the aggregated sequence is computed as follow:

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d}}V) \tag{3}$$

A total of L layers are adopted to progressively learn the contextual embedding. In the following we describe the objective function.

Objective Function: For image classification, the annotation is based on the image level, making it hard to determine the annotation of each patch-level embedding since it is likely to represent pure background. Thus, it is not appropriate to apply token-level supervision directly with the image-level labels. On the other hand, using only sequence-level supervision suffers from overfitting caused by severe sample sparsity. Inspired by studies in unsupervised representation learning, we optimize the Transformer to perform localised self-alignment on each patch-level embedding. Concretely, the objective function consists of a reconstruction loss and a compactness-aware loss, represented as follows:

$$\mathcal{L}_{recons}=\sum_{i=0}^{N}\left\|z_i'-z_i\right\| \tag{4}$$

where z_i represents the input patch-level embedding, and z_i' represents the output embedding of the Transformer. The role of the reconstruction loss can be understood as a constraint to limit the search space, which serves as a key term to avoid the trivial solution leading to collapse of representations. Compactness-aware loss is employed to promote instance-level compactness, which forces the patch-level embeddings derived from the same image to get closer to each other:

$$\mathcal{L}_{compact}=\sum_{i=0}^{N}\left\|z_i'-\bar{z_i'}\right\| \tag{5}$$

where $\bar{z_i'}=\frac{1}{N}\sum_{i=0}^{N}z_i'$ represents the mean feature vector. Thus, the total loss function is formulated as:

$$\mathcal{L}_{total}=\mathcal{L}_{recons}+\alpha\mathcal{L}_{compact} \tag{6}$$

where α denotes the coefficient. To validate the effectiveness of the objective function, we provide further analysis in Section 4.4. Through the information propagation over the augmented local patches, we can now produce a context-aware feature sequence for each novel sample through the Transformer. The calibrated feature of each sample is then obtained by averaging the feature sequence. Notably, we follow the strict inductive learning setting where only samples from the support set are used in the training phase. In the testing phase, we directly use the trained transformer to generate the calibrated features of both support and query samples and then perform nearest-neighbour classification as ProtoNet[38] based on the calibrated features.

4 EXPERIMENTS

4.1 Datasets

The benchmark for evaluation including two widely-used datasets which are **mini-ImageNet**[44] and **tiered-ImageNet**[32]. **mini-imageNet** is a subset of ImageNet[33]. It is composed by 100 categories, and each category contains 600 samples. Training set, validation set and test set contain 64, 16, 20 classes, respectively. **tiered-ImageNet** is a larger dataset containing 608 classes, nearly 1200 samples per class. The split of train, validation and test is 351, 97 and 160, respectively.

4.2 Implementation details

We evaluate the proposed method under strict inductive learning settings. Both 5-way 1-shot and 5-way 5-shot are evaluated. We

Table 1: Comparison results for mini-Imagenet under 5-way 1-shot and 5-way 5-shot, reported by mean accuracies with the 95% **confidence interval**

Method	Backbone	1-shot	5-shot
ProtoNet[38]	Conv4-64	49.42 ± 0.78	68.20 ± 0.66
MatchingNet[44]	Conv4-64	43.56 ± 0.84	55.31 ± 0.73
MAML[12]	Conv4-64	48.70 ± 1.75	63.11 ± 0.92
RelationNet[42]	Conv4-64	50.44 ± 0.82	65.32 ± 0.70
PN+rot[14]	Conv4-64	53.63 ± 0.43	71.70 ± 0.36
CC+rot[14]	Conv4-64	54.83 ± 0.43	71.86 ± 0.33
Chen et al.[4]	Conv4-64	48.24 ± 0.75	66.43 ± 0.63
Centroid[1]	Conv4-64	53.14 ± 1.06	71.45 ± 0.72
Neg-Cosine[26]	Conv4-64	52.84 ± 0.76	70.41 ± 0.66
Self-Calibration (Ours)	Conv4-64	**56.55 ± 0.61**	**72.52 ± 0.48**
ProtoNet[38]	ResNet-12	60.37 ± 0.83	78.02 ± 0.57
TADAM[31]	ResNet-12	58.50 ± 0.30	76.70. ± 0.30
MetaOptNet[25]	ResNet-12	62.64 ± 0.61	78.63 ± 0.46
CAN[20]	ResNet-12	63.85 ± 0.48	79.44 ± 0.34
FEAT[50]	ResNet-12	66.78 ± 0.20	82.05 ± 0.14
DSN-MR[37]	ResNet-12	64.60 ± 0.72	79.51 ± 0.50
DeepEMD[52]	ResNet-12	65.91 ± 0.82	82.41 ± 0.56
E^3BM[28]	ResNet-12	63.80 ± 0.40	80.10 ± 0.30
InfoPatch[27]	ResNet-12	67.67 ± 0.45	82.44 ± 0.31
FRN[46]	ResNet-12	66.45 ± 0.19	82.83 ± 0.13
POODLE[24]	ResNet-12	67.80	83.72
DMF[48]	ResNet-12	67.76 ± 0.46	82.71 ± 0.31
Self-Calibration(Ours)	ResNet-12	**68.62 ± 0.63**	**83.95 ± 0.41**

Table 2: Comparison results for tiered-Imagenet under 5-way 1-shot and 5-way 5-shot, reported by mean accuracies with the 95% **confidence interval**

Method	Backbone	1-shot	5-shot
ProtoNet[38]	ResNet-12	65.65 ± 0.92	83.40. ± 0.65
MetaOptNet[25]	ResNet-12	65.99 ± 0.72	81.56 ± 0.63
CAN[20]	ResNet-12	69.89 ± 0.51	84.23 ± 0.37
AM3[47]	ResNet-12	67.23 ± 0.34	78.95 ± 0.22
FEAT[50]	ResNet-12	70.80 ± 0.23	84.79 ± 0.16
DSN-MR[37]	ResNet-12	67.39 ± 0.82	82.85 ± 0.56
E^3BM[28]	ResNet-12	71.20 ± 0.40	85.30 ± 0.30
DMF[48]	ResNet-12	71.89 ± 0.52	85.96 ± 0.35
InfoPatch[27]	ResNet-12	71.51 ± 0.52	85.44 ± 0.35
POOFLE[24]	ResNet-12	70.42	85.26
RENet[21]	ResNet-12	71.61 ± 0.51	85.28 ± 0.35
SC(Ours)	ResNet-12	**72.08 ± 0.67**	**86.00 ± 0.47**

report the average accuracy and the 95% confidence interval by randomly sampling 1000 tasks from the test dataset. For the feature extractor, two widely-used backbones are included: ResNet12 and Conv4-64.

The training consists of two stages, the pre-training stage on the base dataset and the fine-tuning stage on the novel task. We train the feature extractor with a standard cross-entropy loss in the pre-training stage. Following the settings of the previous studies, we resize the input image to 84 × 84. Adam [23] is used as the optimizer, where the initial learning rate is set to 0.0001, and the batch size is set to 64. Notably, unlike meta-learning based methods, no additional episodic training stage is included in the training of the feature extractor. On the adaption of novel tasks, we fix the feature extractor and calibrate the biased features of novel samples based on the proposed self-calibration framework. We then describe the details of the self-calibration framework. For the incorporation of local features, we generate overlapped patches by simply repeating random cropping and then resizing the cropped patches to 84 × 84. Hyperparameters, including the total number of the patches N, the number of the transformer layers L and the coefficient α in the objective function, are further studied in ablation studies. For the unsupervised training process of Transformer, we follow the strict setting of inductive learning where only support samples are used. We apply a channel-wise whitening transformation to pre-process the raw features before feeding them into the model. This is implemented by subtracting the mean value and dividing by the standard deviation over the channel dimension. we use Adam [23] as the optimizer and train it for 100 iterations. The learning rate is set to 0.0001.

4.3 Comparative Studies

We compare the proposed method with several representative methods, which can be categorized into i) latest methods achieving state-of-the-arts[21, 27, 46, 48, 50, 52]. ii) fine-tuning based methods[1, 4, 24]. We perform evaluation both under RestNet-12 and Conv4-64 on mini-ImageNet to make comprehensive comparisons. The results are shown in Table 1, 2. Concretely, it can be observed that under all settings, the proposed method gains improvements over the latest researches, achieving the new state-of-arts. Compared with recent fine-tuning based methods[1, 24], the proposed method improves the performances ranging from $1\% - 3\%$. Notably, our method does not require the use of base dataset as well as any external data and is thus more data efficient. Additionally, it is also worth noting that the proposed method boost the performance of the standard transfer learning method proposed by Chen et al.[4] by 8.31% on 1-shot and 6.09% on 5-shot, demonstrating the effectiveness. Moreover, when compared with recent methods performing spatial alignment of feature maps based on meta learning, including CAN[20], DeepEMD[52] and DMF[48], our method yields higher performances with just an easily trained auxiliary network instead of relying on a complex meta-learning scheme.

4.4 Ablation Studies

In this section, we make a comprehensive study to further explore the components of the proposed self-calibration framework, including i) effect of the auxiliary network for feature calibration, ii) impact of the hyperparameters α in Equation (6) and the total number N of local patches, iii) analysis of the objective function.

Effect of the auxiliary network: Transformer is employed as the auxiliary network to learn self-aligned features in the proposed method. To explore the importance, we conduct the following ablation studies on mini-ImageNet which analyze i) the impact of different architectures of the auxiliary network and ii) the effect of the number of Transformer layers L. For the architecture of the auxiliary network, to verify the boost provided by the additional auxiliary network, we first test the performance of directly removing the auxiliary network and averaging the raw local features (raw patches). Moreover, to demonstrate the effectiveness of Transformer, we make further comparisons by replacing Transformer with a simple fully-connected network (fc), which can be interpreted as a standard auto-encoder. The results are shown in Table 3. It can be observed that the performance is degraded by 3.20% for 1-shot and 3.19% for 5-shot without the use of the auxiliary network. This demonstrates the effectiveness of using the auxiliary network for feature calibration. Moreover, when replacing Transformer with a fully-connected network, the performance declines by 1.09% for 1-shot and 0.74% for 5-shot. This reflects that using transformer to enable explicit interaction between local features is effective to learn better feature representation. We next study the effect of the number of Transformer layers L. The results are shown in Table 4. It can be observed that simply using a single Transformer layer for feature calibration gains notable improvements, and 2 is the optimal setting of L. This demonstrates the computational efficiency of the proposed method.

Table 3: Ablation studies for the auxiliary network in the proposed method on Mini-ImageNet

	mini-ImageNet	
Method	1-shot	5-shot
raw patches	65.42±0.65	80.76±0.41
fc	67.53±0.61	83.21±0.41
Ours	**68.62±0.63**	**83.95±0.41**

Table 4: Ablation studies for the number of Transformer layers L on mini-ImageNet

Parameter L	1	2	3	4
1-shot	67.98	**68.62**	68.38	68.33
5-shot	83.12	**83.95**	83.65	83.57

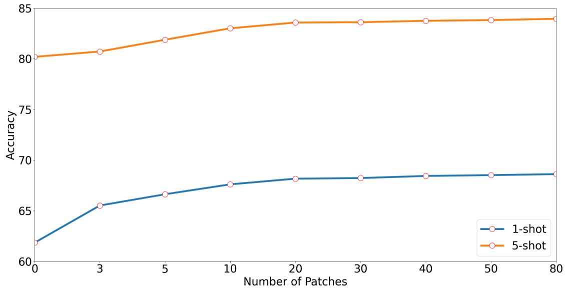

Figure 3: Ablation studies for the total number N of local patches on mini-ImageNet

Table 5: Ablation studies for the coefficient of α on mini-ImageNet

Parameter α	1	0.1	0.01	0.001
1-shot	68.20	68.33	**68.62**	68.12
5-shot	83.31	83.55	**83.95**	83.22

Hyperparameter Analysis: We also study the impact of the total number N of the local patches sliced from a novel sample and the effect of the hyperparameter α in Equation (6), which indicates the coefficient of the loss function. The results are reported in Fig. 3 and Table 5. We find that the accuracy grows as the number of the augmented local patches increases. This demonstrates that improved feature representations are learned by leveraging the detailed local information for self-alignment, validating the effectiveness of the proposed method. On the other hand, it is also notable that competitive results can be achieved with only a very small number of local patches. For the coefficient α, we find that when $\alpha = 0.01$ the proposed method achieves the best performance. Setting a different value only lead to minor performance degradation, demonstrating the robustness of the self-calibration framework.

Unsupervised or Supervised: The proposed self-calibration framework leverages structural information in latent space to perform feature calibration in a self-aligned manner. It is trained in a completely unsupervised manner. In this section, we investigate how performance changes if we employ a supervised loss function for training. It is achieved by directly assigning the image-level label to each patch-level embedding as a non-fine grained annotation. We make comparisons with two commonly-used objective

Table 6: Ablation studies for comparisons with different objective functions

	mini-ImageNet	
Loss function	1-shot	5-shot
CE Loss	67.38±0.62	83.81±0.38
Supcon Loss[22]	67.89±0.63	83.65±0.40
Ours	**68.62±0.63**	**83.95±0.41**

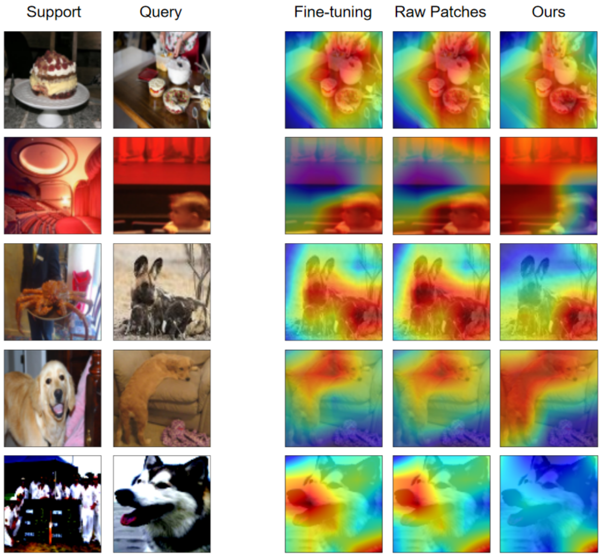

Figure 4: Visualization of class activation maps for 5-way 1-shot tasks on mini-ImageNet

functions for supervised training, including i) cross-entropy loss, ii) supervised contrastive loss[22]. The results are shown in Table 6. It can be observed that in a 1-shot scenario, the proposed self-supervised loss yields a higher performance. On the other hand, in a 5-shot scenario, the three objective functions gain comparable results. This is possibly because in a extreme few-shot scenario, more semantic and structural information can be leveraged from data itself rather than labels. As the sample size grows, the role of labels is becoming increasingly significant.

4.5 Visualization Analysis

In this section, we make a further qualitative evaluation of the proposed self-calibration framework based on the visualization of the class activation maps (CAM)[54]. Comparisons are carried out under the 5-way 1-shot setting among the following methods, including i) the standard transfer learning method proposed by Chen et al.[4] (fine-tuning) and (ii) simply averaging the raw features of local patches (raw patches). The results are shown in Fig. 4, where warmer colors represent higher responses. It demonstrates that with the unsupervised alignment of the auxiliary Transformer network, the features more precisely locate discriminative regions of an input image. For the second row of images representing theaters, it can be observed that without the use of the self-calibration framework, the discriminative region is exactly wrongly predicted, whereas our method avoids this mistake. Moreover, in the last row, we extensively select the query sample that belong to a different category from the support sample. It can be observed that the features clearly reject the confusing regions with the self-calibration framework. This reflects the locally aligned features have greater robustness, demonstrating the effectiveness of the proposed method.

5 CONCLUSION

In this work, we target improving the performance of fine-tuning based methods in few-shot learning without using any additional dataset as well as the base dataset. We point out a dilemma of fine-tuning and novelly reformulate fine-tuning into feature calibration through an auxiliary network. To address this issue, we propose a self-calibration framework that uses Transformer trained in a self-supervised manner to progressively incorporate contextual information in local features and ultimately construct improved image-level features. Extensive experiments demonstrate the proposed method achieves the new state-of-the-arts, and greatly outperforms the conventional fine-tuning method.

REFERENCES

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. 2020. Associative Alignment for Few-Shot Image Classification. In *ECCV*.
[2] Yuexuan An, Hui Xue, Xingyu Zhao, and Lu Zhang. 2021. Conditional Self-Supervised Learning for Few-Shot Classification. In *IJCAI*. 2140–2146.
[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*.
[4] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification. In *ICLR*.
[5] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. 2021. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *ICCV*.
[6] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. 2021. Pareto Self-Supervised Training for Few-Shot Learning. In *CVPR*. 13663–13672.
[7] Rajshekhar Das, Yu-Xiong Wang, and José M. F. Moura. 2021. On the Importance of Distractors for Few-Shot Classification. In *ICCV*.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*.
[9] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. 2020. A Baseline for Few-Shot Image Classification. In *ICLR*.
[10] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *ICCV 2015*.
[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR 2021*.
[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML (Proceedings of Machine Learning Research, Vol. 70)*. 1126–1135.
[13] Weifeng Ge and Yizhou Yu. 2017. Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-Tuning. In *CVPR*.
[14] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. 2019. Boosting Few-Shot Visual Learning With Self-Supervision. In *ICCV*. 8058–8067.
[15] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR 2018*.
[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *NeurIPS*.
[17] Bharath Hariharan and Ross Girshick. 2017. Low-Shot Visual Recognition by Shrinking and Hallucinating Features. In *ICCV*.
[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR 2020*.

[19] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. TransReID: Transformer-Based Object Re-Identification. In *ICCV*.

[20] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross Attention Network for Few-shot Classification. In *NeurIPS*.

[21] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. 2021. Relational Embedding for Few-Shot Classification. In *ICCV*. 8822–8833.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *NeurIPS*.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[24] Duong Hoang Le, Khoi Duc Nguyen, Khoi Nguyen, Quoc-Huy Tran, Rang Nguyen, and Binh-Son Hua. 2021. POODLE: Improving Few-shot Learning via Penalizing Out-of-Distribution Samples. In *Advances in Neural Information Processing Systems 2021*.

[25] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. 2019. Meta-Learning With Differentiable Convex Optimization. In *CVPR*. 10657–10665.

[26] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. 2020. Negative Margin Matters: Understanding Margin in Few-Shot Classification. In *ECCV 2020*.

[27] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. 2021. Learning a Few-shot Embedding Model with Contrastive Learning. In *AAAI*. 8635–8643.

[28] Yaoyao Liu, Bernt Schiele, and Qianru Sun. 2020. An Ensemble of Epoch-Wise Empirical Bayes for Few-Shot Learning. In *ECCV 2020*.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *ICCV*.

[30] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *ECCV 2016*.

[31] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS 2018*.

[32] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. 2018. Meta-Learning for Semi-Supervised Few-Shot Classification. In *ICLR*.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252.

[34] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. 2019. Meta-Learning with Latent Embedding Optimization. In *ICLR*.

[35] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NeurIPS*.

[36] Shuai Shao, Lei Xing, Yan Wang, Rui Xu, Chunyan Zhao, Yanjiang Wang, and Baodi Liu. 2021. MHFC: Multi-Head Feature Collaboration for Few-Shot Learning. In *MM '21: ACM Multimedia Conference, Virtual Event, 2021*.

[37] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. 2020. Adaptive Subspaces for Few-Shot Learning. In *CVPR 2020*.

[38] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. In *NeurIPS*. 4077–4087.

[39] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. Segmenter: Transformer for Semantic Segmentation. In *ICCV*.

[40] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. 2020. When Does Self-supervision Improve Few-Shot Learning?. In *ECCV*.

[41] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV 2018*.

[42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *CVPR*. 1199–1208.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008.

[44] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching Networks for One Shot Learning. In *NeurIPS*. 3630–3638.

[45] Zeyuan Wang, Yifan Zhao, Jia Li, and Yonghong Tian. 2020. Cooperative Bi-path Metric for Few-shot Learning. In *MM '20: The 28th ACM International Conference on Multimedia, 2020*.

[46] Davis Wertheimer, Luming Tang, and Bharath Hariharan. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *CVPR*.

[47] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro O. Pinheiro. 2019. Adaptive Cross-Modal Few-shot Learning. In *NeurIPS*. 4848–4858.

[48] Chengming Xu, Yanwei Fu, Chen Liu, Chengjie Wang, Jilin Li, Feiyue Huang, Li Zhang, and Xiangyang Xue. 2021. Learning Dynamic Alignment via Meta-Filter for Few-Shot Learning. In *CVPR*. 5182–5191.

[49] Shuo Yang, Lu Liu, and Min Xu. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *ICLR*.

[50] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *CVPR*. 8805–8814.

[51] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. 2021. Prototype Completion With Primitive Knowledge for Few-Shot Learning. In *CVPR*.

[52] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *CVPR*. 12200–12210.

[53] Manli Zhang, Jianhong Zhang, Zhiwu Lu, Tao Xiang, Mingyu Ding, and Songfang Huang. 2021. IEPT: Instance-Level and Episode-Level Pretext Tasks for Few-Shot Learning. In *ICLR*.

[54] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*.