Task-level Self-supervision for Cross-domain Few-shot Learning

Wang Yuan[1 †], **Zhizhong Zhang** [1 †], **Cong Wang** [3], **Haichuan Song** [1], **Yuan Xie** [1*], **Lizhuang Ma** [1,2,4*]

[1] East China Normal University
[2] Shanghai Jiao Tong University
[3] Huawei Technologies
[4] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
wangyuan05@stu.ecnu.edu.cn,zzzhang@cs.ecnu.edu.cn

Abstract

Learning with limited labeled data is a long-standing problem. Among various solutions, episodic training progressively classifies a series of few-shot tasks and thereby is assumed to be beneficial for improving the model's generalization ability. However, recent studies show that it is even inferior to the baseline model when facing domain shift between base and novel classes. To tackle this problem, we propose a domain-independent task-level self-supervised (TL-SS) method for cross-domain few-shot learning. **TL-SS** strategy promotes the general idea of label-based instance-level supervision to task-level self-supervision by augmenting multiple views of tasks. Two regularizations on task consistency and correlation metric are introduced to remarkably stabilize the training process and endow the generalization ability into the prediction model. We also propose a high-order associated encoder (**HAE**) being adaptive to various tasks. By utilizing 3D convolution module, HAE is able to generate proper parameters and enables the encoder to flexibly to any unseen tasks. Two modules complement each other and show great promotion against state-of-the-art methods experimentally. Finally, we design a generalized task-agnostic test, where our intriguing findings highlight the need to re-think the generalization ability of existing few-shot approaches.

Introduction

Learning effective prediction model with limited labeled data is a long-standing problem. As a significant advance, few-shot learning (FSL), which generalizes the meta-knowledge in base classes (sufficient samples) to novel classes (few labeled data), has attracted considerable attention. With the prevalence of meta-learning training strategy "episode" (Vinyals et al. 2016), the baseline of few-shot learning has been continuously improved. However, recent studies (Chen et al. 2019b; Tseng et al. 2020; Triantafillou et al. 2020; Guo et al. 2019) show that, when there exists domain shift between training and test data, episodic training will be greatly affected or even inferior to the primary model (*i.e.,* frozen the ConvNet and fine-tune the fully-connected layer with few target data).

Intuitively, "episode" strategy is naturally assumed to be beneficial for the cross-domain problem. By establishing continuous heterogeneous tasks (*i.e.,* N-way K-shot tasks), episode emulates the process that progressively classifies a series of datasets, each including disjoint categories. It is expected that the learned representation can generalize to novel classes. However, as indicated by (Su et al. 2019), episode technique might discard semantic information that is irrelevant for base classes but critical for novel classes. Despite the data availability, training for base class classification on the source data would not be reflective.

To tackle this issue, we find recent proposed "task-specific adaptation" (Ye et al. 2020; Oreshkin et al. 2018; Li et al. 2019; Guo et al. 2020) methods show a promising direction. A key point lies on that these methods raise their focus on the concept of "task", instead of instance-level classification. They utilize the task statistics, *e.g.,* mean and variance in a episode, to abstract a specific task with a brief context description. It allows the model to generate task-specific encoder parameters, and therefore encourages the model to focus on current task. Since various task descriptions results in various task-specific encoders, it provides great diversity for training, and is more in line with the essence of the episodic

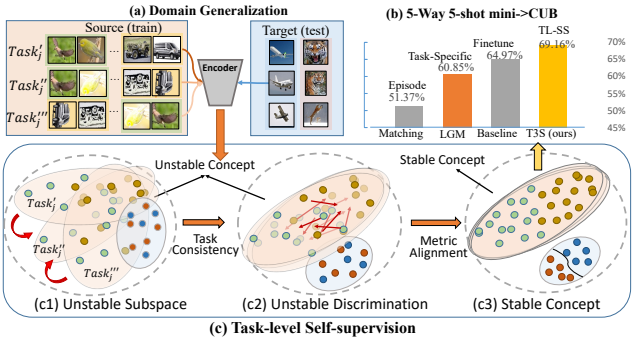

Figure 1: The motivation. Traditional few-shot methods are greatly affected by domain shift. Some task-specific methods initially show some potential, but still have a gap with the baseline (b). We attribute it to its "unstable concept". Our insight is to propose a task-level self-supervised strategy to improve generalization (c). By constraining the task consistency (c1) and metric alignment between different views of the same task, the stability and discrimination (c2) of the subspace are guaranteed respectively.

[1*] corresponding author. † equal contribution.

strategy. As a result, their generalization ability would be remarkably improved. LGM (Li et al. 2019) in Fig. 1(b) also shows a great promotion on the cross-domain task, which consistently support our claims.

Despite their large potential, existing task-specific approaches still haven't achieved satisfactory results under the cross-domain setting. We attribute it in two aspects: 1) Unstable subspace distribution (Fig. 1(c1)). Existing task-specific adaptation methods often use the task statistics as the task abstraction, but it is less robust in the context of few-shot learning (a task only contains very few samples). Since the task context modeling is always ambiguous, the task context would drift with the domain shift. 2) Lack of local feature discrimination (Fig. 1(c2)). Current task-specific approaches, only dependent on instance-level supervision, are easy to overfitting due to the huge domain discrepancy.

In this paper, we propose a novel Task-Level Self-Supervised strategy (**TL-SS**). Our observation is that the label-based supervision is easily influenced by the observation domain, and the instance-level self-supervision is insufficient to handle the domain discrepancy problem. Hence, following the episodic training and task-specific methods, we promote the idea of instance-level to task-level supervision under the assumption that each episode can be considered as a specific domain[1]. A key difference between instance-level and task-level supervision lies on we pay more attention on the task itself rather than the instance. To do so, multiple views of task are constructed via instance-level augmentation and class-level permutation. We force task context modeling results of different views to be similar (Fig. 1(c1)). Thus when episode is coming one by one, we continually train the model on various domains, and this task-level consistency supervision allows us to learn stable concept and can promote more transferable knowledge towards model generalization.

To extract more robust task context, we adopt the methodology of "learning-to-learn" from meta-learning and construct a high-order associated encoder (**HAE**). Different from previous methods (Li et al. 2019), we utilize a 3D convolutional network to abstract task context by capturing the intrinsic data structure within a task. It not only highlights the discrepancy between tasks and enables the encoder to flexibly adapt to any unseen task, but also complement to TL-SS (*e.g.,* class-level permutation). Finally, we introduce a novel weight generator to adaptively generate variable parameters according to the context modeling (Fig. 2(b)).

We evaluate our method on standard benchmarks and design a new task-agnostic test to show the effectiveness of our approach. Our contributions are summarized as:

- We propose a novel task-level self-supervised (**TL-SS**) training strategy. Two task-level regularizations are used to promote the training stability and discriminative representation.
- A high-order associated encoder (**HAE**) is proposed whose parameters can adapt to any unseen domain according to the robust task context modeling.
- Finally, we evaluate our method on several standard benchmarks and design a more generalized task-agnostic test. Sufficient experiments clearly prove the effectiveness of our proposed method.

[1]This class-difference-caused distribution shift among heterogeneous tasks can be considered as a special case of domain shift.

Problem Definition and Related Work

Classification tasks usually contain a set of data D_s (source), typically large-scale, to train the base model. For testing, there is also a set of data D_t (target), which includes labeled samples (support data) and unlabeled ones need to be classified (query data). In the following, $C(\cdot)$ and $P(\cdot)$ denote the categories and distributions of a dataset respectively.

Traditional Cross Domain Classification: $C(D_s) = C(D_t), P(D_s) \neq P(D_t)$. The main motivation is to extract "domain invariant" features of samples with the same categories but different domains. Most of them are only valid for homogeneous tasks.

Few-shot Learning (FSL): $C(D_s) \cap C(D_t) = \emptyset, P(D_s) \approx P(D_t)$, and the support data is few. MatchingNet (Vinyals et al. 2016) proposed the "episode" training strategy and several milestones have emerged after this, including metric based (Snell et al. 2017; Sung et al. 2018; Satorras and Estrach 2018; Kim et al. 2019), optimization based (Finn et al. 2017; Andrychowicz et al. 2016; Lee et al. 2019; Li et al. 2019) and others (Santoro et al. 2016; Sung et al. 2017).

Cross Domain Few-shot Learning (CD-FS)[2]: $C(D_s) \cap C(D_t) = \emptyset, P(D_s) \neq P(D_t)$, and support data is insufficient. (Chen et al. 2019b) that current FSL methods degraded significantly when encountering domain shifts. (Dong and Xing 2018) is the first study to address this issue in the one-shot learning setting, but they assume they can access unlabeled data in target domain. Recently, (Tseng et al. 2020) utilized feature-wise transformation layers to reduce feature distribution shift.

Task-specific & Weight generating Based Methods: Recently, task-specific methods (Oreshkin et al. 2018; Ye et al. 2020) and weight generating (Guo et al. 2020; Li et al. 2019) based methods have shown great potential for CD-FS. They aim to adapt the model dynamically according to the current task. The most inspirational work to us is (Li et al. 2019), which generates encoder parameters according to task context. However, as mentioned above, its task context is unstable. Instead, we design a more reasonable task modeling module and introduce self-supervised regularization to make it suitable for CD-FS task.

Self Supervised learning: Predicting the colors (Larsson et al. 2016), position (Noroozi and Favaro 2016), rotation (Gidaris et al. 2019) and the missing part (Pathak et al. 2016) of the images are some of the many methods for self-supervised learning. Recently, these methods have been proposed for solving FSL (Chen et al. 2019a; Gidaris et al. 2019; Su et al. 2019) and cross-domain (Xu et al. 2019; Carlucci et al. 2019) tasks. But they are all concentrating on instance-level self-supervision. To our best knowledge, we are the pioneer's work for task-level self-supervision. It is

[2]CD-FS in our work is limited to heterogeneous task. Homogeneous task (Motiian et al. 2017; Teshima et al. 2020) which weakens the FSL requirements is not within the scope of this work.

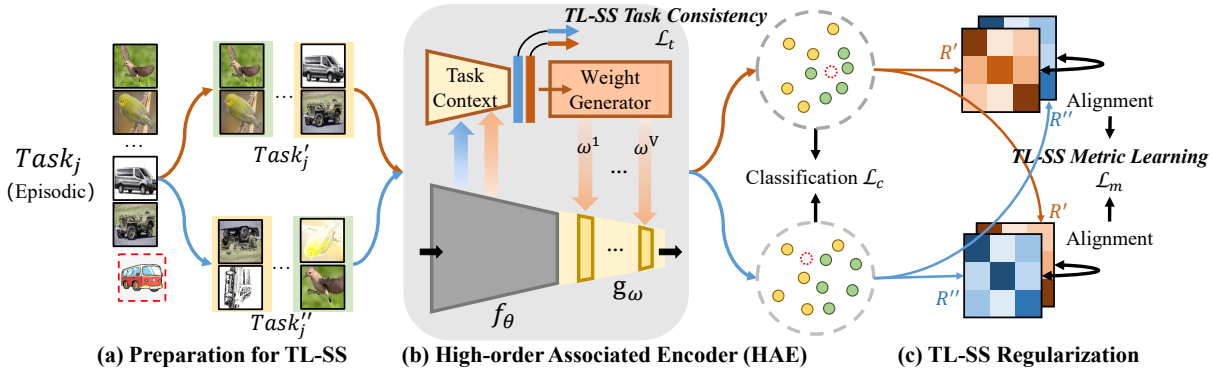

Figure 2: Illustration of the proposed overall framework. We construct a high-order associated encoder via task-level self-supervision to solve CD-FS task. (a) The task-level self-supervised (**TL-SS**) strategy proposed in this paper varies $Task_i$ sampled by the **episode** strategy into multi-views (red and blue data flow). (b) The parameters of high-order associated encoder (**HAE**) are adaptively generated by a weight generator according to task context modeling results. (c) TL-SS introduce two constraints: TL-SS task consistence loss $\mathcal{L}_c$ and TL-SS metric learning loss $\mathcal{L}_m$.

expected that these task consistency can largely stabilize the meta-learning process and hence improve the generalization ability of the trained model.

Methodology

Overview

Under the cross-domain few-shot context, we are given a collection of data $\boldsymbol{\mathcal{X}}_s$ from a specific domain, where we attempt to train a model with these data, but expect it could be generalized to other domains $\boldsymbol{\mathcal{X}}_t$, especially the scale of $\boldsymbol{\mathcal{X}}_t$ is small (*e.g., N-way K-shot*). Furthermore, we follow the general FSL setting that ignores $\boldsymbol{\mathcal{X}}_t$ in the training phase and *does not require any fine-tuning processes*, enabling fast model deployment to the unseen task.

Concretely, we follow the "episode" sampling strategy to simulate a N-way K-shot task, which randomly selects N categories from the training set, with K random support samples for each category and Q query samples in one episode. That is:

$$\boldsymbol{T}_j = \left\{ \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{NK}, \{\boldsymbol{x}_{NK+1}, \ldots, \boldsymbol{x}_{NK+Q}\} \right\}, \quad (1)$$

where $\boldsymbol{x}_i \in \boldsymbol{\mathcal{X}}_s$ denotes the training samples and y_i denotes the corresponding labels. In our task-level supervision framework (Fig. 2), it consists of two key components: 1) A High-order Associated Encoder (**HAE**)). 2) A Task-Level Self-Supervised regularization (**TL-SS**).

High-order Associated Encoder (HAE)

We propose a High-order Associated Encoder (HAE) for extracting domain-adaptive features. Motivated by meta-learning, we concretize the meta-knowledge as the ability to generate the suitable parameters for a given domain/task. Thus the encoder is adaptive to various tasks and yields task-specific subspace. Task-Level Self-Supervised regularization is then used to make these subspace stable, discriminative, and hence improve the generalization ability.

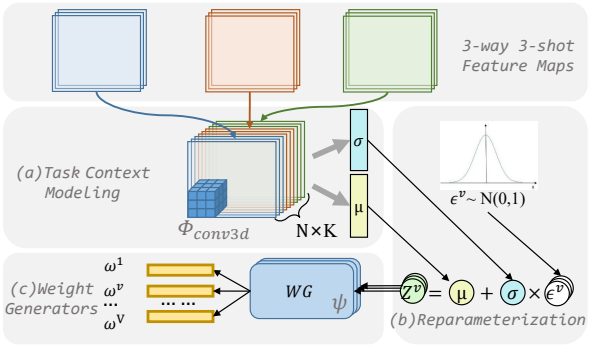

Figure 3: Illustration of the High-order Associated Encoder (HAE). For simplicity, one sample shows only one channel of feature map.

HAE is divided into a set of shallow layers f_θ and a set of deep layers g_ω, where f_θ are the traditional convolutional layers with trainable parameters. For a N-way K-shot task, f_θ first transforms all samples into the feature maps $\left\{ f_\theta(\boldsymbol{x}_i) \in \mathbb{R}^{H \times W \times C} \right\}_{i=1}^{NK+Q}$, where H, W and C denote the height, width and channel number, respectively. Then, the parameters of g_ω are generated by the *Weight Generator* based on the results of *Task Context Modeling*.

Task Context Modeling: We design a task context modeling module $M(\cdot;\phi)$ to abstract a specific task $\boldsymbol{T_j}$ with a fixed-size feature $\boldsymbol{\tau}_j \in \mathbb{R}^d$, $d = 128$ in our case. The main role of $\boldsymbol{\tau}_j$ is to thoroughly reflect differences among tasks. From this perspective, we propose to utilize the relationship among samples as the task context rather than the instance-level representations. In particular, we splice the feature maps of all support samples into a video-like format, where each category serve as a specific action unit. We summarize the task context as a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)$. We use a 3D CNN to jointly learn the

statistics $\boldsymbol{\mu}_j \in \mathbb{R}^{d/2}$ and $\boldsymbol{\sigma}_j \in \mathbb{R}^{d/2}$ of task $\boldsymbol{T_j}$ (Fig. 3(a)):

$$\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j = M\left(\{f_\theta(\boldsymbol{x}_i)\}_{i=1}^{NK}; \phi_{con3d}\right), \tag{2a}$$

$$\boldsymbol{\tau}_j = \left[\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j\right], \tag{2b}$$

where ϕ_{con3d} indicates the parameters of the 3D CNN. $[\cdot]$ indicates the concatenate operation. $\boldsymbol{\tau}_j$, obtained by concatenating $\boldsymbol{\mu}_j$ and $\boldsymbol{\sigma}_j$, refers to *task context feature*. Intuitively, 3D convolution network enjoys the primary advantages: 1) The input of 3D CNN is the combination of all samples which treats episodic training from a task-level aspect; 2) 3D CNN captures the correlations not only in spatial dimension but also among sample relationships, referred to "high-order association". 3) As indicated previously, 3D CNN is able to extract the order information, a unique task-level characteristic. Thus 3D CNN facilities our task augmentation and task-level self-supervision[3] (in Sec.TL-SS).

Weight Generator: Given current task context feature $\boldsymbol{\tau}_j$, the weight generator aims to generate parameters of g_ω in which we suppose there are V layers, and use $G_v(\cdot;\psi)$ to denote the v-th layer generator. In order to ensure the gradient could be back-propagated properly, we refer to the re-parameterization trick in VAE (Kingma and Welling 2014) to make the whole process differentiable (Fig. 3(b)). Formally, for the v-th layer generator, we first define a re-sampling variable :

$$\boldsymbol{z}_j^v = \boldsymbol{\mu}_j + \boldsymbol{\sigma}_j \times \boldsymbol{\epsilon}^v, \text{ with } \boldsymbol{\epsilon}^v \sim \mathcal{N}(0,1), \tag{3}$$

where $\boldsymbol{\epsilon}^v \in \mathbb{R}^{d/2}$ is an auxiliary noise variable, $\boldsymbol{z}_j^v$ denotes the v-th re-sampling instance of task context features. Finally, the parameters of v-th layer in $g_\omega(\cdot)$ corresponds to the v-th layer in $G_v(\cdot;\psi)$ is given as:

$$\boldsymbol{\omega}_j^v = G_v(\boldsymbol{z}_j^v; \psi^v), \tag{4}$$

where ψ^v indicates the parameters of G_v, and $\boldsymbol{\omega}^v$ represents the generated parameters of v-th layer in $g_\omega()$.

To stabilize the training process, we apply L2 weight normalization like (Li et al. 2019) on the generated parameters $\boldsymbol{\omega}$. Besides, to reduce the amount of parameters, the weight generator only produces parameters of deep layers g_ω, based on the assumption that shallow layers f_θ are used to extract low-level features, which is invariant to various tasks.

Task-level Self-supervision (TL-SS)

Although HAE allows our model to quickly adapt to various tasks, as discussed before, the learned subspace for each task is always ambiguous and not ensured to be discriminative. This leads the learned metric can not easily accommodate to novel categories. In the following, we propose a task-level self-supervised (TL-SS) training strategy.

Task Variant: For a N-way K-shot task $\boldsymbol{T}_j$, TL-SS first constructs multiple views for $\boldsymbol{T}_j$ according to the following steps. 1) Instance-level Augmentation. Every sample is randomly rotated, flipped or slightly scaled and cropped. It allows the network to learn diverse representations of the same instance. 2) Class-level Permutation. Scramble the class order, as well as the sample order from a specific category, to construct diverse representations of the same task. Note that different sample and class orders result in different outputs of 3D CNN, and therefore these preparations provide variant versions of certain task for conducting self-supervision. For simplicity, we take two views as an example to clearly introduce our method (red $\boldsymbol{T}'_j$ and blue $\boldsymbol{T}''_j$ data streams in Fig. 2(a)).

TL-SS Task Consistency Loss: With multiple views of $\boldsymbol{T}_j$, we design a regularization term, *i.e.,* TL-SS task consistency loss $\mathcal{L}_t$, with the observation that different views of the same task should present the same task context. So, we adopt cross-entropy loss:

$$\mathcal{L}_t = \sum -\boldsymbol{\tau}'_j \log \boldsymbol{\tau}''_j, \tag{5}$$

where $\boldsymbol{\tau}'_j$ and $\boldsymbol{\tau}''_j$ denote the task context features of different views of $\boldsymbol{T}_j$ in Eq.(2). Eq.(5) directly imposes task context self-supervision, and therefore task context modeling and task consistency loss complement each other (Fig. 2(b)), which significantly stabilize the subspace distribution.

TL-SS Metric Learning Loss: Previous (Dou et al. 2019) have shown that, relationships among class concepts purely exist in semantic space, independent of changes in the observation space. In this work, "relationships between classes" is naturally expanded to "correlation among samples". Inspired by instance discrimination (Wu et al. 2018), we propose to force the relative position among samples to be consistent in different views, allowing variants of task to supervise each other.

Specifically, for a task $\boldsymbol{T}_j$, after task augmentation, we extract the multi-view task embeddings $\boldsymbol{e}'(i) = g(f_\theta(\boldsymbol{x}'_i), \omega_j)$ and $\boldsymbol{e}''(i) = g(f_\theta(\boldsymbol{x}''_i), \omega_j)$, where $\boldsymbol{x}'_i \in \boldsymbol{T}'_j$ and $\boldsymbol{x}''_i \in \boldsymbol{T}''_j$. Then, we calculate their relative positions by Cosine and Euclidean distance:

$$\begin{aligned}\boldsymbol{R}'_j(m,n) &= \left[cos(\boldsymbol{e}'(m), \boldsymbol{e}'(n)), D(\boldsymbol{e}'(m), \boldsymbol{e}'(n))\right], \\ \boldsymbol{R}''_j(m,n) &= \left[cos(\boldsymbol{e}''(m), \boldsymbol{e}''(n)), D(\boldsymbol{e}''(m), \boldsymbol{e}''(n))\right],\end{aligned} \tag{6}$$

where $m \neq n$, $\boldsymbol{R}_j \in \mathbb{R}^{2\times NK\times NK}$. $cos(\cdot)$ and $D(\cdot)$ indicate the Cosine and Euclidean distance, respectively. Finally, we propose to align these two matrices between different views $\boldsymbol{R}'_j$ and $\boldsymbol{R}''_j$ by minimizing their symmetrical Kullback–Leibler (KL) divergence:

$$\mathcal{L}_m = \frac{1}{2}\left[D_{\mathrm{KL}}\left(\boldsymbol{R}'_j \| \boldsymbol{R}''_j\right) + D_{\mathrm{KL}}\left(\boldsymbol{R}''_j \| \boldsymbol{R}'_j\right)\right], \tag{7}$$

where $D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) = \sum_r p_r \log \frac{p_r}{q_r}$ denotes the KL divergence. In this way, the local discrimination is much robust by adding this self-supervised regularization.

In addition, for each query sample $\tilde{\boldsymbol{x}}$, we also calculate the classification loss $\mathcal{L}_c$ for supervised learning:

$$\mathcal{L}_c = \sum_i -\boldsymbol{p}_i \log(\boldsymbol{y}_i), \tag{8}$$

where $\boldsymbol{y}$ denotes the ground-truth and $\boldsymbol{p}$ is the predicted distribution:

$$\boldsymbol{p} = \sum_{i=1}^{N\times K} \frac{e^{d(\tilde{\boldsymbol{x}}, \boldsymbol{x}_i)}}{W} \boldsymbol{y}_i, \tag{9}$$

[3]We do not think there is a sequence among samples. On the contrary, we will constrain the contexts of different task variants to be consistent in the subsequent TL-SS

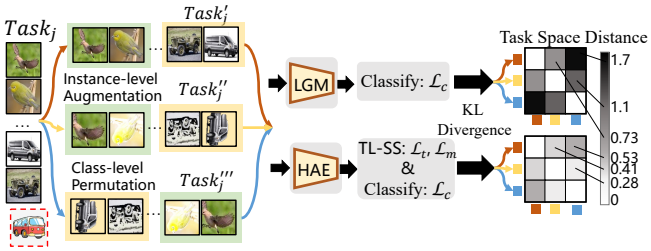

Figure 4: The combination of HAE and TL-SS ensures consistency among different views of a specific task. The results are directly obtained from our experiments.

where $W = \sum_{i=1}^{N\times K} e^{d(\tilde{\boldsymbol{x}},\boldsymbol{x}_i)}$, $d(\cdot)$ is Cosine distance. $\boldsymbol{x}_i$ represents support sample. Overall, the total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_t + \mathcal{L}_m + \mathcal{L}_c. \tag{10}$$

Intuitively, different views of task should be consistent. In Fig. 4, we use KL divergence to measure the distribution discrepancy among different task views. The results show that 3D CNN could capture small modifications from various task augmentations and TL-SS is able to effectively regularize the model, making it stable, consistent.

Experiments

For simplicity, we call our framework **T**ask-**S**pecific **S**elf-**S**upervised Learning as **T3S**. We validate the effectiveness of the proposed T3S under two cross-domain few-shot settings. First, we train the few-shot classification model on the *mini*ImageNet and test the trained model on other eight different benchmarks (four in supplementary material). Second, we design a generalized task-agnostic test, where our intriguing findings highlight the need to re-think the generalization ability of existing FSL methods.

Experimental Settings

Source Dataset: *mini*ImageNet (Vinyals et al. 2016), a subset of the ILSVRC-12 (Deng et al. 2009), is a standard benchmark for few-shot image classification. It consists of 60,000 color images of size 84×84 with 100 classes. We follow the splitting introduced by (Ravi and Larochelle 2017), with 64, 16, and 20 classes for training, validation and testing, respectively. We take the training set as source domain and select the model on the validation data.

Target Dataset: CUB (Welinder et al. 2010), "Caltech-UCSD Birds-200-2011" is a fine-grained dataset containing 200 classes and 11,788 bird images in total. **Cars** (Krause et al. 2013) contains 196 classes and 16,185 car images in total. **Places** (Zhou et al. 2018) "places365standard" contains more than 10 million real-world scenes from 365 categories. **EuroSAT** (Helber et al. 2019) contains 10 classes and 27,000 satellite images in total. With standard splitting, the test set of CUB, Cars and Places as well as the whole EuroSAT serve as the target domain.

Evaluation: To evaluate our approach, all the results are obtained under standard few-shot classification task: 5-way 1-shot/5-shot task. Also, as recommended in (Triantafillou et al. 2020), we analyze the effect of different "way" and "shot" on mini → CUB experiments in Supplementary Material. We use classification accuracy as the evaluation metric and present the average results over 1000 trials.

Implementation Details

Baseline Clarification. The baseline model follows prior work (Chen et al. 2019b) simply including a ResNet-10 backbone and a fully-connected layer. It is trained from scratch with a batch size of 32 by minimizing the standard cross-entropy loss on 64 base classes. Before testing on the target domain, we train a new linear classifier using $N \times K$ support samples to fine-tune the model. We find this baseline beats almost all the episodic training method.

Training T3S. We take first eight layers of baseline as the initialization for shallow layers f_θ. We take C3D (Tran et al. 2015) as our task context model but replace the FC layers with a global average pooling (details in supplemental material). The weight generator contains two single perceptrons, each of which corresponds to a specific layer in g_ω. We totally train 1,000K episodes for our model using the Adam optimizer with initial learning rate 10^{-3} and exponentially decayed by 50% every 50k episode. In each episode, we sample $N \times K$ support samples and 16 query samples. The mini-batch size is empirically set to be 64 for 5-way 1-shot/5-shot task.

Main Results

In this section, we demonstrate the effectiveness of our approach against state-of-the-art methods. The competitors are implemented by official code or our re-implementation if the results are not reported on papers[4].

Cross-domain few-shot results (Table 1). We conduct 5-way 1-shot/5-shot task on four cross-domain experiments: *mini*ImageNet → CUB/Cars/Place/EurSAT. As shown in Table 1, our method nearly achieves the best performance. Especially on the 5-shot task, we evidently outperform other methods with around 3% improvement, which demonstrates the good generalization ability of our approach. Note that the improvement on 1-shot setting is less obvious than the 5-shot task due to there is less correlation information can be utilized by HAE in 1-shot task. Also, as indicated previously, it appears that most few-shot methods are inferior to the baseline model. We conjecture that fine-tuning with only a few target samples is also effective, but this process is tedious and time-consuming (Detailed analysis in supplemental material).

Comparison with Related SOTA Methods (Table 2). We also compare the proposed method with related technologies in Table 2. It turns out that the proposed T3S not only achieves the highest accuracy (69.16%) in cross-domain tasks (in 5$^{\text{th}}$ column), but also is least affected by domain shift (in 6$^{\text{th}}$ column). Compared with other task-specific and weight generated methods (1$^{\text{th}}$ to 4$^{\text{th}}$ rows), T3S

[4]The priority of results presented in this paper: open source models >reported results >open source code >our re-implementation.

Table 1: Classification accuracy on four cross domain experiments. The best one in red and the second one in blue. CUB/Cars/-Places/EurSAT means using *mini*ImageNet to train and CUB/Cars/Places/EurSAT to test. FT means fine-tuning process.

*mini*ImageNet→		5-way 1-shot (%)				5-way 5-shot (%)			
		CUB	Cars	Places	EuroSAT	CUB	Cars	Places	EuroSAT
MatchingNet(Vinyals et al. 2016)	ResNet10	35.89 ± 0.5	30.77 ± 0.5	49.86 ± 0.8	56.61± 0.8	51.37 ± 0.8	38.99 ± 0.6	63.16 ± 0.8	64.45 ± 0.6
MAML(Finn et al. 2017)	ResNet18	37.81 ± 0.7	28.52 ± 0.6	44.96 ± 0.8	51.14 ± 0.7	51.34 ± 0.7	37.98 ± 0.7	60.44 ± 0.8	71.70 ± 0.7
ProtoNet(Snell et al. 2017)	ResNet18	37.50 ± 0.7	29.50 ± 0.6	46.24 ± 0.8	57.34 ± 0.7	62.02 ± 0.7	43.53 ± 0.7	67.83 ± 0.8	73.29 ± 0.7
RelationNet(Sung et al. 2018)	ResNet10	42.44 ± 0.8	29.11 ± 0.6	48.64 ± 0.9	50.59 ± 0.8	57.77 ± 0.7	37.33 ± 0.7	63.32 ± 0.8	61.31 ± 0.7
GNN(Satorras and Estrach 2018)	ResNet10	45.69 ± 0.7	31.79 ± 0.5	53.10 ± 0.8	54.48 ± 0.8	62.25 ± 0.7	44.28 ± 0.6	70.84 ± 0.6	70.35 ± 0.6
MetaOpt(Lee et al. 2019)	ResNet10	44.09± 0.7	32.39± 0.6	49.61± 0.6	61.65 ± 0.7	54.67 ± 0.6	45.90 ± 0.5	65.83 ± 0.6	64.44 ± 0.7
LEO(Rusu et al. 2019)	WRN-28	43.33 ± 0.2	29.80± 0.1	48.14± 0.2	51.89 ± 0.3	61.34 ± 0.2	46.80 ± 0.2	70.05 ± 0.3	67.24 ± 0.2
Baseline w/ FT	ResNet10	44.37 ± 0.6	31.20 ± 0.7	52.38 ± 0.8	61.30 ± 0.8	64.97 ± 0.7	47.77 ± 0.8	71.82 ± 0.7	75.69 ± 0.7
LGM(Li et al. 2019)	ResNet10	44.57± 0.8	31.66± 0.6	53.72± 0.9	60.34 ± 0.8	60.85 ± 0.7	39.20 ± 0.7	66.72 ± 0.8	70.15 ± 0.7
TADAM(Oreshkin et al. 2018)	ResNet12	40.15± 0.4	29.67± 0.4	50.42± 0.4	55.85 ± 0.4	60.22 ± 0.4	43.67± 0.5	70.83 ± 0.5	66.12± 0.5
FEAT(Ye et al. 2020)	ResNet18	42.37 ± 0.3	30.83 ± 0.3	53.99 ± 0.3	57.93 ± 0.4	64.78 ± 0.3	45.42 ± 0.3	71.53 ± 0.4	76.00 ± 0.3
LFT-GNN(Tseng et al. 2020)	ResNet10	47.47 ± 0.7	31.61 ± 0.5	55.77 ± 0.8	64.00 ± 0.8	66.98 ± 0.7	44.90 ± 0.6	73.94 ± 0.7	73.40 ± 0.5
T3S (Ours)	ResNet10	**45.92 ± 0.8**	**33.22 ± 0.8**	**55.83 ± 0.8**	**65.73 ± 0.8**	**69.16 ± 0.9**	**49.82 ± 0.9**	**76.33 ± 0.8**	**79.36 ± 0.8**

Table 2: Comparison with related SOTA on 5-way 5-shot tasks. TS/WG/SS means task-specific/weight generating/ self-supervision based methods; CD/FS represents cross-domain/few-shot task. $\downarrow \Delta\%$ represents the degradation.

	Keywords	Methods	mini	CUB	$\downarrow \Delta\%$
1	TS-FS	TADAM(Oreshkin et al. 2018)	76.70%	60.22%	16.48
2		FEAT (Ye et al. 2020)	82.05%	64.78%	17.27
3	WG-FS	LGM (Li et al. 2019)	71.18%	60.85%	10.39
4		AWGIM (Guo et al. 2020)	78.40%	57.69%	20.71
5	SS-FS	BF3S (Gidaris et al. 2019)	79.87%	48.04%	31.83
6		InfoMax (Chen et al. 2019a)	81.15%	62.73%	18.42
7	SS-CD	Rotation (Xu et al. 2019)	74.62%	49.35%	25.27
8		Jigsaw (Carlucci et al. 2019)	77.89%	42.98%	34.91
9	TL-SS	T3S(Ours)	78.66%	**69.16%**	**9.50**

Table 3: Ablation studies: the influence of TL-SS. IA/CP represent IA/CP represent Instance-level Augmentation and Class-level Permutation respectively in Task Variant.

	Encoder	TL-SS		mini→CUB	Convergence
		Task Variant	Loss		
1	Matching	-	-	51.37%	50K
2		IA	$\mathcal{L}_m$	57.40%	50K
3	LGM	-	-	60.85%	240K
4		IA	$\mathcal{L}_m$	63.41%	240K
5	HAE	-	-	64.87%	1,650K
6		IA	$\mathcal{L}_m$	66.55%	1,650K
7		IA+CP	$\mathcal{L}_m$	68.30%	1,650K
8		IA+CP	$\mathcal{L}_m + \mathcal{L}_t$	69.16%	850K

Table 4: Ablation studies: the influence of 3D CNN in HAE.

Encoder	Task context modeling		mini → CUB
	3D CNN	Average	
MatchingNet	-	-	57.40%
LGM	-	√	63.41%
HAE	√	-	68.30%

is superior in both accuracy and robustness. It is noteworthy that traditional **instance-level self-supervision** (5th to 8th rows) is not friendly to heterogeneous generalization task due to excessive attention to the seen categories and distributions.

Ablation Study

We now present experiments confirming our main claims: 1) TL-SS strategy is able to learn stable concepts and hence achieves better generalization. 2) HAE can generate suitable parameters according to current task context. Note that all ablation studies are conducted on 5-way 5-shot *mini*ImageNet→CUB task.

The effectiveness of TL-SS (Table 3). When adding task-level self supervision $\mathcal{L}_m$, a significant progress (average 3.5 percents) could be observed by using various encoder (2th, 4th and 6th rows). At the same time, class-level permutation in task variant can bring additional improvements (7th row). However, as a byproduct, our meta-learning framework with 3D CNN also brings additional training cost (last column). But it should also be emphasized that the task consistency loss $\mathcal{L}_t$ can greatly accelerate the convergence speed (almost cut in half), and slightly improve the performance (8th row).

The effectiveness of HAE (Table 4). Compared with MatchingNet, the task-specific method LGM shows a huge promotion on the cross-domain task. Furthermore, our HAE arrives at 64.87%@accuracy, outperforming LGM with a 4% improvement by learning robust task context.

Visualizing local discrimination. To demonstrate our assumptions and claims, we also conduct a visualization experiments using t-SNE and take "Baseline" as comparison.

We can draw following conclusion from this figure: 1) Under traditional FSL scope, models enjoy the benefit of pre-training process. They are able to distinguish different classes (left in the figure) due to the absence of domain shift. 2) When testing on another domain, the features from baseline are confused, indicating that without T3S, the model can't mitigate the domain discrepancy and hence leads to inferior performance (Fig. 5(a), (b)). 3) When we project the data from multi target domains together (Fig. 5(c)), T3S is able to preserve local feature discrimination without losing domain information. 4) We also track the subspace changing of three certain tasks during training (Fig. 5(c1,c2)). We can observe TL-SS can maintain the stability (fainter subspace

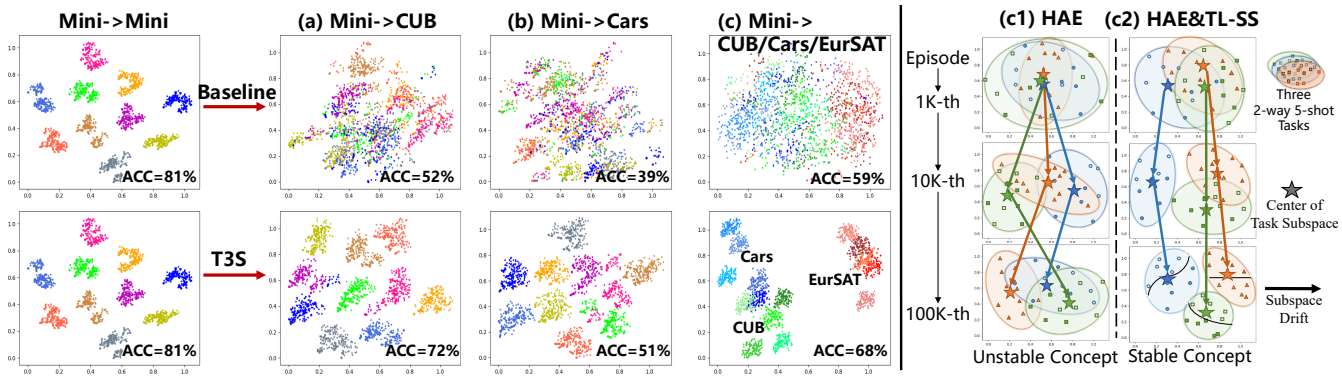

Figure 5: In (a,b), different colors represent different classes. In (c), three colors represent three datasets. In (c1,c2), We track the subspace changing of three certain tasks (3 colors) during training. "1K-th" indicates the thousandth episode.

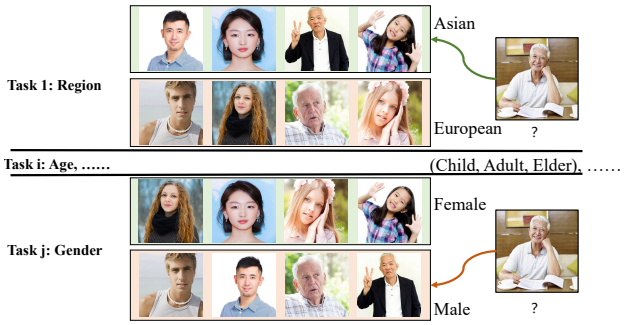

Figure 6: A generalized task-agnostic test. Because the classification angles of unseen tasks are uncertain, specific samples are classified into different clusters in different tasks.

drift) and discrimination of each task subspace, and hence shows better generation ability.

A Generalized Task-agnostic Test

Typically speaking, few-shot learning is a task-agnostic test. That is the categories and distribution of test samples are unpredictable. But from a broader perspective, we argue the classification criterion should also be unpredictable and this problem setting is very usual in real-world applications. For example in Fig. 6, an "Asian grandpa" should be classified into upper (Asian) in task 1 but bottom (Male) in task 2 due to different classification criteria [5]

To mimic this situation, we construct a new few-shot dataset for testing (release later). Each sample in this dataset have three labels in terms of "age, region and gender". For each test episode, we only classify the sample according to a random criterion. There are totally 918 images resized as 84×84 in the test dataset. We design this exam to mainly illustrate two points: 1) Task-level self-supervision avoids the model overfitting to the seen categories/domain. 2) The task context modeled by HAE is more robust.

[5]This task is completely different from the multi-label classification. In the multi-label task, classification criteria are multiple, but known and determined. However, in our task, each classification is carried out from only one criterion, but it is unknown and never seen in the training phase.

Table 5: Results of the generalized task-agnostic test.

5-shot	Age (3 way)	Region (3 way)	Gender (2 way)
Random	*33.33%*	*33.33%*	*50.00%*
Baseline(Chen et al. 2019b)	60.88%	49.52%	**61.67%**
Baseline++(Chen et al. 2019b)	57.36%	50.79%	57.81%
MatchingNet(Vinyals et al. 2016)	51.80%	48.93%	54.79%
ProtoNet(Snell et al. 2017)	55.57%	**51.47%**	55.30%
MAML(Finn et al. 2017)	49.62%	45.31%	52.11%
RelationNet(Sung et al. 2018)	56.52%	49.26%	55.63%
LGM(Li et al. 2019)	57.39%	47.65%	55.28%
LFT-GNN(Tseng et al. 2020)	**60.94%**	50.15%	56.84%
T3S (Ours)	**63.42%** (↑2.5)	**56.09%** (↑4.6)	**69.52%** (↑7.8)

We conduct 5-shot experiments for evaluation and the result is unexpected (Table 5). Taking gender classification (2-way) as an example, most FSL methods can only reach 55%@accuracy, seeming that there is no improvement to random guessing. More importantly, none of them can stand out in all three exams simultaneously. It appears that a fixed feature extractor would embed a sample into a fixed location, regardless of its task-specific conditioning. This will result in significant over-fitting phenomena. Although LGM (Li et al. 2019) also employs a weight generator, due to its simple task context modeling method (average), it appears to produce the changeless parameters in this experiments. This also verifies the advantages of our T3S and hence it achieves the best results in all these three tasks.

Conclusion

In this paper, we for the first time propose a task-level self-supervised framework to solve cross-domain few-shot learning. The core idea of our approach lies in constructing adaptive feature subspace and ensure the stability and discrimination of each subspace via task-level self-supervision. To that end, a high-order associate encoder is proposed to make task context more robust. Extensive experiments demonstrate that the proposed T3S have obvious improvement over the state-of-the-arts methods.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 72192821,61972157), National Key Research and Development Program of China (No. 2019YFC1521104), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Zhejiang Lab (No. 2020NB0AB01), Shanghai Science and Technology Commission (21511101200), Art major project of National Social Science Fund (I8ZD22).

References

Andrychowicz, M.; Denil, M.; Colmenarejo, S. G.; Hoffman, M. W.; Pfau, D.; Schaul, T.; and de Freitas, N. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, 3981–3989.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain Generalization by Solving Jigsaw Puzzles. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2229–2238.

Chen, D.; Chen, Y.; Li, Y.; Mao, F.; He, Y.; and Xue, H. 2019a. Self-Supervised Learning For Few-Shot Image Classification. *CoRR*, abs/1911.06045.

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019b. A Closer Look at Few-shot Classification. In *Proceedings of the 7th International Conference on Learning Representations, ICLR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR)*, 248–255. IEEE Computer Society.

Dong, N.; and Xing, E. P. 2018. Domain adaption in one-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Dou, Q.; de Castro, D. C.; Kamnitsas, K.; and Glocker, B. 2019. Domain Generalization via Model-Agnostic Learning of Semantic Features. In *Advances in Neural Information Processing Systems*, 6447–6458.

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

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2019. Boosting Few-Shot Visual Learning With Self-Supervision. In *Proceedings of IEEE/CVF International Conference on Computer Vision, ICCV*, 8058–8067.

Guo, Y.; Cheung, N.; Guo, Y.; and Cheung, N. 2020. Attentive Weights Generation for Few Shot Learning via Information Maximization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 13496–13505. IEEE.

Guo, Y.; Codella, N. C. F.; Karlinsky, L.; Smith, J. R.; Rosing, T.; and Feris, R. S. 2019. A New Benchmark for Evaluation of Cross-Domain Few-Shot Learning. *CoRR*, abs/1912.07200.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Kim, J.; Kim, T.; Kim, S.; and Yoo, C. D. 2019. Edge-Labeling Graph Neural Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 11–20.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of 2nd International Conference on Learning Representations, ICLR*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Larsson, G.; Maire, M.; Shakhnarovich, G.; and Shakhnarovich, G. 2016. Learning Representations for Automatic Colorization. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *14th European Conference on Computer Vision, ECCV*, volume 9908 of *Lecture Notes in Computer Science*, 577–593.

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

Li, H.; Dong, W.; Mei, X.; Ma, C.; Huang, F.; and Hu, B. 2019. LGM-Net: Learning to Generate Matching Networks for Few-Shot Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97, 3825–3834.

Motiian, S.; Jones, Q.; Iranmanesh, S. M.; and Doretto, G. 2017. Few-Shot Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*, 6670–6680.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Proceedings of the 14th European Conference Computer Vision, ECCV*, volume 9910, 69–84.

Oreshkin, B. N.; López, P. R.; Lacoste, A.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems, NIPS*, 719–729.

Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2536–2544.

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

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, ICLR*.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. P. 2016. Meta-Learning with Memory-Augmented Neural Networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML*, volume 48, 1842–1850.

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

Snell, J.; Swersky, K.; Zemel, R.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 4077–4087.

Su, J.; Maji, S.; Hariharan, B.; and Hariharan, B. 2019. When Does Self-supervision Improve Few-shot Learning? *CoRR*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; 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.

Sung, F.; Zhang, L.; Xiang, T.; Hospedales, T. M.; and Yang, Y. 2017. Learning to Learn: Meta-Critic Networks for Sample Efficient Learning. *CoRR*.

Teshima, T.; Sato, I.; Sugiyama, M.; and Sugiyama, M. 2020. Few-shot Domain Adaptation by Causal Mechanism Transfer. In *Proceedings of the 36th International Conference on Machine Learning, ICML*.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of IEEE International Conference on Computer Vision, ICCV*, 4489–4497.

Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Evci, U.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Manzagol, P.; and Larochelle, H. 2020. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In *Proceedings of the 8th International Conference on Learning Representations, ICLR*.

Tseng, H.; Lee, H.; Huang, J.; and Yang, M. 2020. Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation. In *Proceedings of the 8th International Conference on Learning Representations, ICLR*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Proceedings of the Advances in neural information processing systems*, 3630–3638.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 3733–3742.

Xu, J.; Xiao, L.; López, A. M.; and López, A. M. 2019. Self-Supervised Domain Adaptation for Computer Vision Tasks. *IEEE Access*, 7: 156694–156706.

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.

Zhou, B.; Lapedriza, À.; Khosla, A.; Oliva, A.; and Torralba, A. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell, (PAMI).*, 40(6): 1452–1464.