DOMAIN-INVARIANT AUXILIARY LEARNING FOR ROBUST FEW-SHOT PREDICTIONS FROM NOISY DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern meta-learning approaches produce state-of-the-art performance by imitating the test condition for few-shot learning (FSL) using episodic training. However, overfitting and memorizing corrupted labels has been a long-standing issue. Data cleansing offers a promising solution for dealing with noisy labels. Nevertheless, in FSL, data cleansing exacerbates the severity of the problem as the available training data becomes much more limited and the model is typically inadequately trained. In this work, we address the overfitting in a noisy setting by exploiting auxiliary tasks to learn a better shared representation. Unsupervised auxiliary tasks are designed with no extra labeling overhead and Wasserstein distance is leveraged to align the primary and auxiliary distributions that ensures the learned knowledge is domain-invariant. Building upon the theoretical advances on PAC-Bayesian analysis, we gain ground on deriving novel generalization bounds of meta-learning with auxiliary tasks and under the effect of noisy corruptions. Extensive experiments on FSL tasks with noisy labels are conducted to show the effectiveness and robustness of our proposed method.

1 INTRODUCTION

A gospel for meta-learning is the shared common underlying structure across tasks and the imitation of the test environment during training. The episodic meta-learning methods capture the statistical dependence on the shared latent information by a bi-level optimization problem among a meta-model θ shared across tasks and a set of task-specific models θ_i's for individual tasks. Due to the dependency, a task-specific model θ_i for a novel task can be adapted from θ via few training examples in a few steps. Following the idea of MatchingNet in (Vinyals et al., 2016), an episode is designed to include a support set and a query set to mimic the few-shot task by sub-sampling classes as well as data points. In general, a task-specific model is learned and adapted from the meta-model using the support set data, and the meta-model is updated by the knowledge accumulated from each task and evaluated on the query set data.

While meta-learning models produce state-of-the-art performance for many FSL applications, overfitting and memorizing corrupted labels is an inevitable issue (Vinyals et al., 2016; Zhang et al., 2021; Yao et al., 2021). Lack of training tasks may fail to cover the entire distribution causing rote memorization of a smaller and distorted (training) distribution. Noisy labels can corrupt the shared representation across tasks during meta-learning. Both situations may lead to poor adaptation to new tasks, which hurts the generalization from meta-training to meta-testing. To counter corruption from the noisy data, data cleansing has emerged as an effective means for handling noisy labels and avoiding harmful overfitting (Jiang et al., 2018; Han et al., 2018; Yu et al., 2019). It guides model training by selecting clean instances out of the noisy ones, and then either removes or relabels the noisy ones before updating. The small loss technique is normally used to filter out the noisy samples, which tend to incur a bigger loss. Nevertheless, in FSL, data cleansing exacerbates the severity of the problem as the available training data becomes much more limited and the model is typically inadequately trained.

The ability to exploit more data from one problem that generalizes to another (Ajakan et al., 2014) offers a promising direction to learn from noisy few-shot tasks. Trained alongside the primary task, the auxiliary tasks make up for the limited training data. By choosing auxiliary tasks that are easy to learn from and support the primary tasks, it is instrumental to construct meaningful representations and avoid overfitting to spurious correlations caused by noisy labels. However, designing helpful

auxiliaries for given primary tasks is challenging. For example, in image classification, popular choices of auxiliary tasks include rotation (Gidaris et al., 2018), masking (Doersch et al., 2015), and patch shuffling (Noroozi & Favaro, 2016), which require both domain expertise and additional annotation efforts (Navon et al., 2021). We address the problem by designing an unsupervised auxiliary counterpart, which completely removes the burden of an expensive labeling process (Khodadadeh et al., 2019). One remaining concern is that using extra data from other domains may induce covariate shift due to the mismatch of the extra data and the current task. In this paper, we propose to conduct novel Domain-Invariant Auxiliary Learning, referred to as DIAL, which integrates auxiliary tasks in an adversarial way to ensure domain invariance during auxiliary learning.

To effectively transfer knowledge across the primary and auxiliary tasks while avoiding the covariate shift, the learned representation should avoid domain-specific knowledge from the auxiliary tasks while encouraging common domain-invariant knowledge distilled in this process (Ben-David et al., 2010). According to this theory, unsupervised domain adaptation based on optimal transportation learns representations from the source domain for discrimination in the target domain by measuring and reducing the their disparity, such as Maximum Mean Discrepancy (MMD), Wasserstein distance, or KL-divergence in an adversarial way (Courty et al., 2014; 2017; Shen et al., 2018). Meta-learning shares the intuition of learning from the base classes during training and applying to the novel unseen classes during testing whereas the the base classes and the novel classes are non-overlapped. Therefore, it is intuitive to employ this idea for our primary and auxiliary tasks to make their feature representations indistinguishable. To this end, we propose to minimize the discrepancy between the primary and the auxiliary distributions using optimal transport. Specifically, instead of directly minimizing domain discrepancy, we utilize a domain critic neural network to estimate empirical Wasserstein distance between the primary and auxiliary samples and optimize the feature extractor network to minimize the estimated Wasserstein distance in an adversarial manner (Ajakan et al., 2014). By iterative adversarial training, we finally learn feature representations that are both discriminative and domain-invariant.

Building upon the recent advances in PAC-Bayesian analysis of deep learning and meta learning (Amit & Meir, 2018; Rothfuss et al., 2021; Pentina & Lampert, 2014), we provide novel generalization bounds of meta-learning with key insights to quantify the contribution of auxiliary tasks considering impacts of noisy corruptions. In particular, we theoretically prove that: (i) the auxiliary information helps to tighten the gap between the true generalization risk and the empirical risk; and (ii) the hazard of the noisy labels under the PAC-Bayesian framework. Our contribution is threefold:

- a simple yet effective DIAL framework using auxiliary tasks to learn a robust domain-invariant representation for better generalization and adaptation in unseen few-shot tasks,
- a thorough theoretical analysis of the proposed auxiliary tasks under data cleansing to establish novel PAC-Bayesian bounds that provide key insights on the contribution of the auxiliary tasks and quantify the impact of label noise,
- a series of comprehensive experiments conducted on benchmark datasets with synthetic label noises to demonstrate the effectiveness and robustness of the proposed DIAL framework.

2 Related Works

Few-shot learning. Liu et al. (2019b) first propose a transductive few-shot learning method by utilizing the query set for transductive inference. Li et al. (2019) focus on the semi-supervised few-shot learning which utilizes the unlabeled data by predicting their pseudo labels and iterative self-training. Phoo & Hariharan (2021) propose to self-training the unlabeled data from the test to deal with few-shot learning with extreme differences between the training and testing. Qiao et al. (2019) integrates meta-learning with transductive inference by formulating a semi-definite programming problem for the adaptation procedure. Unlike other related few-shot learning works using unlabeled data such as Ren et al. (2018); Yu et al. (2020); Hou et al. (2019), we explicitly minimize the domain discrepancy between the primary and auxiliary distribution by introducing a regularization based on optimal transport, which guarantees the effectiveness of the auxiliary data due to minimal domain shift. Other works (Han et al., 2021; Sahoo et al., 2018; Motiian et al., 2017) use domain adaptation techniques in few-shot learning. Our work is essentially different since we use unsupervised auxiliary data during training to update the meta-model for the compensation of the lack of training data and utilize adversarial training as a regularizer to ensure that the knowledge extracted from the auxiliary tasks are invariant to different distributions.

Auxiliary learning. Auxiliary learning trains additional auxiliary tasks to improve the generalization ability of the primary task. To better assist the primary task, the auxiliary tasks could be related tasks, such as fine-grained classification of the primary task (Liu et al., 2019a). In (Zhu et al., 2020), the primary task learns to navigate following natural language instructions, while the auxiliary tasks provide additional training signals to help the agent acquire knowledge of semantic representations in order to reason about its activity and build a thorough perception of the environment. In reinforcement learning (RL) (Veeriah et al., 2019; Jaderberg et al., 2016), auxiliary tasks drive representation learning to aid main task. In our work, we utilized unsupervised few-shot tasks to assist the primary ones without introducing extra labeling effort.

Domain Adaptation and Optimal Transport. The optimal transportation (OT) cost is used to measure the difference between distributions supported on high-dimensional space using SGD (Arjovsky et al., 2017; Tolstikhin et al., 2017). For a constant $\xi \geq 1$, the ξ-Wasserstein metric between distributions $\mathbb{P}_X$ and $\mathbb{P}_{X'}$ is defined as:

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

where $\mathbf{x}, \mathbf{x}'$ denotes the random variable 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 $\int \gamma(\mathbf{x}, \mathbf{x}')d\mathbf{x}' = \mathbb{P}_{X'}$. $d(\mathbf{x}, \mathbf{x}')$ is cost function for moving from $\mathbf{x}$ to $\mathbf{x}'$. Eq. (1) is the primal form of the Wasserstein metric, particularly for the case of $\xi = 1$ (Arjovsky et al., 2017). 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 (Courty et al., 2014; Flamary et al., 2016; Courty et al., 2017) . In this paper, we propose to minimize the Wasserstein distance between the primary distribution and the auxiliary distribution to ensure the learned knowledge are domain-invariant and discriminative to future unseen tasks.

3 METHODOLOGY

Meta-learning Given the primary distribution $\mathcal{D}^{pri}$ over $\mathcal{X} \times \mathcal{Y}$, a meta-model θ is trained in an episodic way by sampling batches of the episodes (*i.e.,* tasks) $\mathcal{T}$. The episodic sampling process of a N-way K-shot classification task includes two steps: first randomly samples N classes from the base class $\mathcal{C}_\mathcal{B}$ and then randomly sample K images as the support set $\mathcal{S}^{sup} = \{(x_i^j, y_i^j)\}_{j=1}^K$, Q images as the query set $\mathcal{S}^{que} = \{(x_i^j, y_i^j)\}_{j=1}^Q$. The sampled tasks form the task distribution τ. The objective of episodic meta-learning is a bi-level optimization formulated as a loss over an episode as the outer-level:

$$\min_\theta \mathcal{L}(\theta, \mathcal{D}) = \sum_{\mathcal{T}_i \sim \tau} \mathcal{L}_{\mathcal{T}_i}(\theta_i, \mathcal{S}_i^{sup}, \mathcal{S}_i^{que}) \tag{2}$$

whereas in the inner-level of the optimization, the feature extractor of task $\mathcal{T}_i$ parameterized by θ_i is first initialized as the meta-model θ then adapted to a specific task within a few steps: $\theta_i \leftarrow \theta - \iota \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\theta, \mathcal{S}^{sup})$, where ι is the learning rate. The meta-test is perform similarly to the inner-level except that the episodes are formed by unseen novel classes sampled from $\mathcal{C}_\mathcal{N}$, and $\mathcal{C}_\mathcal{B} \cap \mathcal{C}_\mathcal{N} = \emptyset$.

Episode Sampling To avoid the model being corrupted by the noisy labels, data cleansing technique is utilized to each episode. We call this episode sampling. For both tasks, each example in the query set is re-weighted according to the loss. Taking the primary task as an example, the inner-loss is reformulated as:

$$\mathcal{L}_{\mathcal{T}_i}(\theta_i, \mathcal{S}^{que}) = \frac{1}{|\mathcal{S}_i^{que}|} \sum_{x_i^j, y_i^j \in \mathcal{S}_i^{que}} w_i^j \ell_i^j(x_i^j, y_i^j, \theta_i) \tag{3}$$

where ℓ_i^j corresponds to the loss of j-th example in the i-th task, w_i^j is the corresponding sample weight given by: $w_i^j = 1(\ell_i < \gamma)$ using a predefinded hyperparameter γ, where $1(\cdot)$ is an indicator function.

3.1 DOMAIN-INVARIANT AUXILIARY LEARNING

To introduce extra data without increasing the labeling effort, we use the unsupervised few-shot tasks sampled from an auxiliary distribution $\mathcal{D}^{aux}$, which is defined over $\mathcal{X}$. And its corresponding

task distribution is defined as τ'. With no category information provided, the unsupervised task $\mathcal{T}'$ is constructed by directly samples N images from the $\mathcal{D}^{aux}$ and treat each image as its own class to form the support set of a N-way 1-shot task: $\mathcal{S}_i'^{sup} = \{(\mathbf{x}_i'^1, \tilde{y}_i'^1), ..., (\mathbf{x}_i'^N, \tilde{y}_i'^N)\}$. The query set is acquired by augmenting the images in the support set: $\mathcal{S}_i'^{que} = \{(A(\mathbf{x}_i'^N), \tilde{y}_i'^N), ..., (A(\mathbf{x}_i'^N), \tilde{y}_i'^N)\}$, where A is an augmentation function, such as flipping the image, grayscale, rotation or a combination, and the query examples $A(\mathbf{x}_i'^j)$ are the augmented images of the support example $\mathbf{x}_i'^j$.

Using auxiliary learning, we can update the meta-model with both primary and auxiliary tasks in a bi-level optimization manner as shown in Eq. (2). Consider the potential noise during label collection of the primary task, and the randomness of choosing image as its class in the auxiliary task, we applied episode sampling to each of the primary and auxiliary task. Eq. (2) is then reformulated as:

$$\min_{\theta} \mathcal{L}(\theta, \mathcal{D}^{pri}, \mathcal{D}^{aux}) = \sum_{\mathcal{T}_i \sim \tau} \mathcal{L}_{\mathcal{T}_i}(\theta_i, \mathcal{S}_i^{sup}, \hat{\mathcal{S}}_i^{que}) + \sum_{\mathcal{T}_i' \sim \tau'} \mathcal{L}_{\mathcal{T}_i}(\theta_i, \mathcal{S}_i^{sup}, \hat{\mathcal{S}}_i'^{que}) \tag{4}$$

where $\hat{\mathcal{S}}_i^{que}$ and $\hat{\mathcal{S}}_i'^{que}$ denote the clean query set after episode sampling.

To avoid covariate shift between the prime and auxiliary distributions, the Wasserstein distance between them is incorporated to ensure that the learned feature representation are domain-invariant to different distributions so that the classifier trained on the shared representations can better generalize across domains. The overall objective of `DIAL` is therefore formulated as:

$$\min_{\theta} \alpha_1 \mathcal{L}(\theta, \mathcal{D}^{pri}) + (1 - \alpha_1)\mathcal{L}(\theta, \mathcal{D}^{aux}) + \alpha_2 \mathcal{W}_1(\mathcal{D}^{pri}, \mathcal{D}^{aux}) \tag{5}$$

For simplification, we use $\mathcal{L}(\theta, \mathcal{D}^{pri})$ and $\mathcal{L}(\theta, \mathcal{D}^{aux})$ to denote the loss of the primary tasks and auxiliary tasks and use the hyper-parameter α_1 to balance the contribution of them. The third Wasserstein distance term reflects an optimal transport cost for moving one distribution to another, which is defined in Eq. (1). A smaller transport cost means a better coverage of the distribution. α_2 is a hyper-parameter. By minimizing Eq. (5), we simultaneously minimize the empirical error for learning the primary tasks and the auxiliary tasks, and the Wasserstein-1 distance. The minimization of the Wasserstein-1 distance encourages a better distribution matching of $\mathcal{D}^{pri}$ and $\mathcal{D}^{aux}$ so that the auxiliary data behave more similar to the training data and we essentially enlarge the effective training data as there is minimal domain shift.

Since directly computing the Wasserstein-1 distance is computationally intractable, we recast the objective into the Kantorovich-Rubinstein duality with a 1-Lipschitz neural network called a critic parameterized by θ^d which estimates the supremum of Eq. (1), and the main model optimize:

$$\min_{\theta^f} \max_{\theta^d} \mathcal{R}(\theta^f, \mathcal{D}^{pri}, \mathcal{D}^{aux}) + \alpha_2 \mathcal{E}(\theta^f, \theta^d, \mathcal{D}^{pri}, \mathcal{D}^{aux}) \tag{6}$$

where $\mathcal{R}(\cdot)$ is the prediction loss of the primary and auxiliary tasks and $\mathcal{E}(\cdot)$ is the adversarial loss. By representing the meta-model in a parametrized function $h(\theta^f, \mathbf{x}, y) = h(\mathbf{x}, y)$ and the critic model in function $g(\mathbf{x}, \mathbf{x}', \theta^f, \theta^d) = g(\mathbf{x})$ with restriction that $g(\mathbf{x})$ being 1-Lipschitz, we formulate the terms in Eq. (6) as

$$\mathcal{R}(\theta^f, \mathcal{D}^{pri}, \mathcal{D}^{aux}) = \alpha_1 \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{pri}} \ell(h(\mathbf{x}, y)) + (1 - \alpha_1)\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^{aux}} \ell(h(\mathbf{x}, y)) \tag{7}$$

$$\mathcal{E}(\theta^f, \theta^d, \mathcal{D}^{pri}, \mathcal{D}^{aux}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}^{pri}}[g(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}^{aux}}[g(\mathbf{x})] \tag{8}$$

For simplification, we use $(\mathbf{x}, y) \sim \mathcal{D}$ to denote the two-step task sampling $\mathcal{D}_i \sim \tau$ and $(\mathbf{x}, y) \sim \mathcal{D}_i$ in Eq. (7). In this duality, the critic works as an adversarial discriminator that distinguishes the origin of the samples. The learning proceeds with minimizing the prediction loss to learn discriminative features whereas matching the auxiliary to the primary for learning domain invariant representations. The optimization of Eq. (6) is performed by alternatively updating the model parameters θ^f and the critic parameters θ^d using SGD. During this process, θ^f follows as usual the opposite direction of the gradient whereas θ^d follows gradient direction. To satisfy the 1-Lipschitz constraint, the critic θ^d is implemented with gradient penalty (Gulrajani et al., 2017). The detailed training process is shown in Algorithm 1 in the Appendix.

3.2 PAC-Bayesian Analysis of DIAL

The PAC-Bayesian theory combines the informative priors of Bayesian methods with the distribution-free PAC-guarantees (McAllester, 1999). In general, it first introduces probability measures over the hypothesis space $\mathcal{H}$, and assumes a prior $P \in \mathcal{M}(\mathcal{H})$ independent of the observed

data and a posterior $Q \in \mathcal{M}(\mathcal{H})$ obtained after observing the data. In the context of meta-learning, the meta-model is defined as P and the task-specific model is defined as $Q(S, P)$, meaning that the task-specific model is adapted from P within a few steps. We use $S = \{\mathcal{S}^{sup}, \mathcal{S}^{que}\}$ to denote the training data for simplification. The meta-learning PAC-Bayesian framework presumes a distribution over the meta-model P, named hyper-prior $\mathcal{P}(P) \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$. Given n tasks $S_1, ..., S_n$, the hyper-prior is updated to a hyper-posterior $\mathcal{Q}(P) \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$. In this section, we utilize the PAC-Bayesian framework to provide generalization guarantees for the proposed DIAL framework.

For DIAL, consider n primary tasks with datasets $S_1, ..., S_n$ and n auxiliary task with datasets $S'_1, ..., S'_n$ and n, the expected loss is formulated as:

$$\mathcal{L}(P, \tau, \tau') = \alpha_1 \mathbb{E}_{\mathcal{D}^{pri} \sim \tau} \mathbb{E}_{S \sim \mathcal{D}_m^{pri}} \mathcal{L}(Q, \mathcal{D}^{pri}) + (1 - \alpha_1) \mathbb{E}_{\mathcal{D}^{aux} \sim \tau'} \mathbb{E}_{S' \sim \mathcal{D}_m^{aux}} \mathcal{L}(Q, \mathcal{D}^{aux}) \quad (9)$$

where m is the number of examples in each dataset S. The loss for the primary task is specified as $\mathcal{L}(Q, \mathcal{D}) = \mathbb{E}_{h \sim Q} \mathcal{L}(h, \mathcal{D}) = \mathbb{E}_{h \sim Q} \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z)$, where $h \sim Q$ is a hypothesis sampled from the posterior distribution Q and we use $z = (x, y)$ to denote a input/output pair sampled from the data distribution. The loss for auxiliary is similarly defined with data sampled from the auxiliary distribution. The performance of the hyper-posterior $\mathcal{Q}$ on the task distribution τ, *i.e.,* the primary tasks, is measured by expected loss on new tasks using priors drawn from $\mathcal{Q}$, which is referred to as the transfer error:

$$\mathcal{L}(\mathcal{Q}, \tau, \tau') = \mathbb{E}_{P \sim \mathcal{Q}} \mathcal{L}(P, \tau, \tau') \quad (10)$$

The transfer error is intractable in practice. In PAC-Bayesian framework, it is approximated by the empirical multi-task error as follows:

$$\hat{\mathcal{L}}(\mathcal{Q}, S_{i=1}^n, S_{i=1}^{\prime n}) = \mathbb{E}_{P \sim \mathcal{Q}} \frac{1}{n} \sum_{i=1}^{n} [\alpha_1 \hat{\mathcal{L}}(Q(S_i, P), S_i) + (1 - \alpha_1) \hat{\mathcal{L}}(Q(S'_i, P), S'_i)] \quad (11)$$

where $\hat{\mathcal{L}}(Q(S_i, P), S_i) = \mathbb{E}_{h \sim Q} \hat{\mathcal{L}}(h, S_i) = \mathbb{E}_{h \sim Q} \frac{1}{m} \sum_{j=1}^{m} \ell(h, z_{ij})$ is the empirical error of each primary task $\mathcal{T}_i$ with corresponding data S_i given the prior P by averaging over the posterior distribution Q. And $\hat{\mathcal{L}}(Q(S'_i, P), S'_i)$ is similarly defined for the auxiliary task. Based on the above definitions, we now present a novel bound for DIAL with extra auxiliaries in Theorem 1.

Theorem 1 *(PAC-Bayes bound for vanilla DIAL) Given a hypothesis space $\mathcal{H}$, a base learner $Q : \mathcal{Z}^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, a fixed hyper-prior $\mathcal{P} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$, $\lambda > 0$, $\beta > 0$, a target environment τ and observed environment $\tau \cup \tau'$ where $\mathbb{E}_{\tau \cup \tau'}[\mathcal{D}] \geq \mathbb{E}_{\tau}[\mathcal{D}]$, and $\mathbb{E}_{\tau \cup \tau'}[m] = \mathbb{E}_{\tau}[m]$, then with probability at least $1 - \delta$ over samples $S_1 \in \mathcal{D}_1^{pri,m}, ..., S_n \in \mathcal{D}_n^{pri,m}, S'_1 \in \mathcal{D}_1^{aux,m}, ..., S'_n \in \mathcal{D}_n^{aux,m}$, we have for all base learners Q and hyper-posterior $\mathcal{Q}$, the following inequality holds:*

$$\mathcal{L}(\mathcal{Q}, \tau, \tau') \leq \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{P \sim \mathcal{Q}} \mathbb{E}_{h \sim Q} [\hat{\mathcal{L}}(h, S_i)] + c_1 + 2 \sum_{j=1}^{2} \varrho_j (\mathcal{W}_1(\hat{\mathcal{D}}^{pri}, \hat{\mathcal{D}}^{aux}) + \lambda_{pri} + c_2) \right)$$
$$+ \frac{1}{\lambda} \left(D_{KL}(\mathcal{Q} || \mathcal{P}) + \log \frac{2}{\delta} + \log \frac{\lambda^2}{4n} \right) \quad (12)$$

where $\hat{\mathcal{D}}^{pri} = \frac{1}{N_{pri}} \sum_{i=1}^{N_{pri}} \delta_{\{\mathcal{D}^{pri}\}}$ and $\hat{\mathcal{D}}^{aux} = \frac{1}{N_{pri}} \sum_{i=1}^{N_{aux}} \delta_{\{\mathcal{D}^{aux}\}}$ are the empirical distribution of the primary and auxiliary distribution, respectively, N_{pri} and N_{aux} are the number of examples in the datasets. ϱ_i is the weight of the domain, in DIAL, $\varrho_1 = \alpha_1$ and $\varrho_2 = 1 - \alpha_1$. $\lambda_{pri} = \min_h(\mathcal{L}(h, \mathcal{D}^{pri}) + \mathcal{L}(h, \mathcal{D}^{aux}))$, and c_1 and c_2 are defined in Eq. (20) and Eq. (21). $C(\delta, n, m, \lambda, \beta) = \frac{1}{\beta} \Psi(\beta, m) + \log \frac{4n}{\delta} + \frac{1}{\lambda} \left(\log \frac{2}{\delta} + \log \frac{\lambda^2}{4n} \right)$, and $\Psi(\beta, m) = \log \mathbb{E}_{h \sim P} \mathbb{E}_{S \in \mathcal{D}^m} \exp \left[\beta (\mathcal{L}(h, \mathcal{D}) - \hat{\mathcal{L}}(h, S)) \right]$. β and λ are hyperparameters with their popular choices are $\beta \propto m$ and $\lambda \propto n$. $D_{KL}(\cdot || \cdot)$ is the KL-divergence between two distributions.

Proof sketch. The detailed proof of Theorem 1 is given in Appendix C which includes three key steps. First, we prove the task-specific generalization bound by directly applying multi-source domain adaptation guarantees from Theorem 3 (Redko et al., 2017) to a single task, since the generalization bound for a single task accounts for the task-specific model's ability to perform well on unseen data. Second, for the task environment bound, whose responsibility is to learn an inductive bias for generalizing to new tasks, we use Donsker-Varahan's variational formula (Donsker &

Varadhan, 1975) and Markov's inequality to bound the task-level generalization error. Last, under the assumption that the task-specific model utilizes the inductive bias learned by the meta-model and adapt to a new task, we use a union bound to combine the task-specific and task environment bounds to complete the proof. ■

Remark. According to the choice of the hyperparameter λ, we can see that the term regrading $D_{KL}(\mathcal{Q}||\mathcal{P})$ in the RHS of the bound is reduced considerably since λ is increased as the the number of total task increased. Intuitively, the $D_{KL}(\mathcal{Q}||\mathcal{P})$ term measures the discrepancy between hyper-prior and the hyper-posterior, the more data (*i.e.,* tasks) we observed, the smaller the value of the $D_{KL}(\mathcal{Q}||\mathcal{P})$ will be since we get a richer hyper-posterior and close to the hyper-prior. $D_{KL}(\mathcal{Q}||\mathcal{P})$ is essentially a penalty term to capture overfitting, which accounts for the amount of knowledge we learned from the data. Therefore, our bound in Theorem 1 has successfully proved that by introducing the auxiliary tasks we can reduce this term and decrease the potential risk of overfitting. As in Eq. (5), the minimization of $\mathcal{W}_1(\cdot)$ ensures small domain shift that makes the auxiliary data more effective.The slightly increment $C(\delta, n, m, \lambda, \beta)$ is caused by the side effect of the using of the union bound, which can be compromised. In addition, with the learned domain-invariant representation, the term $\psi(\beta, m)$ can be further reduced.

In Theorem 1, we analyze the positive effect of introducing the auxiliary task without taking the noisy data into account for simplicity. However, the theory stands on a shaky ground since we did not consider the noise. In the next theorem, we take noise into consideration where we analyze their negative effect and how our episode sampling takes care of it. According to episode sampling, we remove the noisy samples in each task using data cleansing, which leads to $\mathbb{E}_{\tau\cup\tau'}[\mathcal{D}] \geq \mathbb{E}_{\tau}[\mathcal{D}]$, and $\mathbb{E}_{\tau\cup\tau'}[m] \leq \mathbb{E}_{\tau}[m]$, *i.e.,* the number of training samples in the meta-training tasks is smaller than the one in the meta-testing tasks, which lead to the following Theorem 2.

Theorem 2 *(PAC-Bayes bound for* `DIAL` *with sampling) Given a hypothesis space $\mathcal{H}$, a base learner $Q : \mathcal{Z}^m \times \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{H})$, a fixed hyper-prior $\mathcal{P} \in \mathcal{M}(\mathcal{M}(\mathcal{H}))$, $\lambda > 0$, $\beta > 0$, a target environment τ and observed environment $\tau \cup \tau'$ where $\mathbb{E}_{\tau\cup\tau'}[\mathcal{D}] \geq \mathbb{E}_{\tau}[\mathcal{D}]$, and $\mathbb{E}_{\tau\cup\tau'}[m] \leq \mathbb{E}_{\tau}[m]$, then with probability at least $1-\delta$ over samples $S_1 \in \mathcal{D}_1^m, ..., S_n \in \mathcal{D}_n^m, S'_1 \in \mathcal{D}_1'^m, ..., S'_n \in \mathcal{D}_n'^m$, and clean datasets $\tilde{S}_1 \subset S_1, ..., \tilde{S}_n \subset S_n, \tilde{S}'_1 \subset S'_1, ..., \tilde{S}'_n \subset S'_n$, where $\tilde{S}_i = \{z_{ij}\}_{j=1}^{m'}$ and $m' \leq m$, we have for all base learners Q and hyper-posterior $\mathcal{Q}$, the following inequality holds:*

$$\mathcal{L}(\mathcal{Q}, \tau) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P\sim\mathcal{Q}}\mathbb{E}_{h\sim Q}[\hat{\mathcal{L}}(h, S_i)] + c_1 + 2\sum_{j=1}^{2}\alpha_j(\mathcal{W}_1(\hat{\mathcal{D}}^{pri}, \hat{\mathcal{D}}^{aux}) + \lambda_{pri} + c_2)$$
$$+\frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \Psi(\lambda, 2n)\right) \tag{13}$$

where $C(\delta, \lambda, n, m') = \frac{1}{\beta}\left(\log\frac{2n}{\delta} + \Psi(\beta, m)\right) + \frac{1}{\lambda}\left(\log\frac{2}{\delta} + \Psi(\lambda, 2n)\right)$ *and* $\Delta_\lambda(\mathcal{P}, \tau, \tau') = \frac{1}{\lambda}\log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}_{S\sim\tau, S'\sim\tau'}\left[e^{\lambda(\mathbb{E}_{S\sim\tau, S'\sim\tau'}\mathcal{L}(P,\tau) - \mathcal{L}(P,\tau,\tau'))}\right]$.

Proof sketch. Detailed proof of Theorem 2 is provided in Appendix D, which includes three key steps. The first and last steps are similar to the proof of Theorem 1. In the second step, we use Markov's inequality to measure the mismatch of task environments caused by the noise and episode sampling and a penalty term $\Delta_\lambda(\mathcal{P}, \tau, \tau')$ is obtained due to the mismatch. ■

Remark. In Eq. (13), the emergence of the penalty term $\Delta_\lambda(\mathcal{P}, \tau, \tau')$ is caused by the mismatch of the training and testing environment by using the data sampling. This demonstrate that for alleviating the negative impact of noisy labels, we inevitably lost some information and cause a lost in the generalization gap as well. However, we manage to neutralize the penalty with clean data and avoid overfitting. That is, with clean data which is close to the true data distribution, we can assure a relatively smaller $D_{KL}(\mathcal{Q}||\mathcal{P})$ to balance the generalization bound.

4 Experiments

Datasets. We evaluate the effectiveness of the `DIAL` framework using the following benchmark datasets for FSL:

- *CUB* contains 11,788 images of 200 bird species (Snell et al., 2017). The classes are split into 100, 50, and 50 for train, test, and validation, respectively.

- *mini-imageNet (mini)* (Sun et al., 2019) contains 100 different classes with 600 images per class, which are split into 64, 20, and 16 for train, test and validation, respectively.
- *tieredImageNet (tiered)* (Ren et al., 2018) has 34 high-level categories and 608 sub-categories, with each sub-category consisting of around 1,300 images (Ren et al., 2018). These categories are split into 20 (351 sub-categories), 8 (160 sub-categories), and 6 (97 sub-categories) for train, test, and validation, respectively.
- *CIFAR-FS (CF)* (Bertinetto et al., 2019) includes 100 different classes with 600 images/class . These classes are further grouped into 20 super-classes. The 100 classes are split into 64, 20, 16 for train, test, and validation, respectively.
- *Cross-Domain (mini-C):* where the training classes are from miniImageNet dataset and the validation and testing classes are from the CUB dataset.

Baselines. We compare DIAL with the following baselines: ProtoNet (Snell et al., 2017), and other noise-robust methods including Weight Decay (Krogh & Hertz, 1991), SPL (Kumar et al., 2010), MixUp (Zhang et al., 2017), CoTeaching (Han et al., 2018), CoTeaching+ (Yu et al., 2019) and FSR-raw (Zhang & Pfister, 2021). For efficiency consideration, all methods are implemented based on ProtoNet (Deleu et al., 2019) with default parameters given by (Medina et al., 2020). All the experiments are conducted on a NVIDIA A100 GPU. A detailed discussion of the comparison methods is presented in the Appendix.

Noise setting. We generate synthetic label noises with random flips. For a more realistic consideration, different ratios of synthetic noises are applied to different classes randomly. The noise ratio in the meta-train classes varies as follows:

- CUB: the number of noisy labels in each class varies from 0 to 20. The first 20 classes contain 20 noisy labels, the 21-40 classes contain 10 noisy labels and the 41-60 classes contain 5 noisy labels in each class. The rest of 40 classes remain clean. The overall noise ratio is around 11.6%.
- miniImageNet: the number of noisy labels in each class varies from 0 to 200. The first 20 classes contain 200 noisy labels, classes 21-30 contain 100 noisy labels, and classes 31-40 contain 50 noisy labels. The last 24 classes remain clean. The overall noise ratio is around 14%.
- tieredImageNet: the number of noisy labels in each class varies from 0 to 500. The first 100 classes contain 500 noisy labels, the 101-200 classes contain 300 noisy labels, and the 201-300 classes contain 100 noisy labels. The remaining 51 classes remain clean. The overall noise ratio is around 20%.
- CIFAR-FS: each training class contains 60 noisy labels. The noise ratio is 10%.

Auxiliary data setting. For methods including DIAL, CoTeaching, CoTeaching+, we utilize the extra data during training. The choice of auxiliary data can be any datasets as long as they can benefit the primary task. For convenience, we use their own test data as the auxiliary data if without particularly specified, which makes it similar to transductive learning or unsupervised learning. However, our setting is essentially different since we only choose test data as auxiliary task for convenience, any close data can be utilized. For example, in the cross-domain mini-CUB benchmarks, using the images of novel classes from the CUB dataset as the auxiliary tasks seems to be a fairly reasonable choice. And in the ablation studies, we also study the effect of using different auxiliary data.

4.1 Experimental Results

Following the experiment design of (Vinyals et al., 2016), we report the average accuracy of 5 runs of each 5-way 1-shot and 5-way 5-shots experiments in Tab. 1. In the experiments, we use $\alpha_1 = 0.5$, $\alpha_2 = 0.2$, and $\gamma = 50\%$ for our methods in all datasets. As shown in Tab. 1, our proposed model outperforms all the baselines, highlighting the effectiveness of the DIAL framework. Compared to SPL and FSR, which use data cleansing techniques (*i.e.,* the small loss strategy or max margin) to remove data noise, our method compensates for the information loss with auxiliary tasks, obtaining a superior performance. DIAL also outperforms methods like Co-teaching and Co-teaching+, which also include extra data, indicating that the superiority of the domain-invariant representation. Other methods, like MixUp, although works well in DNN with large amount of data, lose their power as they limit the flexibility of fast adaptation in the inner-loop of meta-learning. The general-purpose regularization method, such as weight decay, is non-trivial for controlling model complexities of deep networks, since it would hurt the capability of memorizing not only noisy labels but also complex and useful data samples.

Table 1: Model performance with label noises

Methods	5-way 1-shot					5-way 5-shot				
	CUB	mini	mini-CUB	tiered	CF	CUB	mini	mini-CUB	tiered	CF
ProtoNet	42.57	39.95	36.52	31.94	55.46	65.98	63.80	57.92	41.42	71.67
Weight Decay	36.67	26.82	31.07	34.91	52.75	52.32	35.67	37.86	51.46	68.10
SPL	48.55	43.44	37.06	42.11	55.20	68.47	58.67	54.94	59.62	70.93
MixUp	39.22	21.55	21.15	40.46	46.86	48.75	24.88	25.08	40.46	61.28
CoTeaching	49.10	44.44	36.48	40.36	52.51	65.42	59.57	56.90	55.10	66.89
CoTeaching+	49.45	45.08	39.54	40.19	51.27	67.92	60.43	54.29	54.54	67.22
FSR	37.14	44.81	39.82	30.65	37.39	51.48	61.89	57.38	50.40	44.68
DIAL	**54.83**	**51.74**	**41.75**	**47.19**	**61.82**	**70.52**	**66.52**	**58.55**	**60.78**	**75.30**

4.2 Ablation Study

In this section, we conduct an extensive ablation study to investigate the impact of key factors on DIAL, including different noise ratios, choice of auxiliary data, values of hyperparameters for auxiliary tasks and regularization term, and an implementation on another classic episodic meta-learning method (*i.e.,* FOMAML (Nichol et al., 2018)). Additional ablations about episode sampling are presented in the Appendix due to space limit.

Impact of noise ratio. Fig. 1 shows the accuracy of different methods under noise ratios ranging from 10% to 50% on the CIFAR-FS dataset. We observed that our model outperforms all other methods under all noise ratios. Despite that all methods' performance drops as the noise ratio increase, our method decrease the slowest. For the 5-way 1-shot case, the performance decrease 15% and for the 5-way 5-shot classification task the performance decrease 17%. While for the data cleansing method SPL the performance drops 23% and 28%. Even under the extremely 50% noise ratio, our method still outperform the second best by 5% in the 5-way 1-shot classification task, showing that our method is particularly effective with limited training data. As for the 5-way 5-shot classification task, we only outperform the second best by nearly 2% since the data scarce situation is alleviate by extra shots.

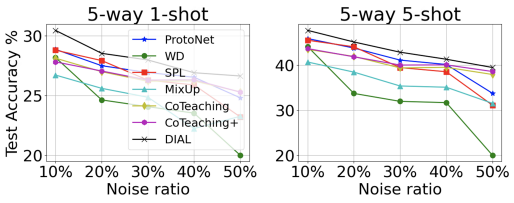

Figure 1: CF dataset with varied noise ratios.

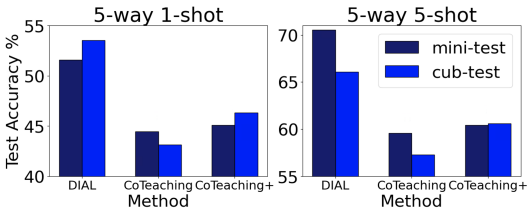

Figure 2: Effect of different auxiliary data.

Impact of auxiliary data. In this study, we test the impact of auxiliary data with the following two experiments: First, we test the influence of auxiliary data on different methods including Co-teaching, Co-teaching+ and DIAL. The experiment is conducted on miniImageNet with two different auxiliary datasets: CUB-test and mini-test. As shown in Fig. 2, for DIAL, two auxiliaries show different performance, with more data available (*i.e.,* 5w5s), the performance of using mini-test is better than using cub-test. This demonstrates the fact that extra data helps during training. While there are only 1 shot available, CUB-test provide a fine-grained classification task to help improve the performance. Our method outperforms other methods in general, indicating that learning a domain-invariant representation is necessary when utilizing extra data. Second, in Fig. 3, we train DIAL with to a various set of auxiliary datasets including CUB-train, CUB-test, CUB-val, and mini-test, mini-val to train on the miniImagenet dataset, and use CUB-test, CUB-val, and mini-train, mini-test, mini-val to train the model on the CUB dataset. We can draw similar conclusion as last experiment, that is, when the data is scarce, the fine-grained classification can help train models with better performance. In general, when choosing the auxiliary datasets, we should consider factors like class relatedness, data size, or types of the training tasks (such as fine-grained vs. coarse-grained), to make the most of the extra data.

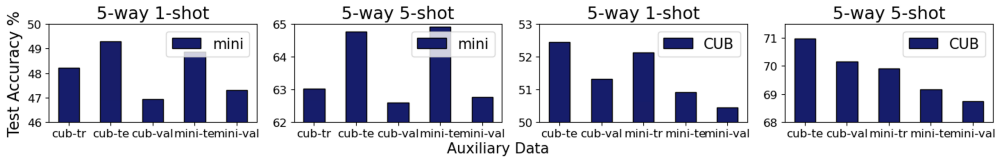

Figure 3: Impact of auxiliary data: 5-way 1-shot and 5-way 5-shot experiments on miniImageNet (left two columns) and CUB (right two columns) using different auxiliary datasets for DIAL.

Impact of hyperparameter α_1 and α_2. In this experiment, we test the effect of α_1 and α_2 in used in the objective Eq. (6) on CUB dataset. When we test the values of α_1, we fix $\alpha_2 = 0.2$. In Fig. 4 the value of α_1 is set among 0.1 to 0.9. The results show that using relatively balanced hyperparameters, specifically, setting α from 0.4 to 0.6, achieves higher performance, which means extra knowledge from the auxiliary task indeed helps the generalization of the model with proper combination with the primary data. Apparently, setting α to a very small value can not achieve desirable result since we focus on the primary task and making little use of auxiliary data. In in Fig. 5, when we test the impact of α_2, we keep α_1 fixed as 0.5 while setting the values of α_2 are set as 0.002, and from 0.1 to 0.6. The results show that a very small α_2 value doesn't work as well as the relatively larger ones considering the divergence between the primary and auxiliary distribution should be constrainted. As for other values, the performance of 5-way 1-shot task are all over 50%, showing a consistently high performance.

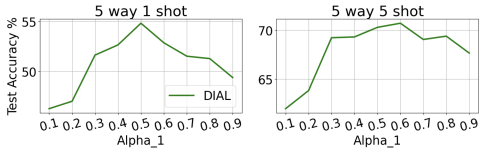

Figure 4: Impact of α_1 on CUB.

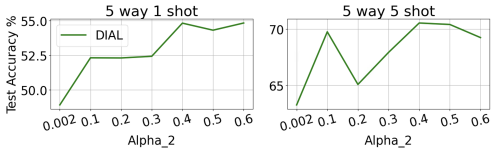

Figure 5: Impact of α_2 on CUB.

Generalization to other meta-learning models. The implementations of above experiments are based on one of the most classic meta-learning method ProtoNet (Snell et al., 2017), in which our model has shown superior performance. Moreover, our framework is model agnostic and it can be easily generalized to other types of meta-learning method. We provide a DIAL implementation based on another classic few-shot learning method FOMAML (Nichol et al., 2018) and test on different datasets with the same settings as mentioned before. The parameters of the base model and our method are set the same as (Finn et al., 2017). From Tab. 2's result, solid performance and advantage over the baseline are observed, which confirms the generalization to other types of meta-learning models.

Table 2: Accuracies (%) of implementations based on FOMAML

	CUB	mini	mini-CUB	tiered	CF
METHOD	5-way 1-shot				
FOMAML	45.11	33.94	31.93	25.95	48.49
DIAL	**45.41**	**39.61**	**35.09**	**34.89**	**50.95**
	5-way 5-shot				
FOMAML	51.06	37.52	37.15	26.11	58.53
DIAL	**52.93**	**47.01**	**43.31**	**44.51**	**62.52**

5 Conclusion

In this paper, we propose a novel meta-learning framework with auxiliary tasks that utilize extra unlabeled data for FSL under noisy settings. An episode sampling process is designed to remove noisy labels. The unsupervised auxiliary task is formed with no additional annotation cost. To better aligned the auxiliary tasks with the primary ones, we propose a regularization term based on the Wasserstein distance for learning a domain-invariant representation. We theoretically and experimentally demonstrate that incorporating auxiliary tasks can be beneficial. Our novel PAC-Bayesian bound also clearly demonstrate the negative impact of noisy labels. This work opens interesting directions for future research. First, when training deep neural networks with the auxiliaries, we observe better performance in terms of generalization bound, which can be further measured and quantified. Second, when the noise is considered in the theoretical analysis, a natural question is that what can be done to reduce the negative impact according to the theoretical insights.

REFERENCES

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pp. 205–214. PMLR, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *Proceedings of the International Conference on Learning Representations*, 2019.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 274–289. Springer, 2014.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

Tristan Deleu, Tobias Würfl, Mandana Samiei, Joseph Paul Cohen, and Yoshua Bengio. Torchmeta: A Meta-Learning library for PyTorch, 2019. URL https://arxiv.org/abs/1909.06576. Available at: https://github.com/tristandeleu/pytorch-meta.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1126–1135, 2017.

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1, 2016.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations*, 2018.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of the Conference on Neural Information Processing Systems*, 2018.

Chengcheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 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*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 1664–1673. Association for Computational Linguistics, 2021.

Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *Proceedings of the International Conference on Learning Representations*, 2016.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Siavash Khodadadeh, Ladislau Bölöni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *Proceedings of the Conference on Neural Information Processing Systems*, 2019.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Proceedings of the Conference on Neural Information Processing Systems*, volume 1, pp. 2, 2010.

Xinzhe Li, Qianru Sun, Yaoyao Liu, Shibao Zheng, Qin Zhou, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In *Proceedings of the Conference on Neural Information Processing Systems*, 2019.

Shikun Liu, Andrew J Davison, and Edward Johns. Self-supervised generalisation with meta auxiliary learning. In *Proceedings of the Conference on Neural Information Processing Systems*, 2019a.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *Proceedings of the International Conference on Learning Representations*, 2019b.

David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.

Carlos Medina, Arnout Devos, and Matthias Grossglauser. Self-Supervised Prototypical Transfer Learning for Few-Shot Classification. *arXiv preprint arXiv:2006.11325*, 2020.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Aviv Navon, Idan Achituve, Haggai Maron, Gal Chechik, and Ethan Fetaya. Auxiliary learning by implicit differentiation. In *Proceedings of the International Conference on Learning Representations*, 2021.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pp. 991–999. PMLR, 2014.

Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *Proceedings of the International Conference on Learning Representations*, 2021.

Limeng Qiao, Yemin Shi, Jia Li, Yaowei Wang, Tiejun Huang, and Yonghong Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3603–3612, 2019.

Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 737–753. Springer, 2017.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *Proceedings of the International Conference on Learning Representations*, 2018.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pp. 9116–9126. PMLR, 2021.

Doyen Sahoo, Hung Le, Chenghao Liu, and Steven CH Hoi. Meta-learning with domain adaptation for few-shot learning under domain shift. 2018.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 403–412, 2019.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Machine Learning*, 2017.

Vivek Veeriah, Matteo Hessel, Zhongwen Xu, Richard Lewis, Janarthanan Rajendran, Junhyuk Oh, Hado van Hasselt, David Silver, and Satinder Singh. Discovery of useful questions as auxiliary tasks. In *Advances in Neural Information Processing Systems*, 2019.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, et al. Improving generalization in meta-learning via task augmentation. In *International Conference on Machine Learning*, pp. 11887–11897. PMLR, 2021.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pp. 7164–7173. PMLR, 2019.

Zhongjie Yu, Lin Chen, Zhongwei Cheng, and Jiebo Luo. Transmatch: A transfer-learning scheme for semi-supervised few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12856–12864, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*, 2017.

Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 725–734, 2021.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10012–10022, 2020.

Appendix

In this appendix, we first provide the detail of the update of model θ^f and θ^d, and the optimization process in Algorithm 1. Then we describe the comparison methods we used in the experiment section, which covers different ranges in Appendix B. The detailed proof of Theorem 1 and Theorem 2 are provided in Appendix C and Appendix D. Additional ablation study about episode sampling is conducted in Appendix E. The link to the source code is provided in Appendix F.

A Optimization Process

The optimization of Eq. (6) is conducted through an alternative training between the optimization of the prediction model parameter θ^f and the critic model parameter θ^d while keeping the other fixed. We first represent the two term in Eq. (6) with their empirical counterparts defined as follows according to the observed data during training:

$$\hat{\mathcal{R}}(\theta^f, \mathcal{D}^{pri}, \mathcal{D}^{aux}) = \frac{\alpha_1}{|\mathcal{D}^{pri}|} \sum_{(x,y)\sim\mathcal{D}^{pri}} \ell(h(x,y)) + \frac{1-\alpha_1}{|\mathcal{D}^{aux}|} \sum_{(x,y)\sim\mathcal{D}^{aux}} \ell(h(x,y)) \tag{14}$$

$$\hat{\mathcal{E}}(\theta^f, \theta^d) = \frac{1}{|\mathcal{D}^{pri}|} \sum_{x\sim\mathcal{D}^{pri}} [g(x)] - \frac{1}{|\mathcal{D}^{aux}|} \sum_{x\sim\mathcal{D}^{aux}} [g(x)] \tag{15}$$

Then the parameters of θ^f and θ^d are updated alternatively as follows:

$$\begin{aligned}\theta^f \leftarrow & \theta^f - \eta^f \left(\frac{\alpha_1}{|\mathcal{D}^{pri}|} \sum_{(x,y)\sim\mathcal{D}^{pri}} \frac{\partial \ell(h(x,y))}{\partial \theta^f} + \frac{1-\alpha_1}{|\mathcal{D}^{pri}|} \sum_{(x,y)\sim\mathcal{D}^{aux}} \frac{\partial \ell(h(x,y))}{\partial \theta^f} \right) \\ & + \left(\frac{\alpha_2}{|\mathcal{D}^{pri}|} \sum_{x\sim\mathcal{D}^{pri}} \frac{\partial g(x)}{\partial \theta^f} - \frac{\alpha_2}{|\mathcal{D}^{aux}|} \sum_{x\sim\mathcal{D}^{aux}} \frac{\partial g(x)}{\partial \theta^f} \right)\end{aligned} \tag{16}$$

$$\theta^d \leftarrow \theta^d + \eta^d \left(\frac{\alpha_2}{|\mathcal{D}^{pri}|} \sum_{x\sim\mathcal{D}^{pri}} \frac{\partial g(x)}{\partial \theta^d} - \frac{\alpha_2}{|\mathcal{D}^{aux}|} \sum_{x\sim\mathcal{D}^{aux}} \frac{\partial g(x)}{\partial \theta^d} \right) \tag{17}$$

where η^f, η^d are the learning rates. The detailed optimization process is shown in Algorithm 1.

Algorithm 1 DIAL Training

INPUT initialized meta-model θ, α_1,α_2, task distributions τ, τ', w
while not done **do**
  Construct a batch of labelled tasks $\{\mathcal{T}_i\} \sim \tau$
  Construct a batch of unlabelled tasks $\{\mathcal{T}_i'\} \sim \tau'$
  for a mini-batch of tasks $\mathcal{T}_i$, $\mathcal{T}_i'$ **do**
    Update primary task-specific model θ_i using data $\mathcal{S}_i^{sup}$
    Update auxiliary task-specific model θ_i' using data $\mathcal{S}_i'^{sup}$
    Use θ_i and θ_i' for episode sampling query data from $\mathcal{S}_i^{que}$ and $\mathcal{S}_i'^{que}$, and obtain the clean query set $\hat{\mathcal{S}}_i^{que}$ and $\hat{\mathcal{S}}_i'^{que}$
  end for
  Fix w and update model parameter θ^f using Eq. (16)
  Fix w and update critic parameter θ^d using Eq. (17)
  Fix θ^f and θ^d and update sample weight w
end while

B Discussion of Comparison Methods

In this work, we compare DIAL with the following baselines: ProtoNet (Snell et al., 2017), (*i.e.,* training only using primary tasks), Weight Decay (Krogh & Hertz, 1991), SPL (Kumar et al., 2010),

MixUp (Zhang et al., 2017), CoTeaching (Han et al., 2018), CoTeaching+ (Yu et al., 2019) and FSR-raw (Zhang & Pfister, 2021). Weight decay and MixUp are typical noisy label methods. SPL, CoTeaching, CoTeaching+ and FSR are methods designed according to data cleansing that focus on removing the potential noisy data. Similar to our method, SPL, CoTeaching and CoTeaching+ use the loss of the example for selecting noisy data. CoTeaching and CoTeaching+ train peer networks that are trained simultaneously and teach each other. Traditional CoTeaching and CoTeaching+ train the peer networks using the same dataset. In this work, for fair comparison, the peer networks are trained on the primary and auxiliary data, respectively. Therefore, extra data are utilized in CoTeaching and CoTeaching+.

C Proof of Theorem 1

In this proof, we first provide the following multi-source domain adaptation bound in Theorem 3, the PAC-Bayesian generalization bound for single task in Theorem 4, and two useful lemmas Lemma 5 and Lemma 6. Then we deduce the detailed proof for Theorem 1 using these results.

We first introduce the multi-source domain adaptation results in Redko et al. (2017). Consider N different source domains, and for each domain a labelled sample set S_j with n_j examples are drawn from the associated unknown distribution μ_{S_j} and labelled by f_j, the empirical weighted multi-source error of a hypothesis h defined for some vector $\alpha = \{\alpha_1, ..., \alpha_N\}$ as follows:

$$\hat{\epsilon}_\alpha(h) = \sum_{j=1}^{N} \alpha_j \hat{\epsilon}_{S_j}(h) \tag{18}$$

where $\sum_{j=1}^{N} \alpha_j = 1$ and each α_j represents the weights of the source domain S_j. $n_j = \beta_j n$, $\sum_{j=1}^{N} n_j = n$. The empirical source distribution is defined as $\hat{\mu}_{S_j} = \frac{1}{N_{S_j}} \sum_{i=1}^{N_{S_j}} \delta_{x^i_{S_j}}$, which is a uniformly weighted sum of N_{S_j} Diracs with mass at locations $x^i_{S_j}$. The empirical target distribution is defined as $\hat{\mu}_T = \frac{1}{N_T} \sum_{i=1}^{N_T} \delta_{x^i_T}$ accordingly. Denote the Wasserstein distance between source domain μ_{S_j} and μ_T as $W_1(\mu_{S_j}, \mu_T)$ and apply it in this multi-source domain adaptation problem, the weighted multi-source error is bounded by the following Theorem 3:

Theorem 3 *(Redko et al., 2017) Assume that the cost function $d(\mathbf{x}, \mathbf{y}) = ||\phi(\mathbf{x}) - \phi(\mathbf{y})||_{\mathcal{H}_{k_l}}$, where $\mathcal{H}_{k_l}$ is a Reproducing Kernel Hilbert Space (RHKS) equipped with kernel k_l and $0 \leq k_l \leq R$. Denote $\hat{h}_\alpha$ as the empirical minimizer of $\hat{\epsilon}_\alpha(h)$ and $h^*_T = \min_h \epsilon_T(h)$ then for any fixed α and $\delta \in (0, 1)$ with probability $1 - \delta$,*

$$\epsilon_T(\hat{h}_\alpha) \leq \epsilon_T(h^*_T) + c_1 + 2 \sum_{j=1}^{N} \alpha_j (W_1(\hat{\mu}_{S_j}, \hat{\mu}_T) + \lambda_j + c_2) \tag{19}$$

where

$$c_1 = 2\sqrt{\frac{2K \sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j} \log(2/\delta)}{n}} + 2\sqrt{\sum_{j=1}^{R} \frac{R\alpha_j}{\beta_j n}} \tag{20}$$

$$c_2 = \sqrt{2\log(\frac{1}{\delta})/\varsigma'}(\sqrt{\frac{1}{N_{S_j}}} + \sqrt{\frac{1}{N_T}}) \tag{21}$$

where $\lambda_j = \min_h(\epsilon_{S_j}(h) + \epsilon_T(h))$ represents the joint error for each source domain j.

The following Theorem 4 is the PAC-Bayesian results for a single task.

Theorem 4 *(Alquier et al., 2016; Rothfuss et al., 2021) Given a data distribution $\mathcal{D}$, hypothesis spaces $\mathcal{H}, \mathcal{F}$, a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, a prior distribution $\pi \in \mathcal{M}(\mathcal{F})$, a confidence level $\delta \in (0, 1]$ and a real number $\eta > 0$, with probability at least $1 - \delta$ over samples $S \sim \mathcal{D}^m$, we have $\forall \rho \in \mathcal{M}(\mathcal{H})$:*

$$\mathcal{L}(\rho, \mathcal{D}) \leq \hat{\mathcal{L}}(\rho, S) + \frac{1}{\eta}\left[D_{KL}(Q||P) + \log\frac{1}{\delta} + \Psi(\eta, m)\right]$$

where $\Psi(\eta, m) = \log \mathbb{E}_{h\sim\pi}\mathbb{E}_{S\in\mathcal{D}^m} \exp\left[\eta(\mathcal{L}(h, \mathcal{D}) - \hat{\mathcal{L}}(h, S))\right]$.

where η is a hyper-parameter with common choice of $\eta = m$. And according to (Rothfuss et al., 2021), $\Psi(\eta, m)$ can be bounded as follows by making additional assumption on the loss function ℓ.

Lemma 5 *(Pentina & Lampert, 2014) For any fixed algorithm A and any λ the following holds:*

$$\mathbb{E}_{E_1,\ldots,E_n} \exp\left(\lambda(b - \frac{1}{n}\sum_{i=1}^{n} g(X_i))\right) \leq \exp(\frac{\lambda^2}{2n})$$

where $b = er(A)$ and $g : X_i \to er_i(A)$ with $X_i = (E_{i-1}, E_i)$

Lemma 6 *(Amit & Meir, 2018) (Union bound) Let $\{E_i\}_{i=1}^n$ be a set of events, which satisfies $Pr(E_i) \geq 1 - \delta_i$, with some $\delta_i \geq 0$, $i = 1, ..., n$. Then $\bigcap_{i=1}^n Pr(E_i) \geq 1 - \sum_{i=1}^n \delta_i$.*

The detailed proof of Theorem 1 is provided as follows, which contains three steps:

Proof: Step 1: Task-specific generalization. In this step, we can directly apply Theorem 3 for a single task $\mathcal{T}_i$ and obtain with probability $1 - \delta_i$:

$$\mathcal{L}(\mathcal{Q}, \mathcal{D}_i^{pri}, \mathcal{D}_i^{aux}) \leq \mathbb{E}_{P\sim\mathcal{Q}}\mathbb{E}_{h\sim Q}[\hat{\mathcal{L}}(h, S_i)] + c_1 + 2\sum_{j=1}^{2}\alpha_j(\mathcal{W}_1(\hat{\mathcal{D}}^{pri}, \hat{\mathcal{D}}^{aux}) + \lambda_{pri} + c_2) \quad (22)$$

where c_1, c_2 are defined in Eq. (20) and Eq. (21) with $\delta = \delta_i$. And $\lambda_{pri} = \min_h(\mathcal{L}(h, \mathcal{D}^{pri}) + \mathcal{L}(h, \mathcal{D}^{aux}))$.

Step 2: Task environment generalization. Define the expected multi-task error as

$$\tilde{\mathcal{L}}(\mathcal{Q}, \tau, \tau') = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathcal{Q}, \mathcal{D}_i^{pri}, \mathcal{D}_i^{aux}) \quad (23)$$

By applying Donsker-Varahan's variational fomula (Donsker & Varadhan, 1975), we have ,

$$\mathcal{L}(\mathcal{Q}, \tau, \tau') - \tilde{\mathcal{L}}(\mathcal{Q}, \tau, \tau') \leq \frac{1}{\lambda}D_{KL}(\mathcal{Q}||\mathcal{P}) + \frac{1}{\lambda}\left(\log \mathbb{E}_{h\sim\mathcal{P}} \exp\lambda(\mathcal{L}(\mathcal{Q}, \tau, \tau') - \tilde{\mathcal{L}}(\mathcal{Q}, \tau, \tau'))\right) \quad (24)$$

Using Lemma 5 and Markov's inequality, we have with probability $1 - \delta_0$

$$\mathbb{E}_{h\sim\mathcal{P}} \exp\lambda(\mathcal{L}(\mathcal{Q}, \tau, \tau') - \tilde{\mathcal{L}}(\mathcal{Q}, \tau, \tau')) \leq \frac{1}{\delta_0}\exp(\frac{\lambda^2}{4n}) \quad (25)$$

Therefore,

$$\mathcal{L}(\mathcal{Q}, \tau, \tau') \leq \tilde{\mathcal{L}}(\mathcal{Q}, \tau, \tau') + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta_0} + \log\frac{\lambda^2}{4n}\right) \quad (26)$$

Step 3: Union bound. To combine the results from the first two steps, Eq. (22), Eq. (23), and Eq. (26), we bound the intersection of the events in them using the union argument in Lemma 6 by setting $\delta_i = \frac{\delta}{4n}$, and $\delta_0 = \frac{\delta}{2}$, we have with probability $1 - \delta$,

$$\begin{aligned}
&\mathcal{L}(\mathcal{Q}, \tau, \tau') \leq \tilde{\mathcal{L}}(\mathcal{Q}, \tau, \tau') + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \log\frac{\lambda^2}{4n}\right) \\
=&\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathcal{Q}, \mathcal{D}_i^{pri}, \mathcal{D}_i^{aux}) + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \log\frac{\lambda^2}{4n}\right) \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{P\sim\mathcal{Q}}\mathbb{E}_{h\sim Q}[\hat{\mathcal{L}}(h, S_i)] + c_1 + 2\sum_{j=1}^{2}\alpha_j(\mathcal{W}_1(\hat{\mathcal{D}}^{pri}, \hat{\mathcal{D}}^{aux}) + \lambda_{pri} + c_2)\right) \\
+&\frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \log\frac{\lambda^2}{4n}\right)
\end{aligned}$$

■

D PROOF OF THEOREM 2

There are three steps to prove Theorem 2:

Proof: Step 1: Task-specific generalization. The same rule is applied to task-specific generalization bound for a single task as in step 1 in the proof of Theorem 1.

Step 2: Task environment generalization. Second, we bound the task-level generalization by relating the transfer error $\mathcal{L}(\mathcal{Q},\tau,\tau')$ to expected multi-task error of primary tasks $\tilde{\mathcal{L}}(\mathcal{Q},\mathcal{D}_i)$, and that of auxiliary tasks $\tilde{\mathcal{L}}'(\mathcal{Q},\mathcal{D}_i')$. Rewrite the meta-training error of a given prior P on observed tasks $\mathcal{D}_i\sim\tau$ and $\mathcal{D}_i'\sim\tau'$ as follows:

$$\begin{aligned}\mathcal{L}_{S,S'}(P) = &\quad \alpha_1\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathcal{L}(Q(S_i,P),\mathcal{D}_i) + (1-\alpha_1)\tfrac{1}{n}\sum_{i=1}^n \mathcal{L}(Q(S_i',P),\mathcal{D}_i') \\ = &\quad \tfrac{\alpha_1}{n}\textstyle\sum_{i=1}^n \mathbb{E}_{z_i\sim\mathcal{D}_i}\mathbb{E}_{h\sim Q}\ell(h_i,z_i) + \tfrac{1-\alpha_1}{n}\sum_{i=1}^n \mathbb{E}_{z_i\sim\mathcal{D}_i'}\mathbb{E}_{h\sim Q}\ell(h_i,z_i)\end{aligned} \tag{27}$$

Similarly, the generalization error on the target task $\mathcal{T}$ is

$$\mathcal{L}(P,\tau) = \mathbb{E}_{(\mathcal{D},m)\sim\tau}\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{z\in\mathcal{D}}\mathbb{E}_{h\sim Q(h|P,S)}\ell(h,z)$$

Using Markov's Inequality, with probability at least $1-\delta_0$:

$$\mathbb{E}_{P\sim\mathcal{P}}\left[e^{\lambda(\mathcal{L}(P,\tau)-\mathcal{L}_{S,S'}(P)}\right] \le \frac{1}{\delta_0}\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}^{i=1,\ldots,2n}_{\mathcal{D}_i\sim\tau,S_i\sim\mathcal{D}_i^{m_i},\mathcal{D}_i\sim\tau',S_i'\sim\mathcal{D}_i'^{m_i}}\left[e^{\lambda(\mathcal{L}(\mathcal{P},\tau)-\mathcal{L}_{S,S'}(P)}\right] \tag{28}$$

The log of the left hand side can be lower bounded by:

$$\begin{aligned}\log\mathbb{E}_{P\sim\mathcal{P}}\left[e^{\lambda(\mathcal{L}(P,\tau)-\mathcal{L}_{S,S'}(P)}\right] &= \log\mathbb{E}_{P\sim\mathcal{Q}}\frac{\mathcal{P}(P)}{\mathcal{Q}(P)}\left[e^{\lambda(\mathcal{L}(P,\tau)-\mathcal{L}_{S,S'}(P)}\right] \\ &\le \mathbb{E}_{P\sim\mathcal{Q}}\log\frac{\mathcal{P}(P)}{\mathcal{Q}(P)} + \lambda\mathbb{E}_{P\sim\mathcal{Q}}\left[\mathcal{L}(P,\tau)-\mathcal{L}_{S,S'}(P)\right] \\ &= -D_{KL}(\mathcal{Q}||\mathcal{P}) + \lambda(\mathcal{L}(\mathcal{Q},\tau)-\mathbb{E}_{P\sim\mathcal{Q}}\mathcal{L}_{S,S'}(P))\end{aligned} \tag{29}$$

where we have $\mathbb{E}_{P\sim\mathcal{Q}}\mathcal{L}(P,\tau)=\mathcal{L}(\mathcal{Q},\tau)$. And the log of the right hand side can be upper bounded by:

$$\begin{aligned}&\log\frac{1}{\delta_0}\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}^{i=1,\ldots,2n}_{\mathcal{D}_i\sim\tau,S_i\sim\mathcal{D}_i^{m_i},\mathcal{D}_i\sim\tau',S_i'\sim\mathcal{D}_i'^{m_i}}\left[e^{\lambda(\mathcal{L}(\mathcal{P},\tau)-\mathcal{L}_{S,S'}(P)}\right] \\ =&\log\frac{1}{\delta_0} + \log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}^{i=1,\ldots,2n}_{\mathcal{D}_i\sim\tau,S_i\sim\mathcal{D}_i^{m_i},\mathcal{D}_i\sim\tau',S_i'\sim\mathcal{D}_i'^{m_i}}\left[e^{\lambda(\mathcal{L}(\mathcal{P},\tau)-\mathcal{L}_{S,S'}(P)}\right] \\ =&\log\frac{1}{\delta_0} + \log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}_{S\sim\tau,S'\sim\tau'}\left[e^{\lambda(\mathcal{L}(\mathcal{P},\tau)-\mathcal{L}_{S,S'}(P)}\right] \\ =&\log\frac{1}{\delta_0} + \log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}_{S\sim\tau,S'\sim\tau'}\left[e^{\lambda(\mathcal{L}(\mathcal{P},\tau)-\mathbb{E}_{S\sim\tau,S'\sim\tau'}\mathcal{L}_{S,S'}(P)}\right] \\ &+\log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}_{S\sim\tau,S'\sim\tau'}\left[e^{\lambda(\mathbb{E}_{S\sim\tau,S'\sim\tau'}\mathcal{L}_{S,S'}(P)-\mathcal{L}_{S,S'}(P)}\right] \\ \le&\log\frac{1}{\delta_0} + \log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}_{S\sim\tau,S'\sim\tau'}\left[e^{\lambda(\mathbb{E}_{S\sim\tau,S'\sim\tau'}\mathcal{L}(P,\tau)-\mathcal{L}(P,\tau,\tau')}\right] + \Psi(\lambda,2n)\end{aligned} \tag{30}$$

where

$$\begin{aligned}&\mathbb{E}_{S\sim\tau,S'\sim\tau'}\mathcal{L}_{S,S'}(P) = \mathbb{E}_{\mathcal{D}\sim\tau,\mathcal{D}'\sim\tau'}\mathbb{E}_{S\sim\mathcal{D},S'\sim\mathcal{D}'}\mathcal{L}_{S,S'}(P) \\ =&\mathbb{E}_{(\mathcal{D},m)\sim\tau}\mathbb{E}_{S\sim\mathcal{D}^m}\mathbb{E}_{(\mathcal{D}',m)\sim\tau'}\mathbb{E}_{S'\sim\mathcal{D}'^m}\mathbb{E}_{z\in\mathcal{D}}\mathbb{E}_{h\sim Q(h|P,S)}\ell(h,z) \\ =&\mathcal{L}(P,\tau,\tau')\end{aligned} \tag{31}$$

Combining (29) and (30), we have

$$\mathcal{L}(\mathcal{Q},\tau) \le \mathbb{E}_{P\sim\mathcal{Q}}\mathcal{L}_{S,S'}(P) + \Delta_\lambda(\mathcal{P},\tau,\tau') + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{1}{\delta_0} + \Psi(\lambda,2n)\right) \tag{32}$$

where

$$\Delta_\lambda(\mathcal{P},\tau,\tau') = \frac{1}{\lambda}\log\mathbb{E}_{P\sim\mathcal{P}}\mathbb{E}_{S\sim\tau,S'\sim\tau'}\left[e^{\lambda(\mathbb{E}_{S\sim\tau,S'\sim\tau'}\mathcal{L}(P,\tau)-\mathcal{L}(P,\tau,\tau')}\right]$$

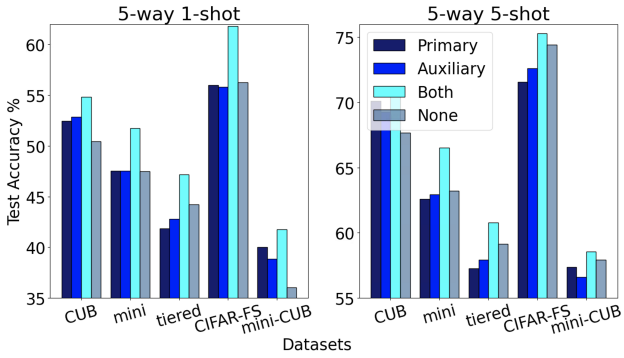

Figure 6: Data cleansing effect on different tasks.

Step 3: Union bound. With the results of step 1&2, by applying the union bound in Lemma 6, and setting $\delta_i = \frac{\delta}{2n}$, $\delta_0 = \frac{\delta}{2}$, we have for any $\delta > 0$,

$$\begin{aligned}
\mathcal{L}(\mathcal{Q}, \tau) \leq & \mathbb{E}_{P\sim\mathcal{Q}} \mathcal{L}_{S,S'}(P) + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \Psi(\lambda, 2n)\right) + \Delta_\lambda(\mathcal{P}, \tau, \tau') \\
= & \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\mathcal{Q}, \mathcal{D}_i^{pri}, \mathcal{D}_i^{aux}) + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \Psi(\lambda, 2n)\right) + \Delta_\lambda(\mathcal{P}, \tau, \tau') \\
\leq & \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P\sim\mathcal{Q}}\mathbb{E}_{h\sim Q}[\hat{\mathcal{L}}(h, S_i)] + c_1 + 2\sum_{j=1}^{2}\alpha_j(\mathcal{W}_1(\hat{\mathcal{D}}^{pri}, \hat{\mathcal{D}}^{aux}) + \lambda_{pri} + c_2) \\
& + \frac{1}{\lambda}\left(D_{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{2}{\delta} + \Psi(\lambda, 2n)\right)
\end{aligned} \tag{33}$$

E Additional Ablation Study

Impact of episode sampling. In Fig. 6, we show the impact of sampling in the DIAL framework by applying it to different tasks, including primary, auxiliary, primary+auxiliary (Both), and vanilla (None). In most cases, applying sampling to both tasks achieves the best performance while vanilla DIAL is the worst, which confirms the effectiveness of our proposed sampling method. This shows that, with additional auxiliary tasks introduced, the potential issue caused by using data cleansing for noisy labels no longer exists.

Impact of threshold γ. In this experiment, we test the effect of the hyperparameter threhold γ in the weight assignment rule on CUB and miniImageNet dataset with synthetic noise, respectively. For simplicity, we set the value of γ by the percentage of noisy examples removed and its value range from 0% to 80%, which is a common practice of self-paced learning approaches. As shown in Fig. 7, for CUB dataset, $\gamma = 30\%$ achieves the highest performance for 5-way 1-shot experiment, and $\gamma = 60\%$ achieves the highest performance for the 5-shot one. For miniImageNet dataset, $\lambda = 10\%$ achieves the highest performance for both 5-way 1-shot and 5-way 5-shot cases. The general trend shows that removing too many examples harms the model's performance, which confirms the hypothesis of information loss over self-paced sampling.

F Source Code

For source code, click here.

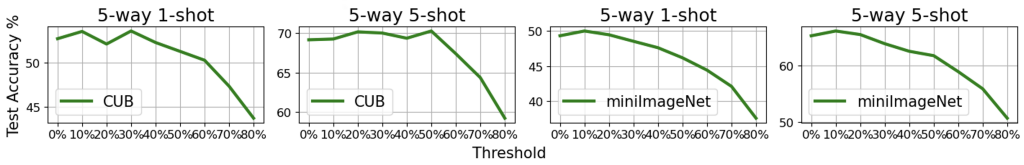

Figure 7: Impact of threshold γ: 5-way 1-shot and 5-way 5-shot experiments on miniImageNet (left two columns) and CUB (right two columns) using different threshold values γ.