Optimal Transport of Diverse Unsupervised Tasks for Robust Learning from Noisy Few-Shot Data

Xiaofan Que[1] and Qi Yu[1]

Rochester Institute of Technology
{xq5054,qi.yu}@rit.edu

Abstract. Noisy few-shot learning (NFSL) presents novel challenges primarily due to the interplay between noisy labels and limited training data. While data cleansing offers a viable solution to address noisy labels in the general learning settings, it exacerbates information loss in FSL due to limited training data, resulting in inadequate model training. To best recover the underlying task manifold corrupted by the noisy labels, we resort to learning from uniquely designed unsupervised auxiliary tasks to compensate for information loss. Using unsupervised tasks can effectively avoid additional annotation costs and minimize the risk of introducing additional label noises. However, a randomly constructed unsupervised task may misguide the model to learn sample-specific features that are likely to compromise the primary few-shot learning task due to the noisy weak learning signals. We propose to conduct novel auxiliary task selection to ensure the intra-diversity among the unlabeled samples within a task. Domain invariant features are then learned from carefully constructed auxiliary tasks to best recover the original data manifold. We conduct a theoretical analysis to derive novel generalization bounds for learning with auxiliary tasks. Extensive experiments are conducted to demonstrate that our method outperforms existing noisy few-shot learning methods under various in-domain and cross-domain few-shot classification benchmarks.

Keywords: Noisy few-shot learning · Auxiliary task · Optimal transport

1 Introduction

Few-shot learning (FSL) aims to utilize knowledge acquired from a large amount of diverse tasks during training and apply it to new, unseen tasks from the same task distribution with only a small amount of labeled data [10, 43, 45]. Therefore, learning an accurate representation of the task manifold that effectively captures the underlying structure and relationships among tasks is essential to ensure good few-shot generalization. On the other hand, noisy labels commonly exist in the task distribution due to the ever increasing demand for large-scale annotated datasets, where labeling errors can be introduced in various means that are impossible to control [4, 30, 34, 47, 58]. Given the few-shot nature of each training task, memorization of noisy labels introduced along with limited training data can exacerbate overfitting, which will disrupt the representation space and severely hinder few-shot generalization.

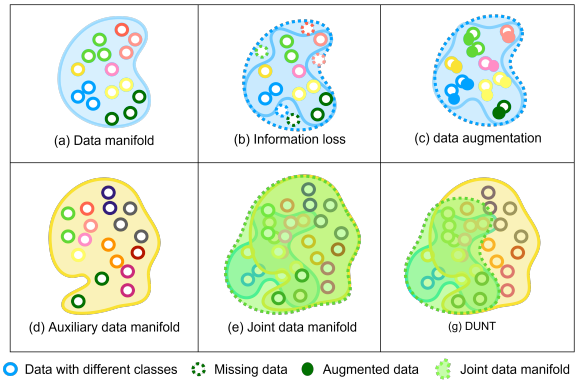


Fig. 1: (a) Data manifold of the primary task distribution. (b) Data manifold after using data cleansing to remove some noisy labels, which is incomplete due to information loss. (c) Using augmentation can make up for the information loss but it is limited by the diversity of existing data. (d) Data manifold of the auxiliary task distribution. (e) Joint data manifold of primary and auxiliary data distribution, in which the domain-specific knowledge of the auxiliary manifold is also included. (f) Through optimal transport of diverse unsupervised tasks, the learned shared representation only includes rich domain-invariant information from the auxiliary manifold.

While training robust models from noisy data has been intensively investigated in standard settings, learning from noisy few-shot data is only sparsely explored in existing literature [28,31]. Noisy few-shot learning (NFSL) presents novel challenges primarily due to the interplay between noisy labels and limited training data. Within each training task, FSL operates with few training samples per class, making each data point highly valuable. Hence, the model tends to be more sensitive to variations in the training data and noisy labels can more easily cause the model to learn incorrect patterns, compromising its ability to accurately generalize to new tasks. Furthermore, identifying and rectifying label noises also becomes more challenging. The scarcity of data makes it harder to distinguish between true class characteristics and erroneous annotations.

Techniques, including robust optimization [28] and data augmentation [31] have been leveraged to handle noises in limited data. However, these techniques, while beneficial, do not entirely resolve the problem. Conventional methods for learning with noisy labels, such as data cleansing, data augmentations, and robust loss functions, are inadequate to tackle the challenges as outlined above. As shown in Fig. 1(a)-(b), data cleansing introduces information loss, further diminishing the already scarce training data in each task. This poses a greater challenge for the model to learn a meaningful and generalizable representation of the task manifold. Data augmentation could alleviate the information loss introduced through data cleansing as shown in Fig. 1(c). Nevertheless, augmentation is performed on the limited training samples, so it may not cover a broad task manifold. Additionally, robust optimization algorithms disregard the concern of information loss, making them less suitable for learning from noisy few-shot data.

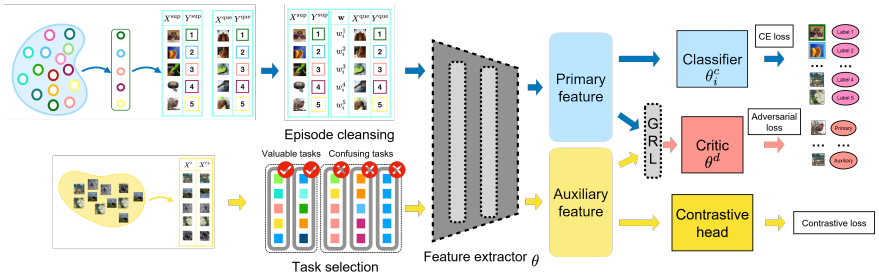


Fig. 2: In DUNT, the feature extractor extract information from both primary and auxiliary tasks, the classifier predicts the labels of the primary and auxiliary examples, and the domain critic outputs the source of the examples. The primary tasks (top) are sampled and cleansed by episode cleansing, the auxiliary tasks (bottom) are constructed and filtered by their diverseness. (Best viewed in color.)

To tackle the unique challenges arising from NFSL, we propose novel extensions to existing techniques designed for standard noisy learning scenarios to more effectively handle the information loss in the few-shot setting. In particular, to best recover the underlying task manifold corrupted by the noisy labels, we resort to learning from uniquely designed auxiliary tasks to compensate information loss. The auxiliary tasks are entirely unsupervised as using supervised tasks can be counterproductive in the NFSL setting for three main reasons: (1) They may incur additional annotation overhead. (2) Noisy labels may continue to be introduced that further compromise the manifold structure. (3) Even clean labels can be guaranteed, these labels may not totally align with the primary tasks that can potentially overwhelm that weak learning signals carried by the limited training samples.

While unsupervised auxiliary tasks have been leveraged in diverse learning scenarios [19, 24, 32, 36, 52], including the few-shot setting, the noisy labels and potential information loss introduced through necessary data cleansing makes NFSL much more vulnerable to the potential noises introduced by the auxiliary data. Since data samples are unlabeled, they are usually randomly sampled into an auxiliary task. When a contrastive loss is used to differentiate unlabeled data samples from the same underlying class, the model may be forced to learn sample-specific (spurious) features that cause overfitting. To this end, we propose to conduct novel auxiliary task selection to ensure the intra-diversity among the chosen unlabeled samples within a task. Domain invariant features are then learned from carefully constructed auxiliary tasks to best recover the original data manifold as shown in Fig. 1(d)-(g).

Fig. 2 shows the proposed framework, referred to as **D**iverse **UN**supervised **T**asks for NFSL (DUNT4NFSL or DUNT for short). The feature extractor of DUNT undergoes updates through three distinct loss components: 1) A cross-entropy loss, integral for extracting discriminative and effective features for classification from the primary tasks; 2) A contrastive loss tailored for the unsupervised auxiliary tasks, designed to encourage an embedding space where similar sample

pairs remain close while dissimilar pairs are distanced; 3) An adversarial loss, implemented to align the unsupervised auxiliary tasks with the primary tasks, employing a domain critic head to reduce the covariate shift across multiple domains. To ensure that the feature representation remains unaffected by label noise, we apply an episode cleaning technique to the primary task, removing noisy examples in each task. For the auxiliary tasks, we introduce a novel criterion for diverseness to guarantee that the unsupervised auxiliary tasks contribute informative and valuable insights. Our contributions are threefolds:

- We present `DUNT`, a framework designed to adeptly utilize unsupervised data from a distinct domain in order to counteract the information loss resulting from data cleansing in the challenging scenario of NFSL.
- We introduce a novel strategy to perform diverse auxiliary task selection to avoid learning sample-specific spurious features from the unlabeled auxiliary data samples.
- We conduct an in-depth theoretical analysis of the auxiliary tasks introduced in the `DUNT` framework, establishing novel generalization bounds, offering valuable insights into the contributions of these auxiliary tasks.

We have conducted comprehensive experiments on five benchmark datasets with synthetic label noises to demonstrate the effectiveness and robustness of `DUNT` and its advantage over SOTA methods by a large margin.

2 Related Work

Noisy few-shot learning. Existing methods for training DNNs with noisy labels usually fall into the following categories: *robust loss* [12,49,50,57,60] designs robust cost functions in the presence of noisy labels; *noise transition estimation* [5,20,53,54] aims to estimate the probabilities that clean labels flip into noisy labels; *loss correction* [17,22,23,39,44,55,59] either explicitly or implicitly relabels or reweights the noisy samples, a.k.a. data cleansing. Noisy label in FSL is an important yet relatively under-explored topic. The scarcity of data in FSL not only heightens its susceptibility to label noise but also adds complexity to the challenge of identifying and correcting such noise. A few previous attempts such as [28,29,31] address mislabeled issue in FSL settings using re-weighting strategy or generating robust prototypes by feature hybrid and augmentation techniques. Other works such as RW-MAML [25], AQ [14], DFSL [27] tackle the out-of-distribution tasks or adversarial attacks in few-shot learning. In this work, we particularly focus on solving the limited data problem by utilizing auxiliary unlabeled data as unsupervised tasks to enhance the robustness of FSL.

Optimal transport and Wasserstein distance. The optimal transport (OT) cost is used to measure the difference between distributions supported on high-dimensional space using SGD [1,46]. For a constant $\xi \geq 1$, the ξ-Wasserstein

metric between distributions $\mathbb{P}_X$ and $\mathbb{P}_{X'}$ is defined as:

$$\mathcal{W}_\xi(\mathbb{P}_X, \mathbb{P}_{X'}) = \left(\inf_{\gamma \in \prod(\mathbb{P}_X, \mathbb{P}_{X'})} \mathbb{E}_{\gamma(\mathbf{x}, \mathbf{x}')}[c^\xi(\mathbf{x}, \mathbf{x}')] \right)^{\frac{1}{\xi}} \tag{1}$$

where $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ denote the random variables in which the distributions $\mathbb{P}_X$ and $\mathbb{P}_{X'}$ are defined. $\prod(\mathbb{P}_X, \mathbb{P}_{X'})$ is the set of all joint distributions (*i.e.* couplings) whose marginals are $\mathbb{P}_X$ and $\mathbb{P}_{X'}$, that is, $\int \gamma(\mathbf{x}, \mathbf{x}')d\mathbf{x} = \mathbb{P}_{X'}$ and $\int \gamma(\mathbf{x}, \mathbf{x}')d\mathbf{x}' = \mathbb{P}_X$. $c(\mathbf{x}, \mathbf{x}') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is the cost function for moving from $\mathbf{x}$ to $\mathbf{x}'$. Intuitively, $\gamma(\mathbf{x}, \mathbf{x}')$ indicates how much "mass" must be transported from $\mathbf{x}$ to $\mathbf{x}'$ in order to transform the distributions $\mathbb{P}_X$ into $\mathbb{P}_{X'}$. The Wasserstein metric then is the "cost" of the optimal transport plan [1]. Eq. (1) is the primal form of the Wasserstein metric, particularly for the case of $\xi = 1$ with $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$ [1]. Some existing efforts propose domain adaptation algorithms using ideas from OT theory, which aims to reduce the divergence between two domains by minimizing the Wasserstein distance between their distributions [7, 8, 11]. Works that use domain adaptation techniques in FSL [6, 18, 33, 40, 56] share the same goal that aims to learn transferable knowledge. In this work, we achieve this goal by leveraging extra data first and then minimize the Wasserstein distance between the primary and auxiliary distributions to ensure the learned knowledge are domain-invariant and discriminative to future unseen tasks.

3 Methodology

Problem formulation. In FSL, the goal is to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ that maps an observation $\mathbf{x}$ to its label $\mathbf{y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space. To train the model with the quick adaptation ability to the new tasks with only a small number of examples, a set of tasks $\mathcal{T}_i = \{\mathcal{S}_i^{\text{sup}}, \mathcal{S}_i^{\text{que}}\}$ is sampled from the data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{S}_i^{\text{sup}} = \{(\mathbf{x}_i^k, y_i^k)\}_{k=1}^{NK}$ denotes the support set and $\mathcal{S}_i^{\text{que}} = \{(\mathbf{x}_i^k, y_i^k)\}_{k=1}^{NQ}$ denotes the query set. N is the number of classes, K is the number of examples in each class of the support set, and Q is that of the query set. In `DUNT`, the meta-training tasks are sampled from the two domains: the primary tasks are sampled from the noisy primary distribution $\tilde{\mathcal{D}}^{\text{pri}}$ over $\mathcal{X} \times \tilde{\mathcal{Y}}$, where the label space $\tilde{\mathcal{Y}}$ is considered noisy, and the unsupervised auxiliary tasks are sampled from the auxiliary distribution $\mathcal{D}^{\text{aux}}$ over $\mathcal{X}'$, where the entire label space is missing. We assume that the primary and the auxiliary distributions are from different domains: $\tilde{\mathcal{D}}^{\text{pri}} \neq \mathcal{D}^{\text{aux}}$. During meta-testing, the tasks are sampled from a new unseen target domain $\mathcal{D}^{\text{tgt}}$ over $\mathcal{X}^{\text{tgt}} \times \mathcal{Y}^{\text{tgt}}$, the label is only used for evaluation purpose. The goal is to learn a meta-model θ trained on $\tilde{\mathcal{D}}^{\text{pri}}$ and $\mathcal{D}^{\text{aux}}$ that can quickly adapt to tasks in $\mathcal{D}^{\text{tgt}}$ using only a few examples.

Framework overview. The `DUNT` framework (see Fig. 2) is composed of a shared feature extractor θ, the task classifier θ_i^c for task i and the domain critic θ^d. The feature extractor takes input from both the primary tasks and the auxiliary

tasks; the classifiers predict the labels for the input and the critic predict the source of the input. The gradient reversal layer (GRL) ensures that the feature distributions over the two domains are made similar.

3.1 Learning from the Primary Tasks

One commonly used FSL strategy that takes advantage of task-based formulation is to learn a meta-model through a bi-level optimization. By defining the task-specific model parameter for task $\mathcal{T}_i$ as θ_i', the meta model θ are trained by optimizing for the performance of θ_i' with respect to θ across tasks sampled from $\mathcal{D}^{\text{pri}}$. The loss is

$$\mathcal{L}^{\text{pri}} = \sum_{\mathcal{T}_i(\mathcal{S}_i^{\text{sup}}, \mathcal{S}_i^{\text{que}}) \sim \tilde{\mathcal{D}}^{\text{pri}}} \mathcal{L}_{\mathcal{T}_i}(\theta_i', \mathcal{S}_i^{\text{que}}) \tag{2}$$

where $\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\theta, \mathcal{S}_i^{\text{sup}})$ is the task-specific model adapted from θ within one gradient step. This is a dual-level optimization where in the outer-level, the meta-model θ is updated with the summation of a series of tasks evaluated on the query sets $\{\mathcal{S}_i^{\text{que}}\}$, with $\mathcal{L}_{\mathcal{T}_i}(\theta_i', \mathcal{S}_i^{\text{que}}) = \sum_{(\mathbf{x}_i^k, \mathbf{y}_i^k) \in \mathcal{S}_i^{\text{que}}} \ell(\theta_i^c(\theta_i'(\mathbf{x}_i^k)), \mathbf{y}_i^k)$. The inner-level objective is computed using the task-specific model θ_i' updated with support set $\mathcal{S}_i^{\text{sup}}$ using loss $\mathcal{L}_{\mathcal{T}_i}(\theta, \mathcal{S}_i^{\text{sup}}) = \sum_{(\mathbf{x}_i^k, \mathbf{y}_i^k) \in \mathcal{S}_i^{\text{sup}}} \ell(\theta_i^c(\theta(\mathbf{x}_i^k)), \mathbf{y}_i^k)$, where ℓ is the cross-entropy loss, θ_i^c is the i-th task classifier.

Episode cleansing. In the dual-level optimization, the meta-model θ is updated by first adapting to different task-specific models θ_i' with the corresponding support set and then updated using the summation of the loss of the task-specific models θ_i' on their corresponding query sets. Since the dataset is noisy, the error may be accumulated through the noisy query set and cause the overfitting of the meta-model. Therefore, we propose episode cleansing to remove the noisy data in the query sets based on the memorization effect which indicates that noisy examples have higher losses than the clean ones. The classification loss $\mathcal{L}_{\mathcal{T}_i}(\theta_i', \mathcal{S}_i^{\text{que}})$ becomes

$$\mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}, \hat{\mathcal{S}}_i^{\text{que}}) = \sum_{\mathbf{x}_i^k, \mathbf{y}_i^k \in \mathcal{S}_i^{\text{que}}} w_i^k \ell_i^k, \tag{3}$$

where $\hat{\mathcal{S}}_i^{\text{que}}$ is the cleansed (weighted) query set for task $\mathcal{T}_i$; for simplification, we use ℓ_i^k to denote $\ell(\theta_i^c(\theta_i'(\mathbf{x}_i^k)), \mathbf{y}_i^k)$. w_i^k is the self-paced sample weight for each example pair, and

$$w_i^k = \mathbb{1}(\ell_i^k < \gamma_{th}),$$

where $\mathbb{1}(\cdot)$ is an indicator function whose value is 1 when $\ell_i^k < \gamma_{th}$ and 0 otherwise; and γ_{th} is a predefined hyperparameter used to filter out high-loss examples. We use the percentage of the ordered loss to avoid extensive hyperparameter search, whose ablation is provided in Appendix C.

The primary loss the becomes:

$$\mathcal{L}^{\text{pri}} = \sum_{\mathcal{T}_i(\mathcal{S}_i^{\text{sup}}, \mathcal{S}_i^{\text{que}}) \sim \tilde{\mathcal{D}}^{\text{pri}}} \mathcal{L}_{\mathcal{T}_i}(\theta_i', \hat{\mathcal{S}}_i^{\text{que}}). \tag{4}$$

Similarly, for the support set label noise, we use meta-model to compute the loss values and remove the noisy ones in the support set.

3.2 Learning from Diverse Unsupervised Tasks

As shown in Fig. 2, the general unsupervised auxiliary task $\mathcal{T}_j$ is constructed by i) first randomly sampling N images from the dataset for the support set $\mathcal{S}_j^{\text{sup}} = \{\mathbf{x}_j^k\}_{k=1}^N$; ii) the query set $\mathcal{S}_j^{\text{que}} = \{A_1(\mathbf{x}_j^k), ..., A_Q(\mathbf{x}_j^k)\}_{k=1}^N$ is obtained by augmenting the N images in Q different ways, where $A_1(\cdot), .., A_Q(\cdot)$ denote Q different augmentation techniques, such as random cropping, translation, flipping. The auxiliary loss is the summation of the contrastive loss of a series of auxiliary tasks:

$$\mathcal{L}^{\text{aux}} = \sum_{\mathcal{T}_j(\mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}) \sim \mathcal{D}^{\text{aux}}} \mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}). \tag{5}$$

The contrastive loss is utilized for the unsupervised tasks:

$$\mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}) = -\frac{1}{NQ} \sum_{k=1}^{N} \sum_{q=1}^{Q} \log \frac{\exp(-d[\theta(A_q(\mathbf{x}_j^k)), \theta(\mathbf{x}_j^k)])}{\sum_{l=1}^{N} \exp(-d[\theta(A_q(\mathbf{x}_j^k)), \theta(\mathbf{x}_j^l)])} \tag{6}$$

where $d[\cdot, \cdot]$ is a distance metric, such as Euclidean distance.

Auxiliary task selection. In contrastive learning, the positive pairs are formulated by the original example and its augmentation, and the negative pairs are formulated by images from two different classes. The contrastive loss is designed to pull the positive pairs close to each other and push the negative ones away. However, since each image in the unsupervised task are randomly sampled and treated as a class, there are chances that the images are sampled from the same class. In some extreme case, all the images are sampled from the same class, which means that the positive pairs being treated as negative pairs and being push away from each other. This would cause the model being forced to overfit to the unnecessary details and creates spurious correlations and harm the generalization ability. We denote those tasks as confusing tasks, as a comparison, the tasks with images sampled from different classes are denoted as valuable tasks. The intricate nature of confusing tasks appears to pose a significant threat to the resilience and robustness of FSL, as shown in our later ablation studies. In this work, we propose a novel task selection metric to remove these confusing unsupervised tasks.

Given two images $\mathbf{x}_j^k$, $\mathbf{x}_j^l$, the model parameterized by θ lead to class probabilities vectors $\mathbf{p}_j^k = \text{softmax}(\theta(\mathbf{x}_j^k))$ and $\mathbf{p}_j^l = \text{softmax}(\theta(\mathbf{x}_j^l))$. If the two images

are from the same class, $\mathbf{p}_j^k$ and $\mathbf{p}_j^l$ are naturally encouraged to be close to the assignment vector $\mathbf{e}_y$ in $\{0,1\}^N$ with a single non-zero entry at position y, where y is the class label associated to $\mathbf{x}$ and N is the number of classes. Therefore, we consider the following diversity criterion:

$$div_j = \frac{1}{C(N,2)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{1}{\cos(\mathbf{p}_j^k, \mathbf{p}_j^l)}.$$

The diverseness of the task $\mathcal{T}_j$ is defined as the average of the reciprocal of the cosine similarity between the class probabilities of all example pairs in the task, where $C(N,2) = \frac{N!}{(N-2)!2!}$ is the 2-combination of set N. According to the definition, the higher the value of div_j, the more diverse the task.

Auxiliary loss function. Auxiliary loss $\mathcal{L}^{\text{aux}}$ is updated as

$$\mathcal{L}^{\text{aux}} = \sum_{\mathcal{T}_j(\mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}) \sim \mathcal{D}^{\text{aux}}} v_j \mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}})$$

where v_j is the diversity weight defined as:

$$v_j = \mathbb{1}(div_j > \gamma_{div}),$$

where γ_{div} is a predefined hyperparameter, which filter out those less diverse, a.k.a., confusing tasks. In practice, we simplify the hyperparameter searching by first upsampling six auxiliary tasks, computes each task diversity metric, and only keep the most diverse task in $\{\mathcal{T}_j\}^{\text{val}}$ for training. The corresponding ablations about the numbers of extra tasks sampled are presented in Sec. 4.3.

3.3 Adversarial Training and the Critic Head

Since the auxiliary tasks are drawn from a different data distribution, it is inevitable that there exist some domain shift from the primary domain to the auxiliary one. While learning from the domain-specific knowledge from the auxiliary domain main cause the learned model biased and deteriorate the model's generalization performance, learning from the domain-invariant knowledge helps improve the generalization ability. To enforce the learning of the domain-invariant representation of the meta-model, we propose a critic model θ^c to discriminate the source of the input data, either primary or auxiliary. By training the model in an adversarial way, the critic model θ^c try to distinguish which domain that the input features come from, while the meta-model θ aims to confuse the critic. In this way, when the training process converge, θ tends to extract domain-invariant knowledge from both domains. Formally, the overall adversarial objective is a min-max problem as follows:

$$\min_{\theta,\theta^c} \max_{\theta^d} \alpha_1 \mathcal{L}^{\text{pri}} + (1-\alpha_1)\mathcal{L}^{\text{aux}} + \alpha_2 \mathcal{E}(\theta, \theta^d, \tilde{\mathcal{D}}^{\text{pri}}, \mathcal{D}^{\text{aux}}), \tag{7}$$

where the critic model θ^d assigns 1 to the example from the primary domain and 0 otherwise based on the feature representation extracted by θ. $\mathcal{E}(\cdot)$ is the adversarial loss

$$\mathcal{E}(\theta, \theta^d, \tilde{\mathcal{D}}^{\text{pri}}, \mathcal{D}^{\text{aux}}) = \mathbb{E}_{\mathbf{x}\sim\tilde{\mathcal{D}}^{\text{pri}}}[g(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}^{\text{aux}}}[g(\mathbf{x})],$$

where $g(\mathbf{x}) = \theta^d(\theta(\mathbf{x}))$, with its empirical counterpart as

$$\hat{\mathcal{E}}(g_{\theta,\theta^d}, \tilde{\mathcal{D}}^{\text{pri}}, \mathcal{D}^{\text{aux}}) = \frac{1}{|\mathcal{D}^{\text{pri}}|}\sum_{x\sim\mathcal{D}^{\text{pri}}}[g(\mathbf{x})] - \frac{1}{|\mathcal{D}^{\text{aux}}|}\sum_{\mathbf{x}\sim\mathcal{D}^{\text{aux}}}[g(\mathbf{x})]$$

The optimization of Eq. (7) is conducted through an alternative training between the optimization of the meta-model feature extractor θ, classifier parameter θ^c and the critic model parameter θ^d while keeping the other fixed. Parameters of θ, θ^c and θ^d are updated alternatively. In practice, θ and θ^c are connected by the GRL, which reverses the gradients that flow through θ^c. The pseudo code of the algorithm can be found in Appendix A.

3.4 Theoretical Analysis

We conduct a theoretical analysis of `DUNT` to show: 1) the equivalence between the adversarial regularizer and the Wasserstein distance, which captures the diversity property of the auxiliary distribution while enforcing the learning of a domain invariant representation; 2) the relationship between the primary and auxiliary distributions; 3) a tight generalization bound of the target based on the expected risk of the primary and auxiliary distributions.

Equivalence to Wasserstein distance. The Kantorovich-Rubinstein of Wasserstein distance takes the form [1]:

$$\mathcal{W}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{\|h\|_L\leq 1} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_1}[h(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_2}[h(\mathbf{x})] \tag{8}$$

where the supremum is over all the 1-Lipschitz functions $h : \mathcal{X} \to \mathbb{R}$. Therefore,if we have a parameterized family of functions $\{h_w\}_{w\in\mathcal{W}}$ that are all 1-Lipschitz, we could consider solving the problem

$$\max \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_1}[h_w(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_2}[h_w(\mathbf{x})] \tag{9}$$

which has the same form of the regularizer in Sec. 3.3. This indicates that our adversarial regularizer has the same theoretical foundation of Wasserstein distance optimal transport. To satisfy the 1-Lipschitz constraint, the critic θ^d is implemented with gradient penalty [15]. The overall objective can be reformulated as

$$\min_{\theta} \alpha_1\mathcal{L}^{\text{pri}} + (1-\alpha_1)\mathcal{L}^{\text{aux}} + \alpha_2\mathcal{W}_1(\mathcal{D}^{\text{pri}}, \mathcal{D}^{\text{aux}}), \tag{10}$$

where the first and second term are the primary and auxiliary loss, resp., and the third term is the Wasserstein distance based regularizer, which aims to extract domain-invariant feature from the primary and auxiliary domains. Wasserstein distance reflects an optimal transport cost for moving one distribution to another. A smaller transport cost means a better coverage of the distribution, which captures the property of diversity [42].

Relationship of the primary and auxiliary distributions. By viewing this problem as a distribution matching that matches the auxiliary distribution with the primary one, we have the generalization property that quantify the relationship between the primary and the auxiliary distribution. Given the primary distribution $\mathcal{D}^{\text{pri}}$ and the auxiliary distribution $\mathcal{D}^{\text{aux}}$, the expected risk w.r.t. is defined as ${}^{\mathcal{D}^{\text{pri}}}R_{\mathcal{D}^{\text{pri}}} = \mathbb{E}_{x\sim\mathcal{D}^{\text{pri}}}\ell(h(x), h^*(x))$, and the expected risk w.r.t. $\mathcal{D}^{\text{aux}}$ is defined as $R_{\mathcal{D}^{\text{aux}}} = \mathbb{E}_{x\sim\mathcal{D}^{\text{aux}}}\ell(h(x), h^*(x))$. We have the following Lemma 1:

Lemma 1. *If the loss ℓ is symmetric and L-Lipschitz, $\forall h \in \mathcal{H}$ is at most H-Lipschitz function and the underlying mapping function h^* is $\phi(\lambda)-(\mathcal{D}^{pri}, \mathcal{D}^{aux})$ joint probabilistic Lipschitz; then the expected risk w.r.t. $\mathcal{D}^{pri}$ is upper bounded*

$$R_{\mathcal{D}^{pri}} \leq R_{\mathcal{D}^{aux}} + L(H+\lambda)\mathcal{W}_1(\mathcal{D}^{pri}, \mathcal{D}^{aux}) + L\phi(\lambda), \tag{11}$$

where $\phi : \mathbb{R} \to [0, 1]$ and a mapping h^* is $\phi(\lambda) - (\mathcal{D}^{pri}, \mathcal{D}^{aux})$ joint probabilistic Lipschitz if $supp(\mathcal{D}^{aux}) \subseteq supp(\mathcal{D}^{pri})$; $\forall\lambda > 0$, all coupling $\gamma \in \prod(\mathcal{D}^{pri}, \mathcal{D}^{aux})$: $\mathbb{P}_{(x_{\mathcal{D}^{pri}}, x_{\mathcal{D}^{aux}})\sim\gamma}[|h^*(x_{\mathcal{D}^{pri}}) - h^*(x_{\mathcal{D}^{aux}})| > \lambda\|x_{\mathcal{D}^{pri}} - x_{\mathcal{D}^{aux}}\|_2] \leq \phi(\lambda)$.

Lemma 1 shows that the expected risk of the primary distributions is upper bounded by the expected risk w.r.t. the auxiliary distribution, the Wasserstein distance, and the labeling function property. The proof of Lemma 1 can be derived with Theorem 1 in [42]. Eq. (10) follows the upper bound from Lemma 1, which means a desirable auxiliary distribution with a better matching with the primary distribution which holds a small expected risk leads to tighter generalization bound for the primary distribution. This further verifies the rationality of our design of using the WD based regularizer.

Generalization bound of DUNT. Furthermore, in the following Lemma 2, we derive the upper bound for the target risk $R_{\mathcal{D}^{\text{tgt}}} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{\text{tgt}}}\ell(\mathbf{x}, y)$.

Lemma 2. *Let $\mathcal{D}^{tgt}$ be the target distribution, h is K-Lipschitz w.r.t. the feature $\mathbf{x}$, and the setting of loss and hypothesis be the same as Lemma 1, the expected risk w.r.t. the target distribution is upper-bounded by:*

$$\begin{aligned} R_{\mathcal{D}^{tgt}} \leq & \alpha_1(R_{\mathcal{D}^{pri}} + LK\mathcal{W}_1(\mathcal{D}^{pri}, \mathcal{D}^{tgt})) \\ & +(1-\alpha_1)(R_{\mathcal{D}^{aux}} + LK\mathcal{W}_1(\mathcal{D}^{aux}, \mathcal{D}^{tgt})), \end{aligned} \tag{12}$$

Lemma 2 shows that the expected risk w.r.t. the target distribution is upper bounded by the expected risk w.r.t. the primary distribution and the auxiliary distribution, and the corresponding Wasserstein distances. The proof of Lemma 2

can be derived with Theorem 2 in [41]. Lemma 2 provide solid theoretical foundation for the design of DUNT with a tight generalization bound to the unseen target domain with primary and auxiliary distributions, which can be extended to multiple sources.

4 Experiments

4.1 Datasets, Baselines, and Experimental Settings

Datasets. We conduct experiments using the following benchmark datasets for FSL:*CUB* [51], *miniImageNet (mini)* [48], *tieredImageNet (tiered)* [38], *CIFAR-FS (CF)* [3], and *Cross-Domain* dataset [16]. Each datasets contains synthetic labels noise with different ratios. More details are in the Appendix.

Baselines. We use ProtoNet [43] as the base few-shot learning method and consider comparing with improving methods with general data cleansing methods including SPL [26] and FSR [61], denoted as *PN+SPL* and *PN+FSR*, respectively. We also consider to improve ProtoNet in transductive settings including CoTeaching [17], CoTeaching+ [59], denoted as *PN+Co*, and *PN+Co+*, respectively, *STARTUP* [35], and *DDNet* [21]. Last but not least, we consider to compare with noisy FSL methods *RNNP* [31], *TraNFS* [28] and *DCML* [37]. The detailed descriptions and settings are provided in the Appendix.

Experimental settings. The feature extractor follows the same architecture as the ProtoNet used by [43], which has 4 modules with a 3×3 convolutions and 64 filters, followed by batch normalization, a ReLU nonlinearity, and 2×2 max-pooling, a.k.a., Conv4. The architecture of classifier have the design with a fully connected layer followed by a ReLU nonlinearity, a dropout layer, and another fully connected layer. The critic architecture has the same architecture with an extra sigmoid function. We've considered transductive and non-transductive in the settings. For the former, unlabeled test data of the same dataset is used as auxiliary data during training. This is applied to our method and other transductive methods including PN+Co, PN+Co+, STARTUP, and DDNet. For the latter, we utilize unlabeled datasets from other different datasets as auxiliary. This setting is applied to all the mentioned methods. During training, we use a learning rate=0.001, and use Adam as the optimizer. The batch size is 4. The meta-model is trained with 40000 randomly selected primary tasks and 40000 randomly selected auxiliary tasks. During evaluation, the average accuracy is computed based on 600 randomly sampled tasks.

4.2 Main Experimental Results

Following the experiment design of [48], we report the average accuracy of 5 runs of each 5-way 1-shot and 5-way 5-shots experiments in Tab. 1 to Tab. 2. We use $\alpha_1 = 0.5$, $\alpha_2 = 0.02$, $\gamma_{th} = 50\%$, γ_{div} as one out of six tasks for our methods

Table 1: Non-transductive auxiliary datasets

Methods	CUB	mini	mini-CUB	tiered	CIFAR-FS
Task	5-way 1-shot				
PN+SPL	48.55 ± 1.51	43.44 ± 1.49	37.06 ± 0.63	42.11± 0.71	55.20 ± 0.69
PN+FSR	37.14 ± 0.60	44.81 ± 0.59	39.82 ± 0.68	30.65 ± 0.49	37.39 ± 0.63
RNNP	55.39 ± 0.49	50.18 ± 0.62	41.75 ± 0.72	43.76 ± 0.48	60.15 ± 0.72
TraNFS	53.02 ± 0.52	46.66 ± 0.63	38.25 ± 0.79	40.32 ± 0.88	60.86 ± 0.63
DCML	54.39 ± 0.87	49.33 ± 1.22	41.77 ± 0.80	48.33 ± 0.80	60.28 ± 0.82
DUNT	**56.94** ± 1.03	**52.22** ± 1.10	**43.01** ± 0.51	**49.12** ± 0.68	**62.80** ± 0.49
Task	5-way 5-shot				
PN+SPL	68.47 ± 0.74	58.67 ± 0.86	54.94 ± 1.03	59.62 ± 1.31	70.93 ± 0.83
PN+FSR	51.48 ± 0.91	61.89 ± 0.80	57.38 ± 1.10	50.40 ± 1.21	44.68 ± 1.09
RNNP	70.37 ± 1.33	63.57 ± 1.28	**58.79** ± 2.36	55.07 ± 1.21	74.31 ± 1.36
TraNFS	68.77 ± 1.17	61.06 ± 1.13	51.24 ± 1.16	48.74 ± 1.20	74.15 ± 0.83
DCML	71.28 ± 1.23	62.77 ± 1.00	57.77 ±1.33	61.19 ± 1.01	72.33 ± 0.90
DUNT	**72.17** ± 1.11	**65.10** ± 1.23	58.50 ± 1.11	**62.01** ± 1.09	**74.44** ± 1.42

in all datasets. In Tab. 1 and Tab. 2, we have provided the test accuracy all in-domain datasets and one cross-domain dataset: mini-CUB for both transductive and non-transductive settings. The performance of cross-domain dataset [16] is reported in Tab. 3.

Non-transductive results. For Tab. 1, `DUNT` uses miniimagenet as auxiliary data for CUB datasets, uses tieredimagnet as auxiliary data for the cross-domain mini-CUB, and uses CUB as auxiliary data for the rest datasets. The results show that, with extra auxiliary datasets, `DUNT` has outperformed the general data-cleansing techniques and other noisy FSL methods in general. In Tab. 1, the performance of SPL and FSR are generally worse than that of TraNFS and RNNP, which confirms our assumption that by removing potential noisy data, the data cleansing technique might harm the performance of FSL due to the lack of information, which can be compensated with our auxiliary data design. And compared to TraNFS and RNNP, `DUNT` has reduced the negative impact of the noisy labels more effectively.

Transductive results. In Tab. 2, we consider the transductive setting where the auxiliary dataset for all the methods are unlabeled test data of the same dataset. The results show that our method ranks the top among all the transductive methods. Compared to Tab. 1, some of the transductive methods underperform non-transductive ones due to the following potential reasons: i) the test dataset is smaller than other auxiliary datasets. Therefore, the non-transductive dataset has more information to provide. Take miniImageNet as an example, the test data of the miniImageNet dataset only contains 1000 images and 10 classes, while the auxiliary dataset CUB contains 200 categories and 12000 images, the abundance and diversity of the auxiliary helps greatly improve the model's performance over the primary tasks more than the close test data does; and ii) the auxiliary dataset provide more valuable fine-grained tasks that benefits representation learning. For instance, when training miniImageNet with CUB as an auxiliary, both datasets contains natural images which are closely related, and

Table 2: Transductive auxiliary datasets

Methods	CUB	mini	mini-CUB	tiered	CIFAR-FS
Task	5-way 1-shot				
PN+Co	49.10 ± 1.21	44.44 ± 0.96	36.48 ± 0.98	40.36 ± 0.81	52.51 ± 0.92
PN+Co+	49.45 ± 1.19	45.08 ± 0.91	39.54 ± 1.09	40.19 ± 1.10	51.27 ± 1.11
STARTUP	52.02 ± 1.20	48.79 ± 1.18	40.08 ± 1.22	44.48 ± 0.92	60.69 ± 1.31
DDNet	52.55 ± 1.39	50.82 ± 0.87	41.11 ± 1.89	45.11 ± 1.33	60.25 ± 1.46
DUNT	**55.01** ± 1.12	**52.38** ± 0.85	**42.07** ± 1.10	**49.29** ± 1.26	**61.90** ± 1.14
Task	5-way 5-shot				
PN+Co	65.42 ± 1.22	59.57 ± 1.20	56.90 ± 1.29	55.10 ± 1.21	66.89 ± 1.26
PN+Co+	67.92 ± 1.11	60.43 ± 1.13	54.29 ± 1.10	54.54 ± 0.84	67.22 ± 0.71
STARTUP	68.74 ± 1.21	62.00 ± 1.24	51.07 ± 1.01	55.92 ± 0.99	74.04 ± 0.84
DDNet	69.21 ± 1.34	63.70 ± 2.01	53.87 ± 0.44	57.22 ± 0.75	73.92 ± 1.20
DUNT	**71.20** ± 1.12	**66.46** ± 0.98	**59.05** ± 1.01	**61.09** ± 0.90	**76.11** ± 0.98

Table 3: Cross-domain Datasets

Methods	EuroSAT	CropDisease	ChestX	ISIC	EuroSAT	CropDisease	ChestX	ISIC
	5-way 1-shot				5-way 5-shot			
ProtoNet	52.52± 0.52	56.32± 0.89	28.90± 0.43	30.11± 1.00	61.60± 0.70	78.89± 0.83	33.04± 0.71	36.95± 0.90
PN+SPL	52.89± 1.21	58.95± 2.04	27.68± 0.52	31.79± 1.09	67.60± 1.23	80.25± 1.19	31.13± 1.18	37.05± 1.15
PN+FSR	47.53± 1.21	56.97± 1.23	28.40± 0.61	33.58± 0.92	61.96± 0.90	75.93± 0.86	31.43± 1.19	36.42± 1.22
RNNP	51.86± 1.20	56.56± 1.21	27.39± 0.61	31.26± 0.80	67.97± 1.34	78.43± 1.29	31.16± 1.25	37.84± 1.19
TraNFS	50.61± 1.10	58.39± 0.91	28.26± 0.98	32.84± 1.11	67.93± 1.11	81.51± 1.23	32.16± 1.16	36.95± 1.18
DCML	52.33± 1.02	58.36± 0.46	27.54± 0.69	31.96± 0.80	68.01± 0.90	79.11± 0.98	31.64± 1.00	35.44± 1.10
PN+Co	56.56± 1.08	59.13± 1.14	26.67± 0.73	29.16± 1.13	72.08± 1.15	78.59± 2.11	30.45± 1.17	34.74± 1.09
PN+Co+	52.18± 1.11	58.18± 1.10	28.23± 0.87	30.63± 1.06	65.57± 1.10	78.47± 0.90	30.77± 0.82	33.68± 0.91
STARTUP	52.74± 1.22	59.43± 1.23	27.75± 0.95	32.84± 1.13	69.78± 1.17	80.12± 1.22	31.69± 1.30	39.58± 0.99
DDNet	55.21± 1.83	60.09± 1.09	28.66± 0.98	34.10± 1.01	70.09± 1.90	81.90± 2.09	32.01± 1.89	39.44± 1.13
DUNT	**57.10**± 1.29	**61.49**± 1.01	**29.78**± 0.90	**35.80**± 0.77	**72.81**± 1.10	**83.20**± 1.08	**33.94**± 1.20	**40.31**± 0.90

CUB dataset is a collection of bird species, which provide fine-grained tasks for learning a better feature representation, while the miniImageNet test data only provide coarse-grained task, which makes it less useful in some extend. As for DUNT, the performance stays stable regardless the two settings, showing that our framework has the potential to fully utilize any extra information provided.

Cross-domain results. In Tab. 3, we present the performance on the cross-domain datasets. The training procedure follows the setting in [16], which uses miniImageNet as base dataset for all the methods. For CoTeaching, CoTeaching+, STARTUP, DDNet and DUNT, we use the unlabeled test data EuroSAT, CropDisease, ChestX, and ISIC as auxiliary dataset, respectively. In cross-domain datasets, the transductive methods generally outperform the non-transductive ones due to the utilization of the unlabeled test data that reduces the domain gap between the source domain and the target domain, which is critical when the they are extremely different. DUNT has used the Wasserstein regularization to further reduce the differences. Hence, it ranks the highest in the cross-domain datasets.

4.3 Ablation Study

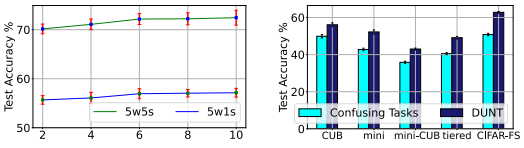
In this section, we study the effectiveness of different components and impact of hyperparameters. For more ablations, please refer to Appendix C.

Effectiveness of different components. Tab. 4 shows the ablation study with different elements of DUNT on CUB and miniImageNet datasets under non-transductive setting. We first show the performance of only using the primary data, which is vanilla ProtoNet (**Primary**). It can be observed that each of these components have a significant impact on the overall performance. Auxiliary data (**Auxiliary**) and episode cleansing (**EC**) serve pivotal functions, with the former introducing additional data and the latter eliminating noisy data. The regularization term (**Reg**) aids in extracting domain-invariant features, contributing to further performance improvement. Task selection (**TS**) mitigates the risk of overfitting, thereby enhancing overall performance.

Table 4: Impact of different components

Primary	Auxiliary	Reg	EC	TS	CUB	mini
✓	✗	✗	✗	✗	$42.57_{\pm 0.34}$	$39.95_{\pm 1.06}$
✓	✗	✗	✓	✗	$48.55_{\pm 1.51}$	$43.44_{\pm 1.49}$
✓	✓	✗	✗	✗	$50.19_{\pm 0.71}$	$48.90_{\pm 0.77}$
✓	✓	✗	✓	✗	$53.55_{\pm 0.97}$	$50.02_{\pm 0.81}$
✓	✓	✗	✗	✓	$52.70_{\pm +1.01}$	$49.50_{\pm 1.66}$
✓	✓	✗	✓	✓	$54.31_{\pm 0.99}$	$50.18_{\pm 0.87}$
✓	✓	✓	✗	✗	$51.90_{\pm 0.89}$	$49.18_{\pm 0.76}$
✓	✓	✓	✓	✗	$55.47_{\pm 0.79}$	$50.73_{\pm 0.60}$
✓	✓	✓	✗	✓	$54.79_{\pm 1.10}$	$49.91_{\pm 1.00}$
✓	✓	✓	✓	✓	**56.94** $_{\pm 1.03}$	**52.22**$_{\pm 1.10}$

Effectiveness of task selection. We study the impact of task selection through the threshold γ_{div}. As mentioned earlier, to streamline the hyper-parameter search process, we opt for an approach that involves upsampling M unsupervised tasks. Subsequently, we select only the most diverse task to be included in the valuable task group, instead of specifying a real value for the diverseness threshold. In this ablation, the number of upsampled tasks varies from 2-10. As depicted in Fig. 3 (a), the performance improves alongside an increase in the parameter M, but this comes at the cost of heightened computational demands. Consequently, considering efficiency, we opt for $\gamma_{\text{div}} = 6$ for all experiments. In Fig. 3 (b), we test the effect of using different unsupervised task: select confusing tasks and select valuable tasks. It is evident from the results that the inclusion of confusing tasks significantly deteriorates performance, whereas the utilization of valuable tasks leads to a substantial improvement.



(a) (b)

Fig. 3: DUNT on CUB with different γ_{div} (a); DUNT with different task selection (b)

5 Conclusion

We propose a novel FSL framework with auxiliary tasks that utilize carefully selected unlabeled data under noisy settings. To better align the auxiliary distribution with the primary one, we propose a regularization term based on the Wasserstein distance for learning a domain-invariant representation. To enhance robustness against label noise, we introduce episode cleansing for examples in the primary task and adopt a diverse task selection strategy for the unsupervised auxiliary tasks. Our framework is theoretically and experimentally proved to be effective and beneficial.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
3. Bertinetto, L., Henriques, J.F., Torr, P.H., Vedaldi, A.: Meta-learning with differentiable closed-form solvers. In: Proceedings of the International Conference on Learning Representations (2019)
4. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020)
5. Cheng, D., Liu, T., Ning, Y., Wang, N., Han, B., Niu, G., Gao, X., Sugiyama, M.: Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16630–16639 (2022)
6. Chobola, T., Vašata, D., Kordík, P.: Transfer learning based few-shot classification using optimal transport mapping from preprocessed latent space of backbone neural network. In: AAAI Workshop on Meta-Learning and MetaDL Challenge. pp. 29–37. PMLR (2021)
7. Courty, N., Flamary, R., Habrard, A., Rakotomamonjy, A.: Joint distribution optimal transportation for domain adaptation. Advances in Neural Information Processing Systems **30** (2017)
8. Courty, N., Flamary, R., Tuia, D.: Domain adaptation with regularized optimal transport. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 274–289. Springer (2014)
9. Deleu, T., Würfl, T., Samiei, M., Cohen, J.P., Bengio, Y.: Torchmeta: A Meta-Learning library for PyTorch (2019), https://arxiv.org/abs/1909.06576, available at: https://github.com/tristandeleu/pytorch-meta
10. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1126–1135 (2017)
11. Flamary, R., Courty, N., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Trans. Pattern Anal. Mach. Intell **1** (2016)
12. Ghosh, A., Kumar, H., Sastry, P.S.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
13. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8059–8068 (2019)
14. Goldblum, M., Fowl, L., Goldstein, T.: Adversarially robust few-shot learning: A meta-learning approach. Advances in Neural Information Processing Systems **33**, 17886–17895 (2020)
15. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)
16. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 124–141. Springer (2020)

17. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Proceedings of the Conference on Neural Information Processing Systems (2018)
18. Han, C., Fan, Z., Zhang, D., Qiu, M., Gao, M., Zhou, A.: Meta-learning adversarial domain adaptation network for few-shot text classification. In: Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021. Findings of ACL, vol. ACL/IJCNLP 2021, pp. 1664–1673. Association for Computational Linguistics (2021)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
20. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise. Advances in neural information processing systems **31** (2018)
21. Islam, A., Chen, C.F.R., Panda, R., Karlinsky, L., Feris, R., Radke, R.J.: Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. Advances in Neural Information Processing Systems **34**, 3584–3595 (2021)
22. Jiang, L., Meng, D., Mitamura, T., Hauptmann, A.G.: Easy samples first: Self-paced reranking for zero-example multimedia search. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 547–556 (2014)
23. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning. pp. 2304–2313. PMLR (2018)
24. Khodadadeh, S., Boloni, L., Shah, M.: Unsupervised meta-learning for few-shot image classification. Advances in neural information processing systems **32** (2019)
25. Killamsetty, K., Li, C., Zhao, C., Iyer, R., Chen, F.: A reweighted meta learning framework for robust few shot learning. arXiv preprint arXiv:2011.06782 (2020)
26. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Proceedings of the Conference on Neural Information Processing Systems. vol. 1, p. 2 (2010)
27. Li, W., Wang, L., Zhang, X., Qi, L., Huo, J., Gao, Y., Luo, J.: Defensive few-shot learning. IEEE transactions on pattern analysis and machine intelligence (2022)
28. Liang, K.J., Rangrej, S.B., Petrovic, V., Hassner, T.: Few-shot learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9089–9098 (2022)
29. Lu, J., Jin, S., Liang, J., Zhang, C.: Robust few-shot learning for user-provided data. IEEE transactions on neural networks and learning systems **32**(4), 1433–1447 (2020)
30. Masi, I., Trn, A.T., Hassner, T., Sahin, G., Medioni, G.: Face-specific data augmentation for unconstrained face recognition. International Journal of Computer Vision **127**, 642–667 (2019)
31. Mazumder, P., Singh, P., Namboodiri, V.P.: Rnnp: A robust few-shot learning approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2664–2673 (2021)
32. Medina, C., Devos, A., Grossglauser, M.: Self-Supervised Prototypical Transfer Learning for Few-Shot Classification. arXiv preprint arXiv:2006.11325 (2020)
33. Motiian, S., Jones, Q., Iranmanesh, S., Doretto, G.: Few-shot adversarial domain adaptation. Advances in neural information processing systems **30** (2017)
34. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749 (2021)

35. Phoo, C.P., Hariharan, B.: Self-training for few-shot transfer across extreme task differences. In: Proceedings of the International Conference on Learning Representations (2021)
36. Qin, T., Li, W., Shi, Y., Gao, Y.: Diversity helps: Unsupervised few-shot learning via distribution shift-based data augmentation. arXiv preprint arXiv:2004.05805 (2020)
37. Que, X., Yu, Q.: Dual-level curriculum meta-learning for noisy few-shot learning tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 14740–14748 (2024)
38. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: Proceedings of the International Conference on Learning Representations (2018)
39. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning. pp. 4334–4343. PMLR (2018)
40. Sahoo, D., Le, H., Liu, C., Hoi, S.C.: Meta-learning with domain adaptation for few-shot learning under domain shift (2018)
41. Shui, C., Li, Z., Li, J., Gagné, C., Ling, C.X., Wang, B.: Aggregating from multiple target-shifted sources. In: International Conference on Machine Learning. pp. 9638–9648. PMLR (2021)
42. Shui, C., Zhou, F., Gagné, C., Wang, B.: Deep active learning: Unified and principled method for query and training. In: International Conference on Artificial Intelligence and Statistics. pp. 1308–1318. PMLR (2020)
43. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
44. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in neural information processing systems (2020)
45. Sun, Q., Liu, Y., Chua, T.S., Schiele, B.: Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 403–412 (2019)
46. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: International Conference on Machine Learning (2017)
47. Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., Madry, A.: From imagenet to image classification: Contextualizing progress on benchmarks. In: International Conference on Machine Learning. pp. 9625–9635. PMLR (2020)
48. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in neural information processing systems. pp. 3630–3638 (2016)
49. Wang, X., Hua, Y., Kodirov, E., Robertson, N.M.: Imae for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters. arXiv preprint arXiv:1903.12141 (2019)
50. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 322–330 (2019)
51. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)

52. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
53. Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., Sugiyama, M.: Part-dependent label noise: Towards instance-dependent label noise. Advances in Neural Information Processing Systems **33**, 7597–7610 (2020)
54. Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., Sugiyama, M.: Are anchor points really indispensable in label-noise learning? Advances in neural information processing systems **32** (2019)
55. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020)
56. Xu, R., Yang, X., Liu, B., Zhang, K., Liu, W.: Selecting task with optimal transport self-supervised learning for few-shot classification. arXiv preprint arXiv:2204.00289 (2022)
57. Xu, Y., Cao, P., Kong, Y., Wang, Y.: L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. Advances in neural information processing systems **32** (2019)
58. Yang, Y., Liang, K.J., Carin, L.: Object detection as a positive-unlabeled problem. arXiv preprint arXiv:2002.04672 (2020)
59. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: International Conference on Machine Learning. pp. 7164–7173. PMLR (2019)
60. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
61. Zhang, Z., Pfister, T.: Learning fast sample re-weighting without reward data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 725–734 (2021)