Bi-level Meta-learning for Few-shot Domain Generalization

Xiaorong Qin[1,2], Xinhang Song[1,2], Shuqiang Jiang[1,2]
[1]Key Lab of Intelligent Information Processing Laboratory of the Chinese Academy of Sciences (CAS),
Institute of Computing Technology, Beijing [2]University of Chinese Academy of Sciences, Beijing
{xiaorong.qin, xinhang.song}@vipl.ict.ac.cn
sqjiang@ict.ac.cn

Abstract

The goal of few-shot learning is to learn the generalization from seen to unseen data with only a few samples. Most previous few-shot learning methods focus on learning the generalization within particular domains. However, the more practical scenarios may also require the generalization ability across domains. In this paper, we study the problem of few-shot domain generalization (FSDG), which is a more challenging variant of few-shot classification. FSDG requires additional generalization with larger gap from seen domains to unseen domains. We address FSDG problem by meta-learning two levels of meta-knowledge, where the lower-level meta-knowledge is domain-specific embedding spaces as subspaces of a base space for intra-domain generalization, and the upper-level meta-knowledge is the base space and a prior subspace over domain-specific spaces for inter-domain generalization. We formulate the two levels of meta-knowledge learning problem with bi-level optimization, and further develop an optimization algorithm without higher-order derivative information to solve it. We demonstrate our method is significantly superior to the previous works by evaluating it on the widely used benchmark Meta-Dataset.

1. Introduction

Traditional few-shot classification addresses the problem of learning to classify unseen classes through knowledge of seen classes, and it is based on the $i.i.d.$ assumption, corresponding to the practice that test (unseen) classes are consistently sampled from the same dataset (domain) as training (seen) classes. We refer to this generalization across classes of the same domain as intra-domain generalization. In real-life, how to achieve further generalization when facing unseen data from unknown domains may be a problem with more practical interest.

The recently attention-grabbing problem few-shot domain generalization (FSDG) [5, 6, 28, 50] is to learn a universal model via multiple training domains (*e.g.*, cat breeds, traffic signs and textures) for good generalization to novel classes from a wide range of domains, both in- and out-of-domain. FSDG learns novel classes across domains, and solves few-shot classification problem at an upper level involving more general domain distribution. It is inter-domain generalization, which is downward compatible with intra-domain generalization.

Typically, the generalization ability of previous few-shot works is obtained by virtue of meta-knowledge [15, 53] learned through meta-learning. ProtoNet [48], a meta-learning metric-based method, learns an embedding space as meta-knowledge, in which instances gather around a prototype representation for each class. Such meta-knowledge is learned with the assumption that all data comes from one domain, and the generalization ability is required within the domain. [19] shows that ProtoNet is only better adapted to intra-domain generalization of a single domain, but leads to a negative gain in inter-domain generalization based on multiple known domains. It is manifested in that the model co-learned by multi-domain data is worse than respective models trained by each single domain on in-domain test data, and that the generalization to out-of-domain data is also poor. Most previous works [6, 11, 44] on FSDG tackle the challenges by learning a flexible embedding space with task-adaptive modules based on ProtoNet, which can be seen as learning meta-knowledge that is more generalized to domain-agnostic data than that of ProtoNet. But we first argue that such a stronger meta-knowledge learned in previous works is not effective enough for solving simultaneously in- and out-of-domain tasks with larger and smaller generalization gaps respectively , and second that the challenges about the negative gain and poor generalization arise because the meta-knowledge learning is not considered at two levels, inter- and intra-domain levels, where inter-domain meta-knowledge applicable to all domains is more meta-level than intra-domain meta-knowledge.

To this end, to solve inter-domain generalization compatible with intra-domain generalization, we propose to

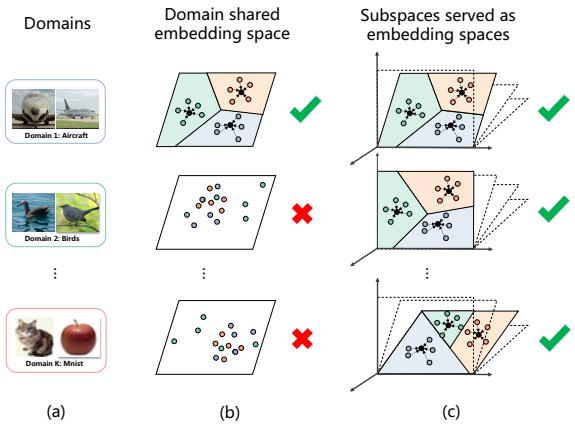


Figure 1. (a) Images from different domains in Meta-Dataset [51]. (b) A shared embedding space for all domains. (c) Subspaces of a base space as domain-specific embedding spaces.

learn two levels of meta-knowledge, where lower-level intra-domain meta-knowledge is built on upper-level inter-domain meta-knowledge. Specifically, we treat a base feature space as inter-domain meta-knowledge, and then, we model the lower-level intra-domain meta-knowledge with low-rank subspaces of the base feature space, *i.e.*, intra-domain embedding spaces, due to the variability of discriminative features across domains, as shown in Figure 1(c). For a test task from one domain, we need to find the optimal domain-specific embedding space based on the base space for classification. Naturally, we use a bi-level optimization framework [9] for it, where the upper layer learns the base feature space and the lower layer learns a subspace for each specific domain. In addition, for fast adaptation or over-fitting prevention on specific domains in the lower level, we not only learn small-capacity subspace orthogonal projections as subspace representation to project base features, but also propose another inter-domain meta-knowledge, namely a prior subspace. It is used for regularization shared across domains such that the lower optimization can converge rapidly to the optimal solution on each domain. We use projection metric on the Grassmann manifold in the regularization term. How to solve this optimization problem with multi-domain data? We develop an algorithm without higher-order derivative information to learn the upper-layer parameters. Through such a mechanism, a more general base feature extractor and a prior subspace are learned, as well as specific subspaces for all seen domains.

2. Related works

Few-shot classification Early few-shot classification methods are mainly divided into metric-based [41, 46, 49, 57] and optimization-based [3, 10, 16, 20, 26, 29, 31] methods. The most representative works are ProtoNet [48] and MAML [12], respectively. ProtoNet tackles the problem by learning a universal embedding space for a specific domain such that the distance between images on the embedding space is consistent with their semantic distance, but it cannot adapt to the problem of multi-domain learning and generalization. Our work proposes an upper-level meta-knowledge over embedding spaces for multiple domains. MAML learns a group of initialization parameters suitable for all tasks through bi-level optimization mechanism. In our method, we also use the mechanism, but we learn two levels of meta-knowledge, both inter- and intra-domain.

Multi-domain learning The goal of multi-domain learning in the visual field is to train a single model to handle multiple visual domains. If data per domain is relatively small and the domains are similar, this common learning of a single model will improve performance across domains than training a separate model for each domain [30]. However, when there is more data and differences between domains are significant, the shared model trained by multiple domains does not perform as well as the individual model of each domain [42, 55]. To solve this problem, most of the traditional methods in multi-domain learning develop a tunable deep network architecture with shared parameters and domain-specific parameters, where shared parameters are generally *universal representations* [42](*i.e.* feature extractor) learned by all domains together, and domain-specific parameters are generally small-volume modules learned by each domain data separately, such as adapter residual modules in [42], serial and parallel residual adapters in [43] and feature critic network in [30].

Few-shot domain generalization Recently, some communities have made different attempts based on the challenging FSDG. The first type is to learn multiple single-domain models and adaptively select features from multiple model outputs for classification by a feature selection strategy during meta-testing. SUR [11] automatically learns the weights of multiple outputs for feature selection based on the support set and URT [33] proposed a meta-learning layer that can dynamically re-weights and composes the most appropriate domain-specific representations in the meta-test phase. The methods have good performance and are simple to implement, but in them, the model storage is large and multiple forward passes lead to high computational efficiency. Another common class of approaches are to learn a shared network with small-capacity task-specific adaptors, whose parameters are usually finetuned or predicted through a meta auxiliary network conditioned on the support set. CNAPS [44] uses a pre-trained feature extractor augmented with FiLM layers [40] and a few-shot adaptive classifier based on conditional neural processes (CNP) [13]. The FiLM layers are adapted for each task using support

images. URT [33] learns a shared feature extractor and respective FiLM layers for seen domains. It can provide the initialization of FiLM parameters for new tasks by convexly combining those parameters of seen domains at meta-test time. In addition to the above methods, URL [27] learns a universal feature extractor by distilling knowledge of multiple domain-specific models. TSA [28] mainly discusses the ways to quickly adapt during meta-testing. It attaches a parametric transformation to each layer based on the existing meta-training models like URL just mentioned. The transformation can be constructed by a serial or a residual topology and can be parameterized with matrix multiplication or channel-wise scaling.

Subspace learning for few-shot learning [4, 47] learn a subspace of the full feature space for each class and compute the projection distances of testing images in each subspace for classification. Our work is to learn a more discriminative subspace for each specific domain. [2] makes the weight vectors of the new classes close to those of the old classes by subspace regularization. However, we bring the subspace representations of different domains close to a central representation. Although [58] also learns a discriminant subspace, we learn intra-domain subspaces and inter-domain subspace based on the geometric relationship.

3. Problem setting and preliminaries

Few-shot classification aims at learning a model by old classes to identify novel classes with a lack of labeled data in a meta-learning paradigm. FSDG inherits the objective and paradigm, except that the generalization gaps across classes are different. In this section, we first review the setting and terminology of few-shot classification. Then, we introduce non-episodic meta-training related to our proposed method, and describe our focus problem, FSDG.

Few-shot classification setting Given a single domain D, we have base classes from D, and expect to learn such a model in the meta-training phase that can be rapidly adapted to identifying novel classes with learned meta-knowledge. The core idea of most current popular few-shot methods is the episodic learning simulating few-shot scenarios. In few-shot studies, meta-training and meta-testing data consist of a large number of tasks (episodes) $\{\mathcal{T}_t^{train}\}_{t=1}^{M}$ and $\{\mathcal{T}_t^{test}\}_{t=1}^{M}$ sampled from base classes and novel classes of D, respectively. Each task $\mathcal{T}$ contains a support set S and query set Q, where S is used for fast adaptation. The well-known N-way K-shot tasks mean that their support sets have N base classes, each of which has K instances, *i.e.*, $S = \{\{(x_i^j, y_i^j)\}_{i=1}^{K}\}_{j=1}^{N}$, where (x_i^j, y_i^j) represent an image and its corresponding label. In order to verify the generalization ability of the model, it is required that base and novel classes do not overlap.

Non-episodic learning in meta-training Most few-shot methods train a meta-learner by episodic learning as mentioned above, even though base classes are indeed many-shot. It has been demonstrated that in the case of many-shot datasets, episodic learning is not necessary. It can lead to a data-inefficient way of exploiting training batches [22], and the feature extractor trained by the mechanism is inferior to that trained in classical multi-class classifier [56]. Accordingly, we follow the non-episodic learning, that is our model will be directly trained in the conventional way of learning a standard multi-class classifier rather than by learning a large number of episodes.

Few-shot domain generalization As a variant of few-shot classification, its challenge lies in more multi-level generalization gaps, both in- and out-of-domain. In particular, given multiple domains $D_1, D_2, ..., D_N$, we need to solve in-domain generalization tasks from novel classes of $D_1, D_2, ..., D_N$, and out-of-domain generalization tasks from unseen domains in the meta-testing phase.

4. Proposed method

For FSDG problem involving multi-level generalization gaps owing to in- and out-of-domain generalization tasks, we propose two levels of meta-knowledge learning relying on bi-level optimization, where the upper optimization learns upper-level inter-domain meta-knowledge and the lower optimization learns lower-level intra-domain meta-knowledge. First, we show the defined notation in §4.1 used in our method. Then in §4.2, we introduce the projection metric of subspaces on the Riemannian manifold and use it to measure the geodesic distance between two subspaces in our regularization norm. In §4.3, we will introduce our meta-problem relying on a bi-level optimization formulation to learn the two levels of meta-knowledge, and we develop an optimization algorithm without higher-order derivative information to solve it.

4.1. Notation definition

In our model as Figure 2, a randomly sampled image $\mathbf{x}$ from a random domain D_k passes through the base feature extractor $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ to be a n-dimensional vector $f_{\boldsymbol{\theta}}(\mathbf{x})$. We define the space spanned by all possible $f_{\boldsymbol{\theta}}(\mathbf{x})$ as a base feature space, also known as inter-domain meta-knowledge for generalization across domains. Base on the base feature space, we expect to learn domain-specific embedding spaces as intra-domain meta-knowledge for in-domain generalization. We model domain-specific embedding spaces using subspaces of the base feature space, so we need to learn orthogonal projection $\mathbf{P}_k \in \mathbb{R}^{n \times n}$ for specific

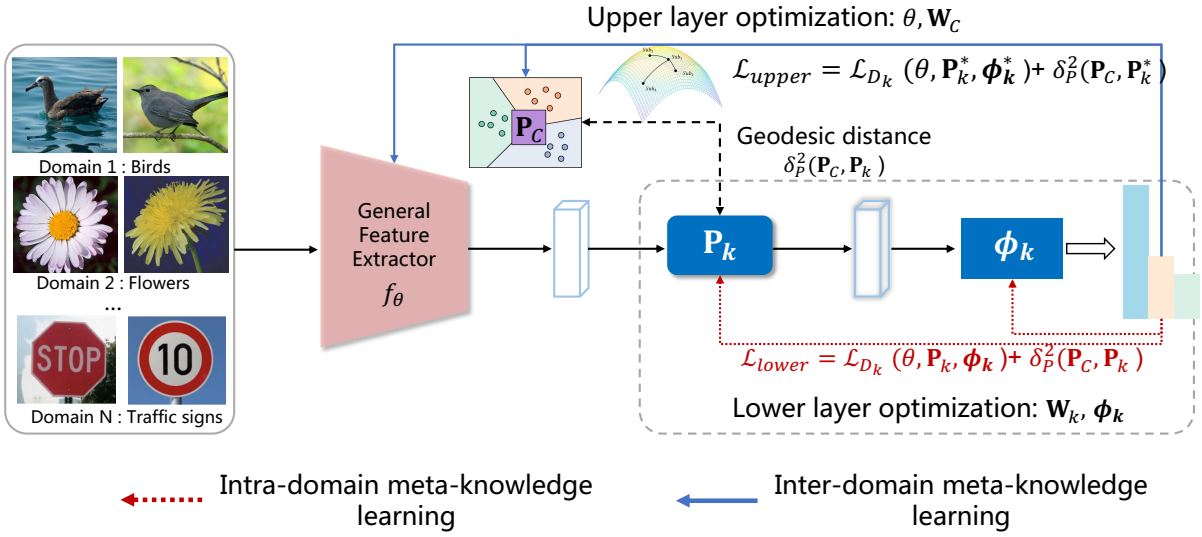


Figure 2. The overall framework of our meta-problem. The lower-layer optimization performs intra-domain meta-knowledge learning and upper-layer optimization aims at learning inter-domain meta-knowledge.

subspace $\mathcal{S}_k$ about D_k and then project $f_{\boldsymbol{\theta}}(\mathbf{x})$ onto $\mathcal{S}_k$ to be $\mathbf{P}_k f_{\boldsymbol{\theta}}(\mathbf{x})$, which is finally taken in the domain classifier h_{ϕ_k} with parameters ϕ_k of D_k.

In addition to inter-domain meta-knowledge $f_{\boldsymbol{\theta}}$, we propose a prior subspace with projection $\mathbf{P}_C$ over all domain-specific subspaces, used in a regularization term of domain-specific subspace learning as another inter-domain meta-knowledge for fast adaptation of intra-domain knowledge.

4.2. Projection metric on Grassmann manifold

The set of all m-dimensional linear subspaces of D-dimensional space $\mathbb{R}^D$ $(0 < m \leq D)$ is not a Euclidean space, but a Riemannian manifold known as Grassmann manifold $\mathcal{G}(m, D)$, which is a smooth surface embedded in a high-dimensional Euclidean space. [1, 25] discuss the geometric properties and structure of Grassmann manifolds. With its smooth characteristic, the distance between two subspaces is a geodesic distance. In addition, subspace orthogonal projections can represent elements on a Grassmann manifold one-to-one. For the convenience of taking derivatives in optimization, we will use the projection metric δ_P^2 [14] as our subspace metric, that is popular and approximates the geodesic metric on $\mathcal{G}(m, D)$. The δ_P^2: $\mathcal{G}(m, D) \times \mathcal{G}(m, D) \to \mathbb{R}^+$ is defined as

$$\begin{aligned} \delta_P^2(U, V) &= \|\mathbf{P}_U - \mathbf{P}_V\|_F^2 \\ &= tr[(\mathbf{P}_U - \mathbf{P}_V)^\top (\mathbf{P}_U - \mathbf{P}_V)], \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and U, V are two subspaces with orthogonal projections $\mathbf{P}_U, \mathbf{P}_V$.

4.3. Meta-problem relying on bi-level optimization

Meta-problem formulation Then we introduce our optimization problem with (1). Given a domain distribution $\mathcal{D}$, we hope to learn a base feature extractor $f_{\boldsymbol{\theta}}$ and a prior projection $\mathbf{P}_C$ of a prior subspace so that when facing with a new domain D_k sampled from $\mathcal{D}$, D_k-specific subspace projection $\mathbf{P}_k$ around the prior and the domain-specific multi-class classifier h_{ϕ_k} taking in feature vectors projected can be found by the following learning:

$$\min_{\mathbf{P}_k, \phi_k} \mathcal{L}_{D_k}(\boldsymbol{\theta}, \mathbf{P}_k, \phi_k) + \frac{\lambda}{2}\|\mathbf{P}_k - \mathbf{P}_C\|_F^2, \ s.t.\ \mathbf{P}_k \in \mathcal{A}, \tag{2}$$

where $\mathcal{L}_{D_k}$ is the empirical classification risk (*i.e.*, cross entropy cost function) for the current domain D_k, and $\mathcal{A}$ is the set whose elements satisfy the properties of $n \times n$ orthogonal projections [37] and the same below.

And then, we naturally use a meta-learning approach relying on bi-level optimization to solve the base feature extractor and prior learning problem, which takes the form stochastically:

$$\begin{aligned} &\min_{\boldsymbol{\theta}, \mathbf{P}_C} \mathbb{E}_{D_k \sim \mathcal{D}}[\min_{\mathbf{P}_k, \phi_k} \mathcal{L}_{D_k}(\boldsymbol{\theta}, \mathbf{P}_k, \phi_k) + \frac{\lambda}{2}\|\mathbf{P}_k - \mathbf{P}_C\|_F^2] \\ &s.t.\ \mathbf{P}_C \in \mathcal{A}, \end{aligned} \tag{3}$$

The goal of the lower level learning in (3) is to find the optimal domain-specific projection matrix around the prior $\mathbf{P}_C$ and ϕ_k, while in the upper level, the models tune $\boldsymbol{\theta}$ and $\mathbf{P}_C$ with a series of biased $\{\phi_k, \mathbf{P}_k\}_{k=1}^N$ for some sampled domains such that $\mathbf{P}_C$ is geometrically closer to all in $\{\mathbf{P}_k\}_{k=1}^N$.

In practice, we have some diverse domains (datasets in experiments) $\{D_k\}_{k=1}^N$ drown from $\mathcal{D}$. Using them, our aims are to learn inter-domain meta-knowledge $f_{\boldsymbol{\theta}}$ and $\mathbf{P}_C$, as well as intra-domain meta-knowledge $\{\mathbf{P_k}\}_{k=1}^N$ for the seen domains. Notably, learning orthogonal projections of the subspaces in (3) directly will result in a constrained optimization problem, so for simplicity, that is done by learning a set of basis vectors $\mathbf{W} \in \mathbb{R}^{n\times m}$ of a $m-$rank subspace ($m \leq n$, is also a hyperparameter) for each domain and further generating an orthogonal projection matrix $\mathbf{P}$ by $g(\mathbf{W})$:

$$g(\mathbf{W}) = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top. \tag{4}$$

The final empirical bi-level optimization problem called 2LM for our objective is:

$$\min_{\boldsymbol{\theta},\mathbf{W}_C} \sum_{k=1}^{N} \min_{\mathbf{W}_k,\phi_k} \{\mathcal{L}_{D_k}(\boldsymbol{\theta}, \mathbf{W_k}, \phi_k) + \frac{\lambda}{2}\|g(\mathbf{W}_k) - g(\mathbf{W}_C)\|_F^2\}. \tag{5}$$

Meta-optimization We use stochastic gradient descent (SGD) method for our meta-problem. In the lower-layer learning, assume that we can find its optimal parameters$(\{\mathbf{W}_k^*, \phi_k^*\})$ approximately after lower optimization for each iteration. Then we can perform SGD on meta-parameters $(\boldsymbol{\theta}, \mathbf{W}_C)$ avoiding higher-order derivative information computing, and we establish the following simple lemma to explain it. First, we define $F_k(\boldsymbol{\theta}, \mathbf{W}_C) = \min_{\mathbf{W}_k,\phi_k} \mathcal{L}_{D_k}(\boldsymbol{\theta}, \mathbf{W}_k, \phi_k) + \frac{\lambda}{2}\|g(\mathbf{W}_k) - g(\mathbf{W}_C)\|_F^2$.

Lemma 4.1. *Assume that $\mathcal{L}_{D_k}$ is differentiable and $(\mathbf{W}_k^*, \phi_k^*)$ is the unique minimizer of $F_k(\boldsymbol{\theta}, \mathbf{W}_C)$. Then the gradient components of the meta-loss F_k with respect to $\boldsymbol{\theta}$ and $\mathbf{W}_C$ are given by $\frac{\partial F_k}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}_{D_k}(\boldsymbol{\theta}, \mathbf{W}_k^*, \phi_k^*)}{\partial \boldsymbol{\theta}}$ and $\frac{\partial F_k}{\partial \mathbf{W}_C} = \lambda(\frac{\partial g(\mathbf{W}_C)}{\partial \mathbf{W}_C})^\top[g(\mathbf{W}_C) - g(\mathbf{W}_k^*)]$, which are no higher-order derivative information.*

The lemma's proof is provided in Appendix. The lemma lays the groundwork for our meta-optimization and shows the gradient expression of the meta-loss in (5) with respect to $\boldsymbol{\theta}$ and $\mathbf{P}_C$ when using mini-batch proximal learning in the lower level optimization. At the t-th iteration, we sample mini-batch images for each domain $\{\mathbf{X}_k^{(t)}\}_{k=1}^N$ and then carry out mini-batch proximal learning for the lower-level parameters. Here we expect to find its optimal parameters $\{(\mathbf{W}_k^{(t)})^*, (\phi_k^{(t)})^*\}_{k=1}^N$, but it is difficult to achieve. We choose an sub-optimal value as a substitute, and to do so, we use a warm-start approach, that is using the result of the previous optimization as the initialization of this optimization, and perform a few gradient updates to get $\{(\mathbf{W}_k^{(t)})^*, (\phi_k^{(t)})^*\}_{k=1}^N$ due to mini-batch data and small capacity nature of these parameters compared to the backbone. According to the lemma 4.1, we can develop an algorithm 1 to optimization (5) without higher-order derivative information.

Algorithm 1 SGD based optimization for 2LM.

Input: N seen domains $\{D_k\}_{k=1}^N$, learning rates α, β
Randomly initialize $\mathbf{W}_C$, θ, $\{\mathbf{W}_k\}_{k=1}^N$,$\{\phi_k\}_{k=1}^N$
while not done **do**
  Sample N mini-batch images $\{(\mathbf{X}_k, \mathbf{Y}_k)\}_{k=1}^N$
  from $\{D_k\}_{k=1}^N$, respectively
  for $(\mathrm{X}_k, \mathbf{Y}_k) \in \{(\mathrm{X}_k, \mathbf{Y}_k)\}_{k=1}^N$ **do**
    Compute a approximate minimizer $\mathbf{W}_k^*$ and ϕ_k^*
    to the lower-level optimization by few-step
    GD update
  end for
  Update upper-level meta parameters by the lemma
  4.1
  $\mathbf{W}_C = \mathbf{W}_C - \alpha\frac{1}{N}\sum_{k=1}^N \frac{\partial F_k}{\partial \mathbf{W}_C}$
  $\boldsymbol{\theta} = \boldsymbol{\theta} - \beta\frac{1}{N}\sum_{k=1}^N \frac{\partial F_k}{\partial \boldsymbol{\theta}}$
end while

4.4. Few-shot tasks adaptation in the meta-testing stage

In the meta-testing phase, we address few-shot tasks as setting in §3, and the few-shot tasks can be drowned from novel classes of the seen domains during meta-training or other unseen domains.

For a task $\mathcal{T}_t$ belonging to the seen domain k, we freeze the general feature extractor $f_{\boldsymbol{\theta}}$ and the k-th domain-specific projection $\mathbf{P}_k$, and then identify query set based on support set. While for a task $\mathcal{T}_t$ of unseen domains, we expect to find an appropriate subspace near the central subspace to prevent overfitting, so we only freeze the general feature extractor and finetune the subspace basis with support set S_t as follows:

$$\min_{\mathbf{W}_t} \{\mathcal{L}_{S_t}(\boldsymbol{\theta}, \mathbf{W}_t) + \frac{\lambda}{2}\|g(\mathbf{W}_t) - g(\mathbf{W}_C)\|_F^2\}. \tag{6}$$

At this stage, we use Nearest-Centroid Classifier (NCC) classifier following those of the previous methods [11, 27, 36, 48, 50]. By (7) an embedding center is calculated for each category of a few-shot task using the support set. And then, the classification probabilities are calculated based on the distances between the query images and embedding centers. The m-th category center in support set S_t is:

$$\mathbf{c}_m = \frac{1}{K}\sum_{k=1}^{K} \mathbf{P} f_{\boldsymbol{\theta}}(\mathbf{x}_k^m), \tag{7}$$

where $f_{\boldsymbol{\theta}}$ and $\mathbf{P}_t$ are the base feature extractor and domain-specific projection generated by the corresponding $\mathbf{W}_t$.

And for an unseen image $\mathbf{x}$, we estimate the probability that it belongs to category m by:

$$p(y=m|\mathbf{x}) = \frac{\exp(d(\mathbf{P}_t f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{c}_m))}{\sum_j \exp(d(\mathbf{P}_t f_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{c}_j))}, \tag{8}$$

where $d(,)$ is the negative cosine similarity.

TSA evaluation TSA [28] specifically studies evaluation methods for meta-testing. It attaches additional adapters like transformation matrices for intermediate feature layer on already meta-trained models (*e.g.*, URL [27]) and updates them by the support sets of new tasks for fast adaptation. During meta-testing, we also apply TSA method to validate the effectiveness of our model than others.

5. Experiments

5.1. Dataset and implementation

Meta-Dataset [51] It is a large-scale benchmark that has been widely used in recent years for few-shot domain generalization through multiple domains. It contains a total of 10 diverse datasets, ImageNet [45], Omniglot [23], Aircraft [35], Birds [54], Textures [8], QuickDraw [18], Fungi [7], VGG Flower [38], Traffic Signs [17] and MSCOCO [32], where the first eight are seen training domains and the last two are unseen testing domains. Following the previous works [27, 28, 50], we also add three additional datasets including MNIST [24], CIFAR10 [21] and CIFAR100 [21], which are used as unseen testing domains. Notably, instead of generating the traditional N-way K-shot few-shot tasks, the benchmark yields realistically imbalanced episodes of variable shots and ways by special sampling procedures [51]. We apply the complex generation approach during meta-testing.

Implementation details We adopt ResNet-18 as the general feature extractor follow the previous few-shot domain generalization works [13, 27, 28]. We use fully connected layers as the classifiers of seen domains during meta-training. And we use $512 \times k$ ($k < 512$, *e.g.*, 384) low-rank matrices as domain-specific subspace basis for generating domain-specific orthogonal projections. To prevent singularity during the process, we replace $(\mathbf{W}^\top \mathbf{W})^{-1}$ in (4) with $(\mathbf{W}^\top \mathbf{W} + \epsilon \mathbf{I})^{-1}$, where ϵ is 10^{-12}. For bi-level optimization in (5), we set $\lambda = 0.001$, and use stochastic gradient descent with momentum as the optimizer and the cosine annealing learning scheduler following the training protocol of SUR [11] in the upper layer. The number of iterations is 240000. While in the lower-layer optimization of each iteration, we use a warm-start approach that is using the last optimization result as the initialization of this optimization, and perform a few stochastic gradient updates for sub-optimal value, *e.g.*, 2 or 5 step. We use the PyTorch package [39] with automatic differentiation in our implementation. We refer to supplementary for more details.

5.2. Evaluation

To evaluate the learned inter-domain meta-knowledge and intra-domain meta-knowledge in our model, we perform 600 testing tasks sampled randomly for each domain of Meta-Dataset. And we show that in- and out-of-domain performance by averaging the results of seen and unseen domains, respectively.

Comparisons with the previous works The models to be compared with ours are shown in the Table 1. First, we introduce two baseline models, SDL and MDL, respectively. SDL represents N models learning N single embedding spaces for N seen domains $\{D_k\}_{k=1}^N$ respectively by optimizing feature extractor $\boldsymbol{\theta}$ and classifier ϕ_k in (9):

$$\min_{\boldsymbol{\theta}, \phi_k} \mathcal{L}_{D_k}(\boldsymbol{\theta}, \phi_k). \tag{9}$$

Similarly, MDL is a multi-domain model trained straightforwardly using the data from N seen domains $\{D_k\}_{k=1}^N$ by optimizing (10):

$$\min_{\boldsymbol{\theta}, \{\phi_k\}_{k=1}^N} \sum_{k=1}^{N} \mathcal{L}_{D_k}(\boldsymbol{\theta}, \phi_k). \tag{10}$$

What's more, we compare to the previous SOTA works, incluing CNAPS [44], ProtoMAML [51], Simple CNAPS [6], SUR [11], URT [33], FLUTE [50], Tri-M [34], Tri CNAPS [5], URL [27], TSA (TSA evaluation on URL) [28]. The last two columns in the table are the results of our method on the Meta-Dataset. 2LM represents our model with our own meta-training and meta-testing method, and 2LM+TSA represents our model applying TSA evaluation method as mentioned in §4.4. It can be seen that our method achieves the state-of-the-art both in- and out-of-domain performance.

Specifically, we can find that despite richer data, MDL does not perform as well as SDL on most datasets, which is due to the difference of embedding spaces across domains.

Our models outperform than MDL on both in- and out-of-domain performance, which proves that our method does effectively learn the base feature space across domains and intra-domain embedding spaces of seen domains, leading to improve inter- and intra-domain generalization. And our models are even superior to SDL of seen domains, so we believe that our upper inter-domain meta-knowledge has a positive impact on lower intra-domain meta-knowledge of seen domains. From the results, it can also be said that the superiority of rich data missed in MDL is excavated by our method. In addition, we can find that the models applying TSA evaluation method including URL and 2LM exhibit

outstanding performance, and we guess that it is related to the addition of an appropriate amount of adaptive parameters, which improves the adaptation space without overfitting. Compared to the last state-of-the-art model URL, which is obtained through knowledge distillation of multiple single-domain models, our model has good scalability for larger datasets, yet set a new state of the art.

Analysis of base feature space learning with intra-domain meta-knowledge We explore the effect of the different subspace projections on the base feature extractor $f_{\boldsymbol{\theta}}$ when performing the upper optimization. First, assume that the outputs of $f_{\boldsymbol{\theta}}$ are directly used as the input of the classifiers across domains. For an instance $\mathbf{x}$ of domain D_k, $\mathbf{y} = f_{\boldsymbol{\theta}}(\mathbf{x})$, and the partial derivatives of classification loss $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ according to the chain rule are:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = (\frac{\partial \mathbf{y}}{\partial \boldsymbol{\theta}})^\top \frac{\partial \mathcal{L}}{\partial \mathbf{y}}. \tag{11}$$

But when we project $\mathbf{y}$ onto specific subspace of D_k by orthogonal projection $\mathbf{P}_k^*$ in the upper layer: $\mathbf{z} = \mathbf{P}_k^* \mathbf{y}$, the partial derivatives become:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = (\frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}})^\top \frac{\partial \mathcal{L}}{\partial \mathbf{z}} = (\frac{\partial \mathbf{y}}{\partial \boldsymbol{\theta}})^\top (\mathbf{P}_k^*)^\top \frac{\partial \mathcal{L}}{\partial \mathbf{z}}. \tag{12}$$

We can note that there is an extra term in the partial derivatives about domain D_k, *i.e.* $(\mathbf{P}_k^*)^\top$, which is learned in the lower layer. We argue that these terms of different domains warp the gradient space to prevent gradient interference and learning instability of feature extractor parameters $\boldsymbol{\theta}$ due to simultaneous learning of different domains, because if the warped gradients of current domain are orthogonal to features of other domains, it will not negatively affect feature learning in other domains. Consequently, our base feature extractor is more general and can learn more wide features. We quantitatively analyze this finding by comparing the experiment with MDL. Specifically, we only use the base feature extractor $f_{\boldsymbol{\theta}}$ of our model and that of MDL to complete few-shot testing with NCC-MD classifiers proposed by [6]. The average results of all domains are 72.60 and 71.38, so our base feature extractor is more expressive.

5.3. Ablation studies

We examine three major determinants of our model: (a) the two levels of meta-knowledge learning setting, (b) the dimension of domain-specific subspaces, *i.e.*, the rank m of basis vector matrix $\mathbf{W}$, and (c) the number of steps for gradient update in the lower layer.

Effectiveness of two levels of meta-knowledge learning setting Depending on the learned meta-knowledge, we build two other comparison models based on 2LM: MDL and 2LM_S. MDL is the vanilla multi-domain model by optimizing (10), which is seen as learning one level of meta-knowledge. MDL deals with both in- and out-of-domain test tasks by learning specific subspaces through support set without regularization.

2LM_S is a model without the inter-domain prior subspace from the upper optimization, and it can learn base feature space and further domain-specific subspaces by (13),

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{N} \min_{\mathbf{W}_k, \phi_k} \mathcal{L}_{D_k}(\boldsymbol{\theta}, \mathbf{W}_k \phi_k), \tag{13}$$

which solves in-domain test tasks like 2LM model does, and out-of-domain tasks by learning specific subspaces without regularization in the same way as MDL. See the results in Table 2. Comparing 2LM_S with MDL, we can find that domain-specific subspaces as intra-domain meta-knowledge of 2LM_S improve the discrimination in their respective domains because 2LM_S outperforms MDL for in-domain generalization, and it promotes the learning of the base feature extractor, as discussed in §5.2, so the results of 2LM_S for out-of-domain generalization are also better than MDL. Furthermore, our model, 2LM has one more kind of inter-domain knowledge namely prior subspace than 2LM_S, and it is superior to 2LM_S overall from the results. This fits with our goal that the prior subspace can be used a regularization to learn a more precise intra-domain projection representation, further contributing to the power of the base feature extractor. Meanwhile, to verify the efficiency of the learned prior subspace of 2LM for out-of-domain generalization, we perform out-of-domain test tasks without the regularization of prior subspace on 2LM and the results are in the last column of Table 2. Compared to 2LM_Noreg, the better out-of-domain performance of 2LM with regularization is in line with our expectation that the prior subspace improves the out-of-domain generalization.

Impacts of the subspace dimension and gradient update step number in the lower-layer optimization In our method, the subspace dimension, *i.e.*, the rank m of the basis vector matrix $\mathbf{W}$ is a hyperparameter. We find that settings with larger-value m give better results, probably because subspaces with higher dimension contain more feature information, while too high values lead to over-fitting. We use 384 in our model. In addition, number of steps n for gradient update in the lower layer is also a hyperparameter. A good choice for n in our model is 2 according to attempts already made. The relevant experimental results are in the Appendix.

5.4. Qualitative results

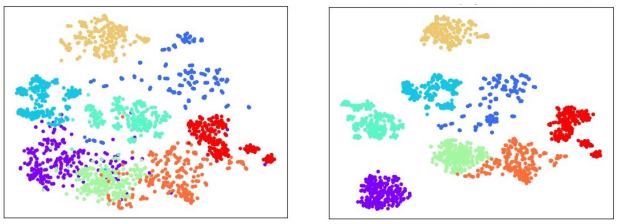
In order to verify the influence of different specific subspace projections on the spatial distribution of feature vectors across domains, we use the t-Distributed Stochastic

Table 1. **Comparisons to baselines and the previous state of the art on Meta-Dataset.** Mean accuracy and 95% confidence interval are reported. All results are obtained during meta-testing phase. The test tasks of the first eight domains seen in the meta-training phase belong to in-domain generalization, while the last five unseen domains are out-of-domain generalization. We also report average accuracy for in- and out-of-domain generalization, as well as overall accuracy for all domains. Following the previous work [27, 51], the average rank of all methods is computed and shown in the table.

Dataset	CNAPS	ProtoMAML	Simple CNAPS	SUR	URT	FLUTE	Tri-M	Tri CNAPS	URL	URL+TSA	MDL	SDL	2LM	2LM+TSA
ImageNet	50.8 ± 1.1	49.5	58.4 ± 1.1	56.2 ± 1.0	56.8 ± 1.1	$\mathbf{58.6 \pm 1.0}$	51.8 ± 1.1	57.9 ± 1.1	57.5 ± 1.1	57.4 ± 1.1	52.3 ± 1.1	55.8 ± 1.0	58.0 ± 3.6	$\mathbf{58.4 \pm 1.6}$
Omniglot	91.7 ± 0.5	63.3	91.6 ± 0.6	94.1 ± 0.4	94.2 ± 0.4	92.0 ± 0.6	93.2 ± 0.5	94.3 ± 0.4	94.5 ± 0.4	95.0 ± 0.4	94.0 ± 0.5	93.2 ± 0.5	$\mathbf{95.3 \pm 1.0}$	$\mathbf{95.4 \pm 0.8}$
Aircraft	83.7 ± 0.6	55.9	82.0 ± 0.7	85.5 ± 0.5	85.8 ± 0.5	82.8 ± 0.7	87.2 ± 0.5	84.7 ± 0.5	88.6 ± 0.5	$\mathbf{89.3 \pm 0.4}$	84.8 ± 0.5	85.7 ± 0.5	88.2 ± 0.5	$\mathbf{89.3 \pm 0.5}$
Birds	73.6 ± 0.9	68.6	74.8 ± 0.9	71.0 ± 1.0	76.2 ± 0.8	75.3 ± 0.8	79.2 ± 0.8	78.8 ± 0.7	80.5 ± 0.7	81.4 ± 0.7	79.8 ± 0.7	71.2 ± 0.9	81.8 ± 0.6	$\mathbf{82.1 \pm 0.7}$
Textures	59.5 ± 0.7	66.4	68.8 ± 0.9	71.0 ± 0.8	71.6 ± 0.7	71.2 ± 0.8	68.8 ± 0.8	66.2 ± 0.8	76.2 ± 0.7	76.7 ± 0.7	70.1 ± 0.7	73.0 ± 0.6	76.3 ± 2.4	$\mathbf{78.2 \pm 1.0}$
Quick Draw	74.7 ± 0.8	51.5	76.5 ± 0.8	81.8 ± 0.6	82.4 ± 0.6	77.3 ± 0.7	79.5 ± 0.7	77.9 ± 0.6	81.9 ± 0.6	82.0 ± 0.6	81.6 ± 0.6	$\mathbf{82.8 \pm 0.6}$	78.3 ± 0.7	$\mathbf{82.8 \pm 0.6}$
Fungi	50.2 ± 1.1	39.9	46.6 ± 1.0	64.3 ± 0.9	64.0 ± 1.0	48.5 ± 1.0	58.1 ± 1.1	48.9 ± 1.2	68.8 ± 0.9	67.4 ± 1.0	63.9 ± 1.1	65.8 ± 0.9	$\mathbf{69.6 \pm 1.5}$	$\mathbf{69.5 \pm 1.2}$
VGG Flower	88.9 ± 0.5	87.1	90.5 ± 0.5	82.9 ± 0.8	87.9 ± 0.6	90.5 ± 0.5	91.6 ± 0.6	$\mathbf{92.3 \pm 0.4}$	92.1 ± 0.5	92.2 ± 0.5	89.8 ± 0.6	87.0 ± 0.6	90.3 ± 0.8	$\mathbf{92.4 \pm 1.6}$
Traffic Sign	56.5 ± 1.1	48.8	57.2 ± 1.0	51.0 ± 1.1	48.2 ± 1.1	63.0 ± 1.0	58.4 ± 1.1	59.7 ± 1.1	63.3 ± 1.2	83.5 ± 0.9	51.5 ± 1.0	47.4 ± 1.1	63.6 ± 1.5	$\mathbf{88.4 \pm 2.1}$
MSCOCO	39.4 ± 1.0	43.7	48.9 ± 1.1	52.0 ± 1.1	51.5 ± 1.1	52.8 ± 1.1	50.0 ± 1.0	42.5 ± 1.1	54.0 ± 1.0	55.8 ± 1.1	47.4 ± 1.0	53.5 ± 1.0	57.0 ± 1.1	$\mathbf{57.3 \pm 1.5}$
MNIST	–	–	94.6 ± 0.4	94.3 ± 0.4	90.6 ± 0.5	96.2 ± 0.3	95.6 ± 0.5	94.7 ± 0.3	94.5 ± 0.5	96.7 ± 0.4	93.0 ± 0.5	89.8 ± 0.5	94.7 ± 0.5	$\mathbf{97.3 \pm 1.2}$
CIFAR-10	–	–	74.9 ± 0.7	66.5 ± 0.9	67.0 ± 0.8	75.4 ± 0.8	78.6 ± 0.7	73.6 ± 0.7	71.9 ± 0.7	$\mathbf{80.6 \pm 0.8}$	65.7 ± 0.8	67.3 ± 0.8	71.5 ± 0.9	76.5 ± 1.4
CIFAR-100	–	–	61.3 ± 1.1	56.9 ± 1.1	57.3 ± 1.0	62.0 ± 1.0	67.1 ± 1.0	61.8 ± 1.0	62.6 ± 1.0	$\mathbf{69.6 \pm 1.0}$	54.9 ± 1.0	56.6 ± 0.9	60.0 ± 1.1	67.7 ± 1.5
In-domain	71.6	60.3	73.7	75.9	77.4	74.5	76.2	75.1	80.0	80.2	77.1	76.8	79.7	**80.9**
Out-of-domain	–	–	67.4	64.1	62.9	69.9	69.9	66.5	69.2	77.2	62.5	62.9	69.4	**77.4**
Average	–	–	71.2	71.4	71.8	72.7	73.8	71.8	75.9	79.0	71.4	71.5	75.7	**79.5**
Rank	11.8	13.5	10.1	8.9	6.8	8.7	8.1	7.9	4.1	3.3	8.2	7.8	4.1	**1.6**

Table 2. **Quantitative analysis of two levels of meta-knowledge setting** on Meta-Dataset. In- and out-of-domain average accuracy are reported.

Domain	MDL	2LM_S	2LM	2LM_Noreg
In-domain	77.05	78.35	**79.72**	–
Out-of-domain	62.53	67.13	**69.35**	–
Average	71.38	74.04	**75.73**	74.85

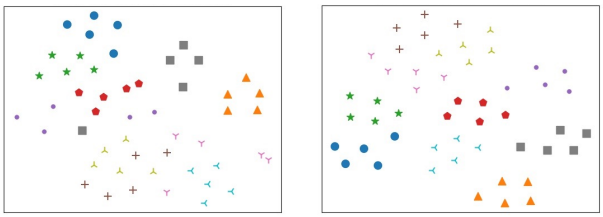


(a) Feature distribution of different domains before projection

(b) Feature distribution of different domains after projection

Figure 3. Distribution visualization of the feature vectors before and after projection about different domains.

Neighbor Embedding (t-SNE) technique [52] to downscale and visualize the base feature vectors outputted by the feature extractor from different domains. The distributions of the feature vectors before and after projection as shown in Figure 3. We can see that the differentiation of the feature vectors after projection is obvious from the right distribution, indicating that our subspace projections have learned the inter-domain variability. And for exploring the effectiveness of a specific projection to the current domain, we do the same visualization for a task sampled randomly from a specific domain in Figure 4. The figure shows that within a given domain, the intra-domain differentiation is improved after projection.



(a) Feature distribution of different classes in a given domain before projection

(b) Feature distribution of different classes in a given domain after projection

Figure 4. Distribution visualization of the feature vectors before and after projection about different classes in a given domain.

6. Conclusion

We have present inter- and intra-domain levels of meteknowledge learning model relying on bi-level optimization formulation for few-shot domain generalization problem, called 2LM. In contrast to the previous methods, 2LM considers that the connection and difference of in- and out-of-domain generalization, and learns a base feature space and a prior subspace for unseen domains with a larger generalization gap, as well as more precise subspaces for seen domains with a smaller generalization gap. Furthermore, we develop an optimization algorithm for model. Our method achieves competitive results compared to the state of the art for FSDG on Meta-Dataset. The ablation studies also have verified the validity of our two levels of meta-knowledge.

7. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62125207, U1936203, 62032022 and 62272443, in part by Beijing Natural Science Foundation under Grant Z190020, JQ22012.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220, 2004. 4

[2] Afra Feyza Akyürek, Ekin Akyürek, Derry Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[3] Kelsey R. Allen, Evan Shelhamer, Hanul Shin, and Joshua B. Tenenbaum. Infinite mixture prototypes for few-shot learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 232–241. PMLR, 2019. 2

[4] Jing Bai, Shaojie Huang, Zhu Xiao, Xianmin Li, Yongdong Zhu, Amelia C. Regan, and Licheng Jiao. Few-shot hyperspectral image classification based on adaptive subspaces and feature transformation. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–17, 2022. 3

[5] Peyman Bateni, Jarred Barber, Jan-Willem van de Meent, and Frank Wood. Enhancing few-shot image classification with unlabelled examples. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1597–1606. IEEE, 2022. 6

[6] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. Improved few-shot visual classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14481–14490. Computer Vision Foundation / IEEE, 2020. 1, 6, 7

[7] Schroeder Brigit and Cui Yin. Fgvcx fungi classification challenge. Online, 2018. 6

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3606–3613. IEEE Computer Society, 2014. 6

[9] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Ann. Oper. Res.*, 153(1):235–256, 2007. 2

[10] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[11] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a multi-domain representation for few-shot classification. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, volume 12355 of *Lecture Notes in Computer Science*, pages 769–786. Springer, 2020. 1, 2, 5, 6

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. 2

[13] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Jimenez Rezende, and S. M. Ali Eslami. Conditional neural processes. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1690–1699. PMLR, 2018. 2, 6

[14] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013. 4

[15] Timothy M. Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5149–5169, 2022. 1

[16] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4005–4016, 2019. 2

[17] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 International Joint Conference on Neural Networks, IJCNN 2013, Dallas, TX, USA, August 4-9, 2013*, pages 1–8. IEEE, 2013. 6

[18] T. Kawashima J. Kim J. Jongejan, H. Rowley and N. Fox-Gieg. The quick, draw!-ai experiment. (2016). (201). 6

[19] Shuqiang Jiang, Yaohui Zhu, Chenlong Liu, Xinhang Song, Xiangyang Li, and Weiqing Min. Dataset bias in few-shot image recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):229–246, 2023. 1

[20] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015. 2

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[22] Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference*

on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 24581–24592, 2021. 3

[23] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 6

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. 6

[25] John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pages 1–31. Springer, 2013. 4

[26] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1–10. Computer Vision Foundation / IEEE, 2019. 2

[27] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Universal representation learning from multiple domains for few-shot classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9506–9515. IEEE, 2021. 3, 5, 6, 8

[28] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7151–7160. IEEE, 2022. 1, 3, 6

[29] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7260–7268. Computer Vision Foundation / IEEE, 2019. 2

[30] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy M. Hospedales. Feature-critic networks for heterogeneous domain generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3915–3924. PMLR, 2019. 2

[31] Yann Lifchitz, Yannis Avrithis, and Sylvaine Picard. Local propagation for few-shot learning. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 10457–10464. IEEE, 2020. 2

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 6

[33] Lu Liu, William L. Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle. A universal representation transformer layer for few-shot image classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3, 6

[34] Yanbin Liu, Juho Lee, Linchao Zhu, Ling Chen, Humphrey Shi, and Yi Yang. A multi-mode modulator for multi-domain few-shot classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8433–8442. IEEE, 2021. 6

[35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 6

[36] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2624–2637, 2013. 5

[37] Carl Dean Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000. 4

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society, 2008. 6

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[40] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press, 2018. 2

[41] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2

[42] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 506–516, 2017. 2

[43] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 8119–8127. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[44] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E. Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7957–7968, 2019. 1, 2, 6

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 6

[46] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2

[47] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4135–4144. Computer Vision Foundation / IEEE, 2020. 3

[48] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087, 2017. 1, 2, 5

[49] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 403–412. Computer Vision Foundation / IEEE, 2019. 2

[50] Eleni Triantafillou, Hugo Larochelle, Richard S. Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10424–10433. PMLR, 2021. 1, 5, 6

[51] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 6, 8

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[53] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artif. Intell. Rev.*, 18(2):77–95, 2002. 1

[54] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):13:1–13:37, 2019. 6

[55] Qianqian Xu, Zhiyong Yang, Yangbangyan Jiang, Xiaochun Cao, Yuan Yao, and Qingming Huang. Not all samples are trustworthy: Towards deep robust SVP prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6):3154–3169, 2022. 2

[56] Han-Jia Ye, Lu Ming, De-Chuan Zhan, and Wei-Lun Chao. Few-shot learning with a strong teacher. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[57] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7343–7353, 2018. 2

[58] Hao Zhu and Piotr Koniusz. EASE: unsupervised discriminant subspace learning for transductive few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9068–9078. IEEE, 2022. 3