ESPT: A Self-Supervised Episodic Spatial Pretext Task for Improving Few-Shot Learning

Yi Rong[1,2,3,4], **Xiongbo Lu**[1], **Zhaoyang Sun**[1], **Yaxiong Chen**[1,2], **Shengwu Xiong**[1,2,3,4*]

[1]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China
[2]Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China
[3]Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China
[4]Shanghai Artificial Intelligence Laboratory, Shanghai 200240, China
{yrong, luxiongbo, zhaoyangsun, chenyaxiong, xiongsw}@whut.edu.cn

Abstract

Self-supervised learning (SSL) techniques have recently been integrated into the few-shot learning (FSL) framework and have shown promising results in improving the few-shot image classification performance. However, existing SSL approaches used in FSL typically seek the supervision signals from the global embedding of every single image. Therefore, during the episodic training of FSL, these methods cannot capture and fully utilize the local visual information in image samples and the data structure information of the whole episode, which are beneficial to FSL. To this end, we propose to augment the few-shot learning objective with a novel self-supervised Episodic Spatial Pretext Task (ESPT). Specifically, for each few-shot episode, we generate its corresponding transformed episode by applying a random geometric transformation to all the images in it. Based on these, our ESPT objective is defined as maximizing the local spatial relationship consistency between the original episode and the transformed one. With this definition, the ESPT-augmented FSL objective promotes learning more transferable feature representations that capture the local spatial features of different images and their inter-relational structural information in each input episode, thus enabling the model to generalize better to new categories with only a few samples. Extensive experiments indicate that our ESPT method achieves new state-of-the-art performance for few-shot image classification on three mainstay benchmark datasets. The source code will be available at: https://github.com/Whut-YiRong/ESPT.

Introduction

Deep learning (LeCun, Bengio, and Hinton 2015) based approaches have achieved impressive results in various image classification tasks, such as face recognition (Meng et al. 2021), object recognition (Krizhevsky, Sutskever, and Hinton 2012) and person re-identification (Li et al. 2021). However, the success of these methods relies heavily on the availability of massive training data with reliable annotations. Unfortunately, in many practical image classification applications, collecting and manually labeling sufficient training samples are not only expensive and time-consuming, but also may not be feasible for some rare object categories due to the scarcity of data. Training deep neural networks in such low-data regimes will inevitably lead to overfitting problems, which can greatly reduce the generalization ability of the learned models and finally limit their applicability in real-world scenarios. On the contrary, humans can rely on past experience to accurately identify new objects by only observing a small number of reference samples. By emulating such ability of human intelligence, few-shot learning has recently shown promising results in learning novel concepts from a few training images, and thus has become an effective approach to address the data scarcity problem in deep learning fields.

Few-shot learning (FSL) (Li, Fergus, and Perona 2006; Lake, Salakhutdinov, and Tenenbaum 2015; Koch et al. 2015; Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Finn, Abbeel, and Levine 2017) aims to learn transferable prior knowledge from a set of "base classes" with sufficient training samples, and then utilize such knowledge to recognize unseen "novel classes" that only have a few reference images per class. For this purpose, the episodic learning strategy is often employed, which samples a series of few-shot episodes from the base classes in the training phase (Snell, Swersky, and Zemel 2017; Finn, Abbeel, and Levine 2017). Each episode consists of a small "support set" and a relatively large "query set" that simulate the setup of the target few-shot classification tasks encountered during the evaluation procedure. Then with these episodes, a generic deep model or a common optimization method (act as the prior knowledge) is typically learned for fast adaptation to the testing few-shot classification tasks with unseen categories. One of the main problems with most FSL methods is that they usually only optimize a single categorical objective (e.g. cross-entropy loss). As a result, the learned models will only capture the necessary knowledge for the classification tasks over the training classes. Therefore, these models tend to have excessive discriminability for the base classes but limited transferability to unseen categories, which will finally lead to a decrease of few-shot classification performance on the novel classes.

To alleviate this problem, several recent studies (Gidaris et al. 2019; Rajasegaran et al. 2021; Lee, Hwang, and Shin 2020; Su, Maji, and Hariharan 2020; Liu et al. 2021) propose to integrate self-supervised learning (SSL) techniques

*Corresponding Author

into the FSL framework. They augment traditional FSL optimization objective with an annotation-free pretext task (e.g. rotation prediction or jigsaw puzzle task), which acts as an auxiliary regularization term and is jointly optimized with the classification loss. Since solving these pretext tasks depends only on the visual information present in images, the learned feature representations will be more inclined to capture low-level visual patterns in image samples, and thus will be more generalizable that can transfer well to the novel classes (Islam et al. 2021). However, existing SSL approaches typically process each input sample individually and seek the supervision signals from every single image. So the pretext task objective of these SSL methods cannot fully exploit the interrelationships of multiple image samples within a few-shot episode, which will lead to the loss of data structure information of the whole episode that may be beneficial to the FSL training process. Moreover, most SSL pretext tasks are constructed based on global image embeddings and tend to ignore the local spatial image features that contain richer and more transferable low-level visual information, which may limit their effectiveness in improving the transferability of FSL models.

In order to address the above problems, in this paper, we propose a novel self-supervised Episodic Spatial Pretext Task (ESPT) for few-shot image classification, where the supervision information is derived from the *relationships between local spatial features of multiple image samples* in each learning episode. Specifically, given a few-shot episode, we first apply a random geometric transformation to all the samples in it to generate its corresponding transformed episode. These two episodes are then fed into a deep network with two identical branches, one for the original episode and the other for the transformed one. In addition to outputting label predictions for the query samples, each branch calculates the relationships between local features extracted at different spatial locations of multiple images in the input episode. After that, our ESPT objective is defined as maximizing the local spatial relationship consistency between the same images in the original and transformed episodes. The intuition behind such definition is that the transformation imposed on the images should not significantly change the local spatial relationships among them. With this objective, the introduced self-supervised pretext task is able to fully exploit the local spatial features of different images and their inter-relational structural information in the input episode. Finally, by jointly optimizing the few-shot classification loss and the proposed ESPT objective, the resulting model will benefit from the above self-supervision information to learn the visual representations with stronger transferability that can be well adapted to novel classes with few training examples, which will effectively improve its few-shot classification performance.

The main contributions of our method can be summarized as: (1) To the best of our knowledge, the proposed ESPT method is the first attempt to augment traditional few-shot image classification model using the self-supervision information obtained from the local spatial relationships among multiple image samples in each learning episode. (2) The proposed ESPT method does not introduce any additional network structures and extra trainable parameters. Therefore it does not increase the model complexity and the risk of data overfitting, which is especially important for solving few-shot learning problems with limited training samples. (3) Extensive experiments on the miniImageNet, tieredImageNet and CUB-200-2011 datasets show that our method outperforms several benchmark approaches and achieves the state-of-the-art few-shot image classification performance.

Related Work

Few-shot Learning

Existing few-shot classification approaches can be roughly divided into three categories: **(1) Metric/Embedding-based** methods project input samples into a discriminative embedding space and then calculate the distance between them and the categories to be classified. MatchingNet (Vinyals et al. 2016) and ProtoNet (Snell, Swersky, and Zemel 2017) learn an embedding space where a predefined metric (e.g., Euclidean distance and cosine similarity) can be used as the distance measurement. However, since a common embedding space cannot be equally effective for all few-shot classification tasks, MetaOptNet (Lee et al. 2019), CTM (Li et al. 2019) and FEAT (Ye et al. 2020) propose to learn task-specific embeddings (or classifiers) to capture the most discriminative information for each target task. Besides, there are also some other FSL methods that design learnable distance metrics via nonlinear relation modules (Sung et al. 2018; Doersch, Gupta, and Zisserman 2020; Kang et al. 2021), ridge/logistic regression (Chen et al. 2019; Bertinetto et al. 2019; Tian et al. 2020), and graph neural networks (Satorras and Estrach 2018; Tang et al. 2021). **(2) Optimization-based** methods typically meta-learn an optimizer or a model that can quickly adapt to the unseen novel classes. MAML (Finn, Abbeel, and Levine 2017) and some of its variant methods (Rusu et al. 2019; Jamal and Qi 2019; Hu et al. 2020; Liu, Schiele, and Sun 2020) attempt to learn a good parameter initialization such that the adapted model of each input task can be rapidly obtained by performing only a few stochastic gradient descent (SGD) steps on support samples. In (Ravi and Larochelle 2017; Li et al. 2017), LSTM based meta-learners are trained to generate update rules to replace the SGD optimizer for model parameter training. **(3) Generation-based** methods aim to increase the number of training samples through data generation and augmentation. (Hariharan and Girshick 2017) and (Wang et al. 2018) propose a hallucinator module that maps real training images and randomly sampled noise to hallucinated samples. The generated samples may not be realistic, but are useful to refine the decision boundary of the learned FSL model. In (Zhang et al. 2018; Schwartz et al. 2018; Park et al. 2020; Yang, Liu, and Xu 2021), the intra-class variance and data distribution of the base classes are transferred to the examples in novel classes to produce augmented data.

SSL Augmented Few-shot Learning

To facilitate the transferability of learned feature representations, several recent works incorporate existing SSL techniques into the FSL framework by introducing auxiliary pre-

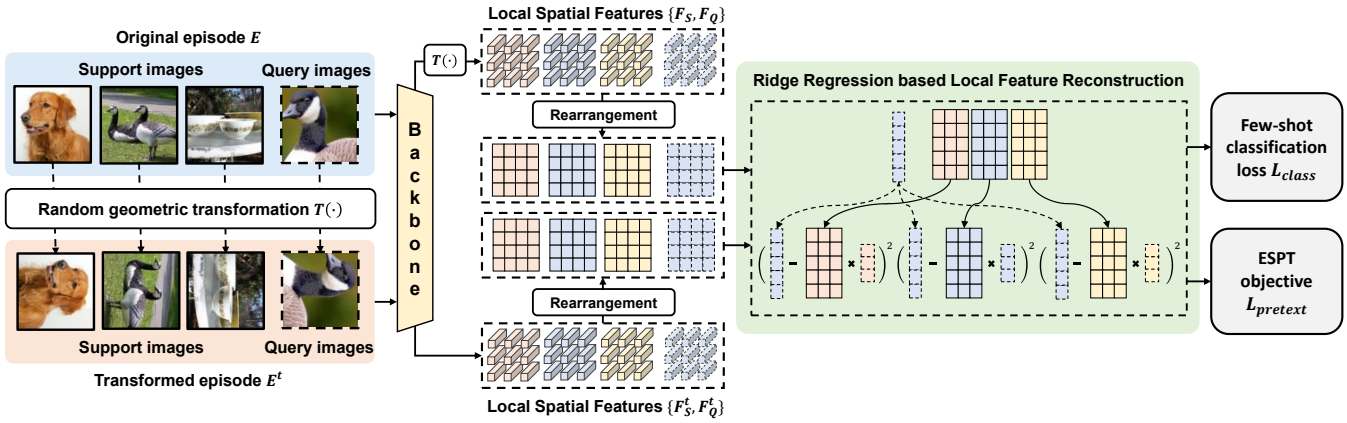


Figure 1: Illustration of our ESPT method in the 3-way 1-shot setting. Given an input episode E, we first generate a transformed episode E^t from it by using a random geometric transformation $T(\cdot)$. Then we feed the image samples within these two episodes into a two-branch network to extract their local spatial features $\{F_S, F_Q\}$ and $\{F_S^t, F_Q^t\}$. After that, we establish the local spatial relationships between the support and query images from each of the two episodes by solving a ridge regression based feature reconstruction problem. Finally, for model training, the few-shot classification loss L_{class} and the proposed ESPT objective $L_{pretext}$ are defined based on the reconstruction residuals and reconstruction coefficients, respectively.

text task objectives, such as predicting image rotation angles (Rajasegaran et al. 2021; Lee, Hwang, and Shin 2020), solving image patch jigsaw puzzles (Su, Maji, and Hariharan 2020), estimating image patch relative locations (Gidaris et al. 2019), and completing contrastive learning tasks (Rizve et al. 2021; Islam et al. 2021; Ouali, Hudelot, and Tami 2021). However, since these pretext tasks are not specially designed for the few-shot classification problem, they may not be able to capture and utilize the episode-specific information during the episodic training process of FSL models. To this end, IEPT (Zhang et al. 2021) designs an episode-level pretext task based on the label prediction probability of each query sample. InforPatch (Liu et al. 2021) proposes a new contrastive learning scheme that defines the positive and negative sample pairs for each anchor query image from the support samples. Different from these existing approaches, our ESPT method seeks the self-supervision signals from the local spatial relationships among multiple images in each episode. It therefore can capture and exploit the low-level visual features of image samples and the data structure information of the whole episode to learn more generalizable feature representations.

Method

Preliminary

In this work, we follow a standard setup of the few-shot classification problem, which is typically described as an n-way k-shot task. That is, in each task, there are n categories to be classified and each category contains k training samples. With this setting, few-shot learning aims to train a deep model on the data of the base classes C_b, with the hope that the learned model can generalize well to the few-shot classification tasks sampled from the novel classes C_n that are not overlapped with C_b, i.e., $C_b \cap C_n = \emptyset$. For this purpose, the episodic learning strategy is employed, which constructs a series of few-shot episodes for model training. Each episode $E = \{S, Q\}$ contains a support set $S = \{(x_s, y_s)|y_s \in C_e, s = 1, ..., n \times k\}$ and a query set $Q = \{(x_q, y_q)|y_q \in C_e, q = 1, ..., n \times l\}$ that are drawn from the same label space to simulate the n-way k-shot setting of the testing classification tasks. Here, S and Q are totally disjoint, satisfying $S \cap Q = \emptyset$, and C_e is a set of n classes that are randomly sampled from C_b. At each training iteration, the model first adapts to the input episode by performing an update using its support set S. Then the performance of the resulting model is evaluated on the corresponding query set Q to produce an optimization loss that is used to update the global model parameters for all episodes.

The Proposed Two-Branch Network

The main purpose of our ESPT method is to augment FSL by integrating an auxiliary self-supervised pretext task, so that the learned model can benefit from class-agnostic self-supervision information to learn more transferable feature representations. To achieve this, as shown in Figure 1, we construct a deep network with two identical branches that share the same feature extractor f_θ, which is implemented as a convolutional neural network (e.g., ResNet, WRN). For each input episode $E = \{S, Q\}$, we first apply a random geometric transformation $T(\cdot)$ to all the samples in it to generate its corresponding transformed episode $E^t = \{S^t, Q^t\}$, where $S^t = \{(T(x_s), y_s)|y_s \in C_e, s = 1, ..., n \times k\}$ and $Q^t = \{(T(x_q), y_q)|y_q \in C_e, q = 1, ..., n \times l\}$. Then, we feed E and E^t into the two branches of our model respectively, and obtain the convolutional feature maps of their image samples through the feature extractor f_θ as follows:

$$F_S = \{(f_s, y_s)\} = \{(T(f_\theta(x_s)), y_s)\}, \quad (1)$$

$$F_Q = \{(f_q, y_q)\} = \{(T(f_\theta(x_q)), y_q)\}, \quad (2)$$

$$F_S^t = \{(f_s^t, y_s)\} = \{(f_\theta(T(x_s)), y_s)\}, \quad (3)$$

$$F_Q^t = \{(f_q^t, y_q)\} = \{(f_\theta(T(x_q)), y_q)\}. \quad (4)$$

Here we apply the same geometric transformation $T(\cdot)$ to the feature map of each sample in the original episode E (see equations (1), (2)), ensuring that the spatial locations in the feature maps $T(f_\theta(x))$ and $f_\theta(T(x))$ of the same image are aligned. The size of each image feature map is $h \times w \times d$, i.e., $\{f_s, f_q, f_s^t, f_q^t\} \in \mathcal{R}^{h\times w\times d}$, where h and w denote its height and width respectively, and d is the dimension of its local feature vector at each spatial location. With these local image features, the objective function of our ESPT method is defined in the following sections.

Episodic Spatial Pretext Task

We construct our episodic spatial pretext task based on the local spatial relationships among multiple image samples in each training episode to effectively capture and exploit its data structure information. To establish such relationships for an input episode E, we propose to reconstruct the local feature vectors of the query images by using the spatial features of the support samples. Concretely, for each class $c \in C_e$, we first rearrange the feature maps of its k support samples ($y_s = c$) into a single spatial feature matrix $X_c \in \mathcal{R}^{khw\times d}$. Then, each local feature vector $(f_q)_{ij} \in \mathcal{R}^d$ of the query image x_q is reconstructed by solving the following linear least-squares problem:

$$\min_{(w_q)_{ij}^c} \left\|(f_q)_{ij} - X_c^T (w_q)_{ij}^c\right\|_2^2 + \lambda \left\|(w_q)_{ij}^c\right\|_2^2, \quad (5)$$

where $\|\cdot\|_2$ denotes the l_2 norm of a vector and $(\cdot)^T$ is the transpose operator of a matrix. $\lambda > 0$ is a trade-off parameter that balances the importance of the regularization term. Since the size of spatial feature matrix X_c changes depending on the number of support samples k , the feature map size hw and the local feature vector dimensions d, we also need to adaptively adjust the value of λ according these variables to guarantee the effectiveness of the reconstruction. Therefore, the parameter λ can be formulated as:

$$\lambda = \frac{khw}{d}\bar{\lambda}, \quad (6)$$

we can control λ by setting different value of $\bar{\lambda}$. $(f_q)_{ij} (i = 1, ..., h, j = 1, ..., w)$ is the local feature vector at the (i, j)-th spatial location of f_q, and $(w_q)_{ij}^c \in \mathcal{R}^{khw}$ denotes its reconstruction coefficients corresponding to X_c. Therefore, $(w_q)_{ij}^c$ can be used to represent the local spatial relationships between $(f_q)_{ij}$ and the support images of the c-th class. The above optimization problem in equation (5) is also known as the ridge regression problem, which has a differentiable closed-form solution that can be rapidly calculated as:

$$(w_q)_{ij}^c = (X_c X_c^T + \lambda I)^{-1} X_c (f_q)_{ij}. \quad (7)$$

For the transformed episode E^t, by applying the same operations, we can also get the reconstruction coefficients $(w_q^t)_{ij}^c$ for the local feature vector $(f_q^t)_{ij}$ of the same query image x_q. With these coefficients, we promote the spatial relationship consistency between each query image in the origin episode and the transformed episode by minimizing the distance between $(w_q)_{ij}^c$ and $(w_q^t)_{ij}^c$ for all classes at different spatial locations in the input image, which can be formulated as the following consistency loss:

$$L_{cons}^q = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \sum_{c\in C_e} dis(sg[(w_q)_{ij}^c], (w_q^t)_{ij}^c), \quad (8)$$

where $dis(\cdot, \cdot)$ denotes a distance function (the cosine distance is used in our implementation). We perform a stop gradient operation $sg[\cdot]$ on $(w_q)_{ij}^c$ to prevent its discriminability from being affected. By averaging the above loss over all query images, the self-supervised objective function of the propose episodic spatial pretext task is defined as:

$$L_{pretext} = \frac{1}{n \times l} \sum_{q=1}^{n\times l} L_{cons}^q. \quad (9)$$

It can be seen that calculating and optimizing the above objective do not require any additional network structures and extra trainable parameters. This can make the proposed ESPT method more flexible and more suitable for solving few-shot classification problems, since it does not increase the model complexity and the risk of data overfitting.

Algorithm 1: Training process of our ESPT method

Input: The training set of the base classes C_b, the transformation set U, the hyperparameters $\bar{\lambda}$ and α
Output: The learned feature extractor f_θ

```
Initialize all learnable parameters Φ = {γ, θ}
while Maximum number of iterations is not reached do
    Randomly sample an episode E from the training set and a geometric transformation T(·) from U
    Generate the transformed episode E^t by applying T(·) to the image samples in E
    Calculate the episodic spatial pretext task objective L_pretext using equation (9)
    Calculate the few-shot classification loss L_class using equation (12)
    Calculate the total loss L_total = L_class + αL_pretext
    Update the parameters Φ base on ∇_Φ L_total
end while
return The updated Φ
```

Few-Shot Classification Loss

For few-shot image classification, the main idea is that the residual of reconstructing the query images of the c-th class with support samples of the same class ($y_s = c$) should be much smaller than that using support images from other classes ($y_s \neq c$). Therefore, we utilize the reconstruction error $\|(f_q)_{ij} - X_c^T (w_q)_{ij}^c\|_2^2$ (in equation (5)) over all feature map locations of each query image x_q to compute its prediction probability over all classes as follows:

$$\langle f_q, c\rangle = -\frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} \|(f_q)_{ij} - X_c^T (w_q)_{ij}^c\|_2^2, \quad (10)$$

$$p(y = c|x_q) = \frac{\exp(\gamma\langle f_q, c\rangle)}{\sum_{c_i\in C_e} \exp(\gamma\langle f_q, c_i\rangle)}, \quad (11)$$

where γ is a trainable temperature parameter that controls the scale of probability logits. Based on these definitions, the few-shot classification loss for all the query samples in the input episode can be formulated as a cross-entropy loss:

$$L_{class} = -\frac{1}{n \times l} \sum_{q=1}^{n \times l} [\log p(y = y_q | x_q)]. \quad (12)$$

Total Loss and Full Algorithm

By linearly combining the episodic spatial pretext task objective and the few-shot classification loss in equations (9) and (12), the total loss of the proposed ESPT method can be defined as follow:

$$L_{total} = L_{class} + \alpha L_{pretext}, \quad (13)$$

where $\alpha > 0$ is a trade-off parameter for weighing the two objectives. The detailed training process of our ESPT method is summarized in Algorithm 1. After training, since the two branches of our model share the same parameters, we drop the branch for the transformed episodes and use the remaining one in the learned model for inference. During the evaluation phase, given an n-way k-shot image classification task, we calculate the prediction probability distribution for each query sample using equation (11), and classify it into the category with the highest probability value.

Experiments

Datasets

We verify the effectiveness of our ESPT method on three widely-used datasets for few-shot image classification, including miniImageNet (Vinyals et al. 2016), tieredImageNet (Ren et al. 2018) and CUB-200-2011 (Wah et al. 2011).

miniImageNet consists of 100 object classes randomly selected from ILSVRC-12 dataset, with each class containing 600 image samples. Using the class split in (Ravi and Larochelle 2017), we take 64, 16 and 20 classes to construct the training set, validation set and testing set, respectively.

tieredImageNet is a much larger subset derived from ILSVRC-12 dataset. It consists of over 779k images from 608 classes, where each class is drawn from one of 34 super-categories according to the ImageNet category hierarchy. We follow the settings in (Zhang et al. 2020; Kang et al. 2021) by dividing this dataset into 20/351, 6/97 and 8/160 super-categories/classes for training, validation and testing, respectively. Performing few-shot image classification under this setting will be more challenging and typically requires stronger generalization ability of the classification model, since the training and testing classes are sampled from different super-categories.

CUB-200-2011 (CUB) is a fine-grained image dataset of different birds. It contains 11,788 image samples of 200 bird species, which are partitioned into 100 training classes, 50 validation classes and 50 testing classes, as in (Chen et al. 2019; Wertheimer, Tang, and Hariharan 2021).

All the images used for our experiments in the above three datasets are manually cropped and resized to 84 $\times$ 84 pixels before input into the feature extractor.

Implementation Details

Backbone For a fair comparison, the ResNet-12 network (He et al. 2016) with the same architecture as previous works (Bertinetto et al. 2019; Ye et al. 2020) is used as the feature extractor f_θ of our model. This ResNet-12 network consists of 4 residual blocks, each containing 3 convolutional layers with the kernel size of 3 $\times$ 3. The number of filters in each convolutional layer of the four blocks is set to 64, 160, 320, and 640, respectively. Each residual block is followed by a 2 $\times$ 2 max-pooling layer for down-sampling the feature maps. We remove the global average-pooling layer on top of the network to preserve the local spatial features. Therefore, for each input image with 84 $\times$ 84 pixels, the feature extractor f_θ will output a feature map with the size of 640 $\times$ 5 $\times$ 5, i.e., $h = 5$, $w = 5$ and $d = 640$.

Image Transformation During the training of our ESPT method, for each input episode, we randomly select a geometric transformation $T(\cdot)$ from a predefined image transformation set U. In this work, we define U as a collection of 2D rotation transformations with one or more rotation degrees, i.e., $U \subset \{90°, 180°, 270°\}$, since the spatial location alignment between feature maps $T(f_\theta(x))$ and $f_\theta(T(x))$ (see equations (1),(3) and (2),(4)) can be easily achieved under this definition. We will discuss the effect of U with different rotation transformations in subsequent ablation studies. It is worth noting that other image transformations (e.g., horizontal/vertical flipping, scaling and color jittering, etc.) can also be used in the proposed ESPT method, only if they do not change the alignment of feature map spatial locations between the original input images and the transformed ones.

Training Details In order to stabilize the training process, same as (Wertheimer, Tang, and Hariharan 2021), we rescale the extracted ResNet-12 local spatial image features by a factor of $1/\sqrt{640}$, which is algebraically equivalent to the prediction logit normalization technique used in existing approaches (Snell, Swersky, and Zemel 2017; Simon et al. 2020). The same **stochastic gradient descent (SGD) optimizer with Nesterov momentum of 0.9 and weight decay of 5e-4** is utilized for model training on the three datasets, except that the initial learning rate is set to different values. For **miniImageNet** dataset, we first pre-train our models for 350 epochs with an initial learning rate of 0.1 and a mini-batch size of 128. The learning rate is decayed by multiplying 0.1 after 200 and 300 epochs. In the subsequent episodic learning phase, the models are fine-tuned for 400 epochs with each epoch containing 100 few-shot episodes. We initialize the learning rate as 0.001 and cut it by a factor of 10 at 200 and 300 epochs. For **tieredImageNet** dataset, the pre-training process runs for 90 epochs, also using the initial learning rate of 0.1 and the mini-batch size of 128. Such initial learning rate is decreased after every 30 epochs by a factor of 10. Similar to miniImageNet, the episodic learning process runs for 450 epochs with 100 few-shot episodes in each epoch. The learning rate is initialized as 0.001 and reduced by multiplying 0.1 at 50 and 250 epochs. For **CUB-200-2011** dataset, we episodic-train our models from scratch for 800 epochs, where each epoch consists of 100 few-shot

Model	Backbone	miniImageNet 5-way 1-shot	miniImageNet 5-way 5-shot	tieredImageNet 5-way 1-shot	tieredImageNet 5-way 5-shot
FSL methods w/o SSL					
MatchingNet (Vinyals et al. 2016)	ResNet-12	65.64±0.20	78.72±0.15	68.50±0.92	80.60±0.71
ProtoNet (Snell, Swersky, and Zemel 2017)	ResNet-12	62.39±0.21	80.53±0.14	68.23±0.23	84.03±0.16
MetaOptNet (Lee et al. 2019)	ResNet-12	62.64±0.61	78.63±0.46	65.99±0.72	81.56±0.53
Baseline (Chen et al. 2019)	ResNet-18	51.75±0.80	74.24±0.63	-	-
Baseline++ (Chen et al. 2019)	ResNet-18	51.87±0.77	75.68±0.63	-	-
Neg-Cosine (Liu et al. 2020)	ResNet-12	63.85±0.81	81.57±0.56	-	-
E^3BM (Liu, Schiele, and Sun 2020)	ResNet-12	64.09±0.37	80.29±0.25	71.34±0.41	85.82±0.29
FEAT (Ye et al. 2020)	ResNet-12	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16
RFS-simple (Tian et al. 2020)	ResNet-12	62.02±0.63	79.64±0.44	69.74±0.72	84.41±0.55
RFS-distill (Tian et al. 2020)	ResNet-12	64.82±0.60	82.14±0.43	71.52±0.69	86.03±0.49
Meta-Baseline (Chen et al. 2021)	ResNet-12	63.17±0.23	79.26±0.17	68.62±0.27	83.29±0.18
DeepEMD† (Zhang et al. 2020)	ResNet-12	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
FRN† (Wertheimer, Tang, and Hariharan 2021)	ResNet-12	66.45±0.19	82.83±0.13	72.06±0.22	86.89±0.14
RENet† (Kang et al. 2021)	ResNet-12	67.60±0.44	82.58±0.30	71.61±0.51	85.28±0.35
TPMN† (Wu et al. 2021)	ResNet-12	67.64±0.63	83.44±0.43	72.24±0.70	86.55±0.63
SSL augmented FSL methods					
CC+rot (Gidaris et al. 2019)	WRN-28-10	62.93±0.45	79.87±0.33	70.53±0.51	84.98±0.36
SLA (Lee, Hwang, and Shin 2020)	ResNet-12	62.93±0.63	79.63±0.47	-	-
SCL (Ouali, Hudelot, and Tami 2021)	ResNet-12	65.69±0.81	83.10±0.52	71.48±0.89	86.88±0.53
SKD (Rajasegaran et al. 2021)	ResNet-12	67.04±0.85	**83.54±0.54**	72.03±0.91	86.50±0.58
CPLAE (Gao et al. 2021)	ResNet-12	67.46±0.44	83.22±0.29	72.23±0.50	**87.35±0.34**
IEPT (Zhang et al. 2021)	ResNet-12	67.05±0.44	82.90±0.30	**72.24±0.50**	86.73±0.34
InfoPatch (Liu et al. 2021)	ResNet-12	**67.67±0.45**	82.44±0.31	71.51±0.52	85.44±0.35
ESPT† (Ours)	ResNet-12	**68.36±0.19**	**84.11±0.12**	**72.68±0.22**	**87.49±0.14**

Table 1: Performance comparison on miniImageNet and tieredImageNet. The mean 5-way few-shot classification accuracies (%, top-1) with the 95% confidence intervals are reported. † denotes the methods using local image representations.

episodes as well. The initial learning rate is set as 0.05 and decreased by a factor of 10 at 500 and 650 epochs. During training on the above three datasets, we evaluate the classification performance of learned models on the validation set after every 10 epochs, and select the best-performing model throughout the training process as our final result model.

Evaluation Metric Same as (Zhang et al. 2021; Ouali, Hudelot, and Tami 2021), we take the top-1 classification accuracy as the evaluation metric, and evaluate the performance of our method under standard 5-way 1-shot and 5-way 5-shot settings. For each experiment, we randomly sample 10,000 few-shot image classification tasks from the testing split of the dataset used, where each category to be classified contains 16 query samples. The mean top-1 accuracy and the 95% confidence intervals over these sampled tasks are calculated and reported.

Experimental Environment The Pytorch framework is used for our programming implementation and all the experiments are conducted in the following environment: Intel(R) Xeon(R) Gold 5117 @2.00GHz CPU, NVIDIA A100 Tensor Core GPU, and Ubuntu 18.04.6 LTS operation system. Under the above environment settings, our method takes about 162 ms for training on each 5-way 5-shot episode and such training iteration costs about 84 ms for FRN (Wertheimer, Tang, and Hariharan 2021), 97 ms for RENet (Kang et al. 2021) and over 240000 ms for DeepEMD (Zhang et al. 2020). But for each test iteration, since Eq. (5) is only applied on original image features, the computational cost of our method is reduced to be similar to that of FRN and RENet, all around 30 ms. Our model typically converges after 75000 iterations on CUB and 35000 iterations on miniImageNet and tieredImageNet.

Main Results

To evaluate the effectiveness of the proposed ESPT approach, we compare its performance with several representative and state-of-the-art FSL methods (Snell, Swersky, and Zemel 2017; Zhang et al. 2020; Wertheimer, Tang, and Hariharan 2021; Kang et al. 2021) as well as some recently proposed SSL augmented FSL models (Gidaris et al. 2019; Lee, Hwang, and Shin 2020; Zhang et al. 2021; Liu et al. 2021). The experiments are conducted on three different few-shot learning tasks: (1) the general image classification task on miniImageNet and tieredImageNet datasets, (2) the fine-grained image classification task on CUB-200-2011 dataset, and (3) the cross-domain miniImageNet→CUB few-shot image classification task.

Model	Backbone	CUB-200-2011 5-way 1-shot	5-shot
MatchingNet	ResNet-18	73.49±0.89	84.45±0.58
ProtoNet	ResNet-18	72.99±0.88	86.64±0.51
MAML	ResNet-18	68.42±1.07	83.47±0.62
Baseline	ResNet-18	65.51±0.87	82.85±0.55
Baseline++	ResNet-18	67.02±0.90	83.58±0.54
RelationNet	ResNet-18	68.58±0.94	84.05±0.56
Neg-Cosine	ResNet-18	72.66±0.85	89.40±0.43
FEAT	Conv4-64	68.87±0.22	82.90±0.15
CPLAE	Conv4-64	69.77±0.50	84.57±0.33
IEPT	Conv4-64	69.97±0.49	84.33±0.33
ProtoNet	ResNet-12	78.60±0.22	89.73±0.12
RFS-simple	ResNet-12	72.78±0.86	87.24±0.50
FEAT	ResNet-12	73.27±0.22	85.77±0.14
DeepEMD†	ResNet-12	75.65±0.83	88.69±0.50
RENet†	ResNet-12	79.49±0.44	91.11±0.24
FRN†	ResNet-12	**83.55±0.19**	**92.92±0.10**
ESPT† (Ours)	ResNet-12	**85.45±0.18**	**94.02±0.09**

Table 2: Performance comparison on CUB-200-2011. † denotes the methods using local image representations.

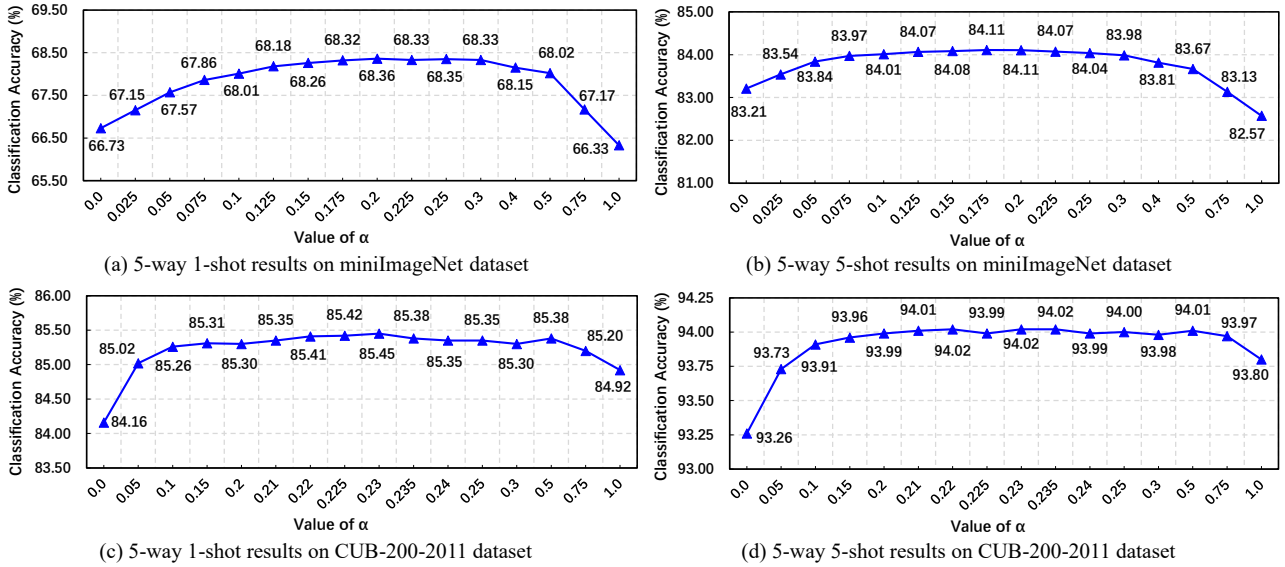
Model	Backbone	miniImageNet→CUB 1-shot	5-shot
MatchingNet	ResNet-10	35.89±0.51	51.37±0.77
RelationNet	ResNet-10	42.44±0.77	57.77±0.69
ProtoNet	ResNet-18	-	62.02±0.70
MAML	ResNet-18	-	51.34±0.72
Baseline	ResNet-18	-	65.57±0.70
Baseline++	ResNet-18	-	62.04±0.76
Neg-Softmax	ResNet-18	-	69.30±0.73
ProtoNet	ResNet-12	47.51±0.72	67.96±0.70
MetaOptNet	ResNet-12	44.79±0.75	64.98±0.68
FEAT	ResNet-12	50.67±0.78	71.08±0.73
SCL	ResNet-12	49.58±0.70	67.64±0.70
SCL-Distill	ResNet-12	50.09±0.70	68.81±0.60
FRN†	ResNet-12	51.60±0.21	**72.97±0.18**
TPMN†	ResNet-12	**52.83±0.65**	72.69±0.52
ESPT† (Ours)	ResNet-12	**54.14±0.21**	**74.91±0.18**

Table 3: Performance comparison on the cross-domain miniImageNet→CUB setting. † denotes the methods using local image representations.

General Few-Shot Image Classification To speed up the training process, we pre-train the feature extractor f_θ of our method before the episodic learning, as in (Wertheimer, Tang, and Hariharan 2021). From the experimental results reported in Table 1, we can obtain the following observations: (1) Among all FSL methods w/o SSL, the approaches using local image representations outperform the others, demonstrating the importance of capturing the local spatial information for FSL. (2) The SSL augmented FSL methods generally achieve better classification performance than the FSL methods w/o SSL, suggesting that SSL can promote FSL to learn more transferable feature representations. (3) The proposed ESPT method achieves the highest 1-shot and 5-shot classification accuracies of 68.36%, 84.11% on miniImageNet dataset and 72.66%, 87.49% on tieredImageNet dataset. Compared with the competing methods, ESPT obtains performance gains of at least 0.69%, 0.44% in 1-shot setting and 0.57%, 0.14% in 5-shot setting for the two datasets, respectively. Note that ESPT achieves such improvements without using any extra network structures (e.g., the attention module used in FEAT, RENet, SCL, IEPT or the rotation classifier used in CC+rot, SLA, SKD) or technical tricks (e.g., training with larger-way episodes or higher resolution input images). Therefore, these experimental results can show the effectiveness of our ESPT method, and also demonstrate the superiority of constructing self-supervised pretext task based on the local spatial relationships among multiple image samples in each episode.

Fine-Grained Few-Shot Image Classification For a fair comparison, we follow the prior works (Chen et al. 2019; Kang et al. 2021; Wertheimer, Tang, and Hariharan 2021) and directly episodic-train our model from scratch without using pre-training techniques. The fine-grained classification results on CUB-200-2011 dataset are presented in Table 2. It can be observed that (1) The local representation based-approaches once again achieve higher accuracies than other benchmark methods, which indicates that fine-grained FSL can also benefit from the local spatial image features. (2) Our proposed ESPT method obtains the best classification results of 85.45% and 94.02% for the 1-shot and 5-shot settings, respectively, outperforming the second best FRN method by a significant margin of 1.90 % and 1.10 %, and is far superior to other competing methods. This shows that our ESPT can also be effective on the fine-grained few-shot image classification tasks.

Cross-Domain Few-Shot Image Classification Following the setup in (Chen et al. 2019; Wu et al. 2021), we evaluate the proposed ESPT method in a more challenging cross-domain miniImageNet→CUB setting. Concretely, the model is trained on all 100 classes in the miniImageNet dataset, but validated and evaluated on 50 validation classes and 50 testing classes from the CUB-200-2011 dataset, respectively. The obtained cross-domain few-shot image classification results are shown in Table 3. We can see that our ESPT method outperforms all benchmark approaches by a large margin. Specifically, the improvements achieved by ESPT over other competing methods range from 1.31% (vs. TPMN) to 18.25% (vs. MatchingNet) in 5-way 1-shot setting and from 1.94% (vs. FRN) to 23.57% (vs. MAML) in 5-way 5-shot setting. These experimental results indicate that augmenting FSL with our proposed ESPT objective promotes learning more transferable feature representations, which can generalize well to novel unseen categories even under domain shift.



(a) 5-way 1-shot results on miniImageNet dataset

(b) 5-way 5-shot results on miniImageNet dataset

(c) 5-way 1-shot results on CUB-200-2011 dataset

(d) 5-way 5-shot results on CUB-200-2011 dataset

Figure 2: Effect of the proposed ESPT objective on the performance of our method on miniImageNet and CUB-200-2011.

Transformation			miniImageNet		CUB-200-2011	
90°	180°	270°	1-shot	5-shot	1-shot	5-shot
✓			68.02	83.86	84.92	93.76
	✓		68.19	84.03	85.06	93.88
		✓	68.05	83.83	84.88	93.78
✓	✓		68.32	84.10	85.22	93.97
✓		✓	68.18	84.03	**85.45**	**94.02**
	✓	✓	**68.36**	**84.11**	85.24	93.95
✓	✓	✓	**68.36**	84.07	85.41	**94.02**

Table 4: Effect of different choices of the transformation set U on the performance of our method.

Ablation Studies

Effect of the Proposed ESPT Objective We analyze the effect of the proposed ESPT objective by comparing the few-shot image classification performance of our method with different loss weight (i.e., α in equation (13)) values on miniImageNet and CUB-200-2011 datasets, as shown in Figure 2. It can be seen that, with the value of α increases from zero, the accuracy curves of our method on both datasets rise gradually at first, then reach their peak values and remain relatively stable at a high level when α is within the range of $[0.1, 0.5]$. This shows that the proposed ESPT objective can effectively and steadily improve the few-shot classification performance of our method. In addition, it can also be found that the best models of our full approach significantly outperform our method without the ESPT objective (i.e., $\alpha = 0$). The improvements on the 5-way 1-shot and 5-way 5-shot tasks are 1.63%, 0.90% on miniImageNet dataset and 1.29%, 0.76% on CUB-200-2011 dataset, respectively. These results can further demonstrate the effectiveness of the proposed ESPT objective.

Effect of the Transformation Set The 5-way classification results obtained with different transformation sets U on miniImageNet and CUB-200-2011 datasets are summarized in Table 4. As shown, it can be seen that the transformation sets with multiple rotation transformations typically produce better classification results than those containing only a single rotation transformation. It is mainly because that more transformations will bring more different data variants, which will benefit our ESPT method to learn feature representations with stronger generalization ability. Moreover, we can also observe that the models trained with the transformation sets $U = \{180°, 270°\}$ and $\{90°, 270°\}$ have the highest classification accuracies on the two datasets, respectively. We analyze the reason why $U = \{90°, 180°, 270°\}$ cannot lead to the best results is because that these three rotation transformations may introduce some redundant variant information that is not useful for FSL.

Conclusion

In this paper, we augment the few-shot classification objective with a newly proposed Episodic Spatial Pretext Task (ESPT) to learn more transferable image representations. By leveraging the local spatial relationships between the support and query samples in each learning episode, ESPT can effectively capture the low-level visual information in different images and the data structure information of the whole episode, which will bring significant benefits to FSL. Extensive experiments on three widely used benchmark datasets demonstrate the effectiveness and the superiority of our ESPT method. And importantly, while achieving the state-of-the-art classification performance, the proposed ESPT method does not increase the model complexity and the risk of data overfitting, which makes ESPT more suitable for solving FSL problems with limited training data.

Acknowledgments

The research was supported by the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City (Grant No: 2021JJLH0099), Project of Sanya Yazhou Bay Science and Technology City (Grant No: SCKJ-JYRC-2022-76), Postdoctoral project of Hainan Yazhou Bay Seed Laboratory (Grant No: B22E18102).

References

Bertinetto, L.; Henriques, J. F.; Torr, P.; and Vedaldi, A. 2019. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A Closer Look at Few-shot Classification. In *International Conference on Learning Representations*.

Chen, Y.; Liu, Z.; Xu, H.; Darrell, T.; and Wang, X. 2021. Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9062–9071.

Doersch, C.; Gupta, A.; and Zisserman, A. 2020. CrossTransformers: spatially-aware few-shot transfer. In *Advances in Neural Information Processing Systems*, volume 33, 21981–21993. Curran Associates, Inc.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1126–1135. PMLR.

Gao, Y.; Fei, N.; Liu, G.; Lu, Z.; and Xiang, T. 2021. Contrastive prototype learning with augmented embeddings for few-shot learning. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, 140–150. PMLR.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Perez, P.; and Cord, M. 2019. Boosting Few-Shot Visual Learning With Self-Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8059–8068.

Hariharan, B.; and Girshick, R. 2017. Low-Shot Visual Recognition by Shrinking and Hallucinating Features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3018–3027.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, S. X.; Moreno, P. G.; Xiao, Y.; Shen, X.; Obozinski, G.; Lawrence, N.; and Damianou, A. 2020. Empirical Bayes Transductive Meta-Learning with Synthetic Gradients. In *International Conference on Learning Representations*.

Islam, A.; Chen, C.-F. R.; Panda, R.; Karlinsky, L.; Radke, R.; and Feris, R. 2021. A Broad Study on the Transferability of Visual Representations With Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8845–8855.

Jamal, M. A.; and Qi, G.-J. 2019. Task Agnostic Meta-Learning for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11719–11727.

Kang, D.; Kwon, H.; Min, J.; and Cho, M. 2021. Relational Embedding for Few-Shot Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8822–8833.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Lee, H.; Hwang, S. J.; and Shin, J. 2020. Self-supervised Label Augmentation via Input Transformations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5714–5724. PMLR.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-Learning With Differentiable Convex Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10657–10665.

Li, F.-F.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4): 594–611.

Li, H.; Eigen, D.; Dodge, S.; Zeiler, M.; and Wang, X. 2019. Finding Task-Relevant Features for Few-Shot Learning by Category Traversal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–10.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2898–2907.

Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. *CoRR*, abs/1707.09835.

Liu, B.; Cao, Y.; Lin, Y.; Li, Q.; Zhang, Z.; Long, M.; and Hu, H. 2020. Negative Margin Matters: Understanding Margin in Few-Shot Classification. In *Computer Vision – ECCV 2020*, 438–455. Cham: Springer International Publishing.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a Few-shot Embedding Model with Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8635–8643.

Liu, Y.; Schiele, B.; and Sun, Q. 2020. An Ensemble of Epoch-Wise Empirical Bayes for Few-Shot Learning. In *Computer Vision – ECCV 2020*, 404–421. Springer International Publishing.

Meng, Q.; Zhao, S.; Huang, Z.; and Zhou, F. 2021. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14225–14234.

Ouali, Y.; Hudelot, C.; and Tami, M. 2021. Spatial Contrastive Learning for Few-Shot Classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 671–686. Springer International Publishing.

Park, S.-J.; Han, S.; Baek, J.-W.; Kim, I.; Song, J.; Lee, H. B.; Han, J.-J.; and Hwang, S. J. 2020. Meta Variance Transfer: Learning to Augment from the Others. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7510–7520. PMLR.

Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F. S.; and Shah, M. 2021. Self-supervised Knowledge Distillation for Few-shot Learning. In *British Machine Vision Conference*.

Ravi, S.; and Larochelle, H. 2017. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations*.

Ren, M.; Ravi, S.; Triantafillou, E.; Snell, J.; Swersky, K.; Tenenbaum, J. B.; Larochelle, H.; and Zemel, R. S. 2018. Meta-Learning for Semi-Supervised Few-Shot Classification. In *International Conference on Learning Representations*.

Rizve, M. N.; Khan, S.; Khan, F. S.; and Shah, M. 2021. Exploring Complementary Strengths of Invariant and Equivariant Representations for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10836–10846.

Rusu, A. A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; and Hadsell, R. 2019. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*.

Satorras, V. G.; and Estrach, J. B. 2018. Few-Shot Learning with Graph Neural Networks. In *International Conference on Learning Representations*.

Schwartz, E.; Karlinsky, L.; Shtok, J.; Harary, S.; Marder, M.; Kumar, A.; Feris, R.; Giryes, R.; and Bronstein, A. 2018. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Simon, C.; Koniusz, P.; Nock, R.; and Harandi, M. 2020. Adaptive Subspaces for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4136–4145.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Su, J.-C.; Maji, S.; and Hariharan, B. 2020. When Does Self-supervision Improve Few-Shot Learning? In *Computer Vision – ECCV 2020*, 645–666. Springer International Publishing.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1199–1208.

Tang, S.; Chen, D.; Bai, L.; Liu, K.; Ge, Y.; and Ouyang, W. 2021. Mutual CRF-GNN for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2329–2339.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking Few-Shot Image Classification: A Good Embedding is All You Need? In *Computer Vision – ECCV 2020*, 266–282. Springer International Publishing.

Vinyals, O.; Blundell, C.; Lillicrap, T.; kavukcuoglu, k.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Wang, Y.-X.; Girshick, R.; Hebert, M.; and Hariharan, B. 2018. Low-Shot Learning From Imaginary Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7278–7286.

Wertheimer, D.; Tang, L.; and Hariharan, B. 2021. Few-Shot Classification With Feature Map Reconstruction Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8012–8021.

Wu, J.; Zhang, T.; Zhang, Y.; and Wu, F. 2021. Task-Aware Part Mining Network for Few-Shot Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8433–8442.

Yang, S.; Liu, L.; and Xu, M. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *International Conference on Learning Representations*.

Ye, H.-J.; Hu, H.; Zhan, D.-C.; and Sha, F. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8808–8817.

Zhang, C.; Cai, Y.; Lin, G.; and Shen, C. 2020. DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12203–12213.

Zhang, M.; Zhang, J.; Lu, Z.; Xiang, T.; Ding, M.; and Huang, S. 2021. IEPT: Instance-Level and Episode-Level Pretext Tasks for Few-Shot Learning. In *International Conference on Learning Representations*.

Zhang, R.; Che, T.; Ghahramani, Z.; Bengio, Y.; and Song, Y. 2018. MetaGAN: An Adversarial Approach to Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.