

Supplementary Material: Unsupervised Domain Adaptation of Black-Box Source Models

BMVC 2021 Submission # 404

A Analyses on the estimation of ε

In order to investigate the influence of estimation precision of noise rate ε on the final performance, we do some experiments that set ε as a fixed value during different iterative steps, and the result is showed on Figure 5(b) and 5(c). From Figure 5(a) we know the ground truth noise rate in VisDA-2017 task is 0.485, and Figure 5(b) shows a fixed ε of value 0.3, 0.4, 0.5 can obtain similar good results, which means IterLNL is robust under the situation that estimated noise rate is not fairly precised. When the estimation error go larger, performance goes down. Using estimated ε , IterLNL achieve the same results as the best result of setting fixed ε on VisDA-2017 task (see Figure 5(b)) and better results on M→U task (see Figure 5(c)), showing that our noise rate estimation method is reliable.

B More analyses on balancing different categories

We propose the category-wise sampling to tackle the unbalanced label noise, i.e., promoting the prediction balance among different categories, and verify its efficacy in Table 3. We also note that existing methods [14, 15] (cf. Equation (3) in [14] and the appendices in [15]) typically promote the prediction balance among categories with a global diversity loss, which aims to assign target samples to each class equally:

$$\mathcal{L}_{gd}(F, \mathcal{T}) = \sum_{k=1}^K \bar{p}_k \log \bar{p}_k, \quad (1)$$

where \bar{p}_k is k -th element of \bar{p} , $\bar{p} = \mathbb{E}[F(\mathbf{x}_i^t)]$, and \mathbf{x}_i^t indicates the selected sample in each training batch. Minimizing \mathcal{L}_{gd} leads to more balanced model predictions among categories.

As illustrated in Table B1, both category-wise sampling and the global diversity loss alleviate the problem of unbalanced category predictions and the proposed category-wise sampling largely outperforms the global diversity loss in terms of category-wise mean accuracy (e.g., from 79.0% to 83.1%). We also find that combining the category-wise sampling and the global diversity loss results in boosting results, showing the effectiveness of our proposed category-wise sampling strategy.

Methods	plane	bicycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	cate. avg	ins. avg
IterLNL (w/o CateS)	95.6	87.2	86.0	90.0	96.9	0.0	93.7	51.2	93.3	88.1	87.4	0.0	72.4	75.6
IterLNL (+ \mathcal{L}_{gd} , w/o CateS)	96.2	86.6	81.8	76.8	95.6	0.0	92.0	87.2	93.6	87.9	90.0	59.9	79.0	81.3
IterLNL	88.7	83.4	78.3	67.7	91.4	87.6	91.8	79.5	86.2	86.7	78.7	77.2	83.1	81.2
IterLNL (+ \mathcal{L}_{gd})	90.0	87.2	77.8	63.3	93.1	90.6	90.8	87.1	88.7	90.6	81.1	76.7	84.8	81.8

Table B1: Ablation study on VisDA-2017 dataset (ResNet-101).

C Comparison with Self-training

As we discussed in Section 3.3, self-training is proposed for tasks with labeled and unlabeled training data and could not be directly applied to the B²UDA task, where only unlabeled (or noisy labeled) data are available. However, given the seemingly similar solutions between our IterLNL and the popular self-training, we also try our best to adapt the self-training methods for the B²UDA task. Specifically, given unlabeled target samples and their noisy labels, we first warm up the target model with supervised classification loss by using the noisy labels as clean ones; then we start self-training based on the warmed-up model by achieving noisy labels from model predictions of high confidence. As illustrated in Table C1, our IterLNL significantly outperforms the self-training on tasks of S→M and A→W, justifying the efficacy of our proposed IterLNL. Note that our IterLNL is significantly different from the self-training: samples with high prediction confidence are directly used for model training in self-training while LNL methods and our IterLNL use samples to learn models only if their current model predictions are consistent with the given noisy labels.

Methods	S→M	A→W
Self-training	93.6±0.3	84.0±0.5
IterLNL	97.7±0.1	92.2±0.0

Table C1: Comparison of IterLNL and self-training.

References

[1] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. *arXiv preprint arXiv:2002.08546*, 2020.

[2] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.