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

Xiaofan Que and Qi Yu
Rochester Institute of Technology

September 9, 2024

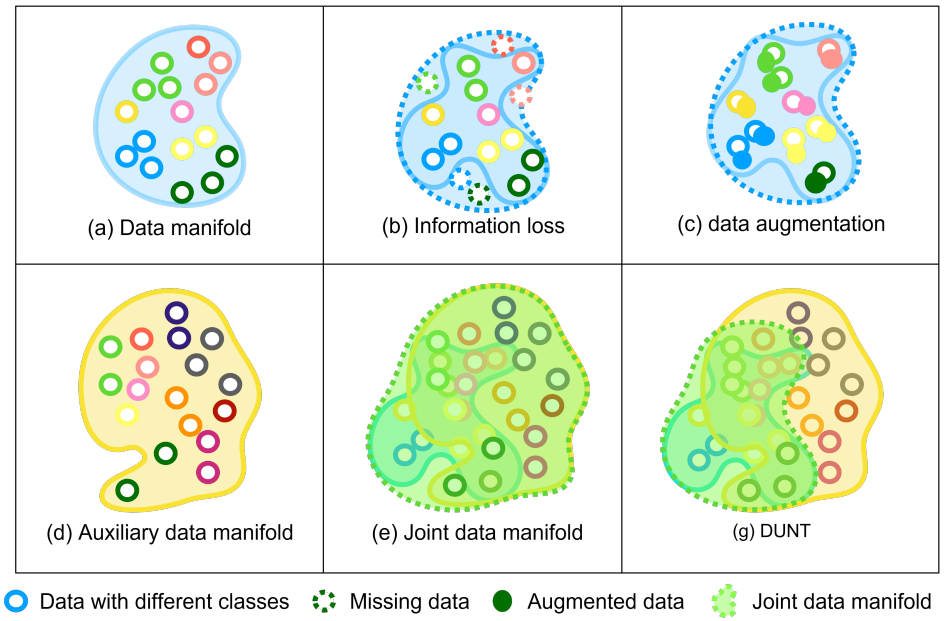
Outlines

1. Motivation and Introduction
   - Key challenges
   - Summary of Contributions
2. Methodology: DUNT
   - Learning from Primary Tasks
   - Learning from the Diverse Unsupervised Tasks
   - Adversarial Training and the Critic Head
3. Experiments
   - Main Results
   - Ablation Studies
4. Conclusion

Motivation and Introduction

- The **overall goal** of this work is to address the noisy few-shot learning (NFSL) problems with uniquely designed unsupervised auxiliary tasks to compensate for information loss.
- Key challenges include:
  - Few-shot learning inherently involves a small number of labeled examples per class. The scarcity of labeled data makes it challenging for models to generalize well to unseen instances.
  - The presence of noisy labels, where the provided annotations may be incorrect or unreliable, can significantly impact model performance. Models need to be robust to label noise to effectively learn from the limited labeled examples available.
  - 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.

Motivation


(a) Data manifold
(b) Information loss
(c) data augmentation
(d) Auxiliary data manifold
(e) Joint data manifold
(g) DUNT
Data with different classes
Missing data
Augmented data
Joint data manifold

Contributions

- Our proposed framework **D**iverse **UN**supervised **T**asks for NFSL (DUNT) includes threefold contributions:
  - 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.
  - a novel strategy to perform diverse auxiliary task selection to avoid learning sample-specific spurious features from the unlabeled auxiliary data samples.
  - 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.

Framework Overview

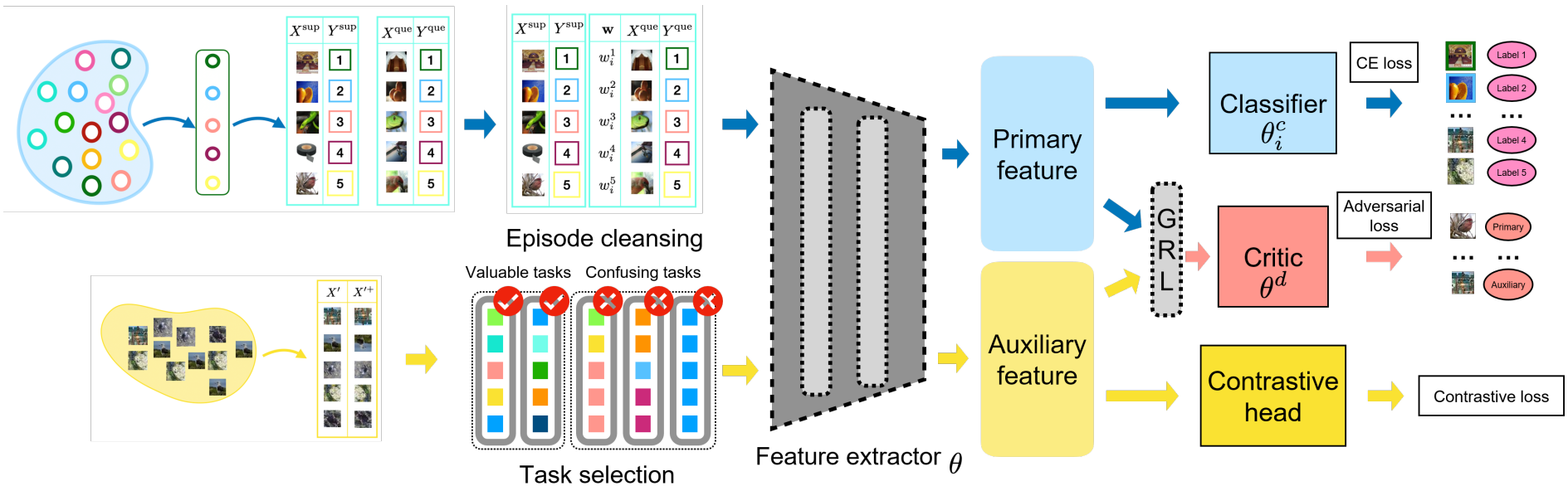


Figure: DUNT overview: The primary tasks (top) are sampled and cleansed by episode cleansing, the auxiliary tasks (bottom) are constructed and filtered by their diverseness.

Learning from Primary Tasks: Episode cleansing

- To remove the noisy labels in the query sets, episode cleansing designs a self-paced sample weight for each example pair:

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

  - $\ell_i^k = \ell(f_{\theta_i'}(\mathbf{x}_i^k), \mathbf{y}_i^k)$ is the loss value of k-th example pair (x_i^k, y_i^k) in task $\mathcal{T}_i$
  - $1(\cdot)$ is an indicator function whose value is 1 when $\ell_i^k < \gamma_{th}$ and 0 otherwise
  - γ_{th} is a predefined hyperparameter used to filter out high-loss examples
- The classification loss $\mathcal{L}_{\mathcal{T}_i}(\theta_i', \mathcal{S}_i^{\text{que}})$ becomes

$$\mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}, \hat{\mathcal{S}}_i^{\text{que}}) = \sum_{\mathbf{x}_i^k, \mathbf{y}_i^k \in \mathcal{S}_i^{\text{que}}} w_i^k \ell_i^k, \tag{1}$$

- For support set, meta-model is used to compute the loss values and remove the noise in the support set.

Learning from Diverse Unsupervised Tasks

- **Construction** of an unsupervised N-way 1-shot auxiliary task $\mathcal{T}_j$:
  - first randomly sampling N images from the dataset $\mathcal{D}^{\text{aux}}$ for the support set $\mathcal{S}_j^{\text{sup}} = \{\mathbf{x}_j^k\}_{k=1}^N$;
  - 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 $A_1(\cdot), .., A_Q(\cdot)$, A denote 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:

$$\mathcal{L}^{\text{aux}} = \sum_{\mathcal{T}_j(\mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}) \sim \mathcal{D}^{\text{aux}}} \mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}), \tag{2}$$

$$\mathcal{L}_{\mathcal{T}_j}(\theta, \mathcal{S}_j^{\text{sup}}, \mathcal{S}_j^{\text{que}}) = -\frac{1}{NQ} \sum_{k=1}^{N} \sum_{q=1}^{Q} \log \frac{\exp(-d[f_\theta(A_q(\mathbf{x}_j^k)), f_\theta(\mathbf{x}_j^k)])}{\sum_{l=1}^{N} \exp(-d[f_\theta(A_q(\mathbf{x}_j^k)), f_\theta(\mathbf{x}_j^l)])} \tag{3}$$

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

Diversity Criterion

- 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 each other.
- Diversity criterion:

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

  - $C(N,2) = \frac{N!}{(N-2)!2!}$ is the 2-combination of set N.
  - According to the definition, *the higher the value of div_j, the more diverse the task*.

Adversarial Training and the Critic Head

- 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. Formally, the overall adversarial objective is a min-max problem as follows:

$$\min_{\theta,\theta^c} \max_{\theta^d} \alpha_1 \mathcal{L}^{\text{pri}} + (1-\alpha_1)\mathcal{L}^{\text{aux}} + \alpha_2 \mathcal{E}(\theta, \theta^d, \tilde{\mathcal{D}}^{\text{pri}}, \mathcal{D}^{\text{aux}}), \quad (4)$$

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.

Results on Non-Transductive Setting

- Our method outperforms the baselines across different few-shot benchmark datasets.

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

Results on Transductive Setting

- Our method outperforms the baselines across different few-shot benchmark datasets.

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

Effectiveness of task selection.

- We study the impact of task selection through the threshold γ_{div}.

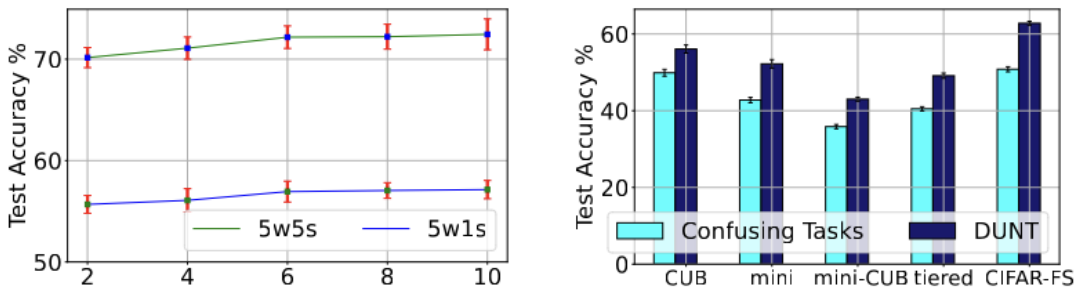


Fig. 3: `DUNT` on CUB with different γ_{div} (a); `DUNT` with different task selection (b)

Conclusion

- In this work, We propose a novel FSL framework with auxiliary tasks that utilize carefully selected unlabeled data under noisy settings.
- In DUNT, we introduce episode cleansing for examples in the primary task and adopt a diverse task selection strategy for the unsupervised auxiliary tasks to enhance robustness against label noise. 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.
- Our framework is theoretically and experimentally proved to be effective and beneficial.