ADU: Adaptive Detection of Unknown Categories in Black-Box Domain Adaptation

Yushan Lai, Guowen Li, Haoyuan Liang, Juepeng Zheng*, and Zhiyu Ye
School of Artificial Intelligence, Sun Yat-Sen University
{laiysh6, ligw8, lianghy68, yezhy26}@mail2.sysu.edu.cn, zhengjp8@mail.sysu.edu.cn

Abstract

*Black-box Domain Adaptation (BDA) utilizes a black-box predictor of the source domain to label target domain data, addressing privacy concerns in Unsupervised Domain Adaptation (UDA). However, BDA assumes identical label sets across domains, which is unrealistic. To overcome this limitation, we propose a study on BDA with unknown classes in the target domain. It uses a black-box predictor to label target data and identify "unknown" categories, without requiring access to source domain data or predictor parameters, thus addressing both data privacy and category shift issues in traditional UDA. Existing methods face two main challenges: (i) Noisy pseudo-labels in knowledge distillation (KD) accumulate prediction errors, and (ii) relying on a preset threshold fails to adapt to varying category shifts. To address these, we propose **ADU**, a framework that allows the target domain to autonomously learn pseudo-labels guided by quality and use an adaptive threshold to identify "unknown" categories. Specifically, ADU consists of Selective Amplification Knowledge Distillation (SAKD) and Entropy-Driven Label Differentiation (EDLD). SAKD improves KD by focusing on high-quality pseudo-labels, mitigating the impact of noisy labels. EDLD categorizes pseudo-labels by quality and applies tailored training strategies to distinguish "unknown" categories, improving detection accuracy and adaptability. Extensive experiments show that ADU achieves state-of-the-art results, outperforming the best existing method by 3.1% on VisDA in the OPBDA scenario.*

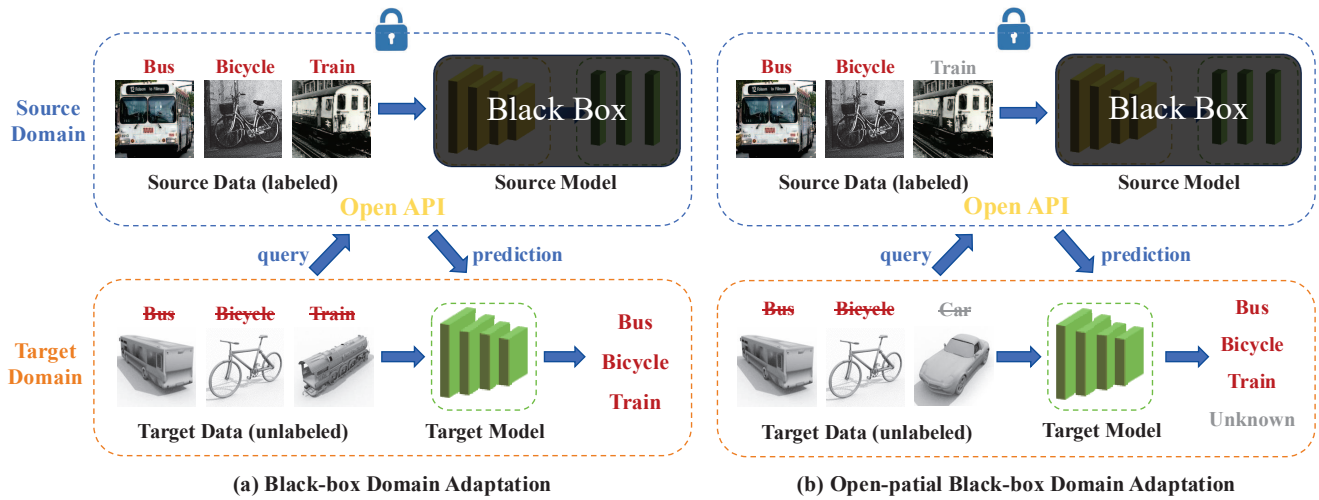
1. Introduction

Unsupervised domain adaptation (UDA) [12] aims to transfer knowledge from a well-labeled source domain to an unlabeled target domain, which can ease the burden of manual labeling. Recently, UDA has been applied in a range of computer vision tasks, including image classification [13, 33, 48], objection detection [7, 20, 56] and semantic segmentation [6, 35, 47]. However, UDA methods may raise concerns about data privacy and portability issues due to their requirement for access to raw source data and source model parameters. Therefore, source-free domain adaptation (SFDA) [19, 27, 58] is proposed to protect the source data privacy. In the SFDA scenario, only the source model is provided to the target domain without access to source data. However, it still faces the issue of source information privacy, which can be compromised through techniques such as white-box attacks [45, 46]. To mitigate these concerns, Black-box Domain Adaptation (BDA) [28, 61] has been proposed recently, as shown in Figure 1(a), which aims to learn a model solely using the unlabeled data from the target domain, based on the predictions from a black-box predictor trained on the source data. This setting can effectively mitigate data privacy issues related to data and model parameter leakage.

However, traditional BDA [28, 57, 60] always assumes that the source and target domains share identical category sets, which frequently fails to apply in practice. In real-world scenarios, the target domain is typically unlabeled, making it difficult to satisfy this assumption due to potential category shifts. Currently, there are two UDA settings that involve unknown classes in the target domain: Open-Set Domain Adaptation (OSDA) [36, 42] and Open-Partial Domain Adaptation (OPDA) [25, 41, 59]. OSDA deals with scenarios where the target domain contains private classes that are unknown to the source domain, while OPDA handles cases where both the source and target domains each have their own private classes. Black-box Domain Adaptation has been applied to OPDA recently [9]. As shown in Figure 1(b), this setting is designed to learn a robust model for the target domain that not only recognizes classes shared by two domains but also identifies "unknown" categories absent in the source domain despite having no information about difference of two label sets.

Currently, only one study has addressed the above problem. [9] applies knowledge distillation to train the target model to mimic source predictor outputs and uses a man-

*Corresponding author



(a) Black-box Domain Adaptation (b) Open-patial Black-box Domain Adaptation

Figure 1. Black-box domain adaptation and Open-partial black-box domain adaptation settings with respect to label sets of source and target domains (red labels indicate common labels of two domains). Compared to BDA, ADU is able to deal with BDA with unknown classes in the target by adaptively detecting unknown categories.

ually preset threshold to identify "unknown" categories. Though inspiring, it still has the following limitations. (i) Due to domain and category shifts between the source and target domains, predictions from the source model are inevitably noisy. Directly utilizing these noisy pseudo-labels will accumulate model prediction errors, making the adaptation process unreliable. (ii) Employing a preset threshold fails to accommodate the variability and complexity of category shifts in different target domains, which is inadequate for accurately detecting "unknown" classes across diverse domains, often resulting in misclassification and reduced adaptability.

To address the issues mentioned above, we propose a simple yet effective framework called **ADU**, specifically designed for Open-Set BDA (OSBDA) and Open-Partial BDA (OPBDA). ADU incorporates two core modules: Selective Amplification Knowledge Distillation (SAKD) and Entropy-Driven Label Differentiation (EDLD). For the first challenge, SAKD enhances traditional knowledge distillation techniques, specifically tailoring KD to BDA with unknown classes in the target domain by amplifying learning from high-quality pseudo-labels produced by source API. This refinement ensures that the target model emphasizes learning from high-quality pseudo-labels, effectively mitigating the impact of noisy data. For the second challenge, EDLD enhances the framework's ability to handle diverse domain conditions. Initially, EDLD categorizes pseudo-labels based on their quality and then applies tailored training strategies to widen the distance between "unknown" classes and the others, while minimizing the impact of noisy pseudo-labels. This adaptive differentiation of labels heightens the effectiveness of employing the average entropy of the target model's predictions as a threshold. Consequently, this refined approach significantly improves the detection accuracy of "unknown" categories and adapts more adeptly to category shifts across various target domains. Additionally, we iteratively refine the pseudo-labels generated by the source API, which can significantly enhance their quality.

Our main contributions in this paper could be summarized as follows:

1. We propose Selective Amplification Knowledge Distillation (SAKD), a refined knowledge distillation technique specifically designed for the OPBDA and OSBDA scenarios, which can effectively mitigate the impact of noisy pseudo-labels.
2. We introduce Entropy-Driven Label Differentiation (EDLD), which categorizes pseudo-labels by quality and applies customized training strategies to enhance the distinction between "unknown" and others, thereby improving detection accuracy and domain adaptability through adaptive entropy-based thresholding.
3. Extensive experiments on four public benchmarks demonstrate the superior performance of our proposed method compared with existing SOTA works, surpassing the best existing method by 3.1% on VisDA in the OPBDA scenario.

2. Related work

Black-box domain adaptation. Unsupervised Domain Adaptation (UDA) [12] aims to adapt a model trained on a labeled source domain to an unlabeled target domain. Many early methods relied on techniques such as instance weighting [52, 55], feature transformation [18, 26, 43], and feature

space [30, 51]. Despite their effectiveness, these methods require access to source domain data, raising privacy and portability concerns [22]. To address privacy issues associated with UDA, Source-Free Domain Adaptation (SFDA) methods [19, 27, 58] have been proposed. These methods adapt models using only the source model and unlabeled target data, eliminating the need for source data during adaptation. Techniques such as entropy minimization [2] and pseudo-labeling [62] have been explored. However, SFDA methods still face potential privacy risks due to the use of generative models and other techniques that might inadvertently reveal source data characteristics. Therefore, Black-box Domain Adaptation (BDA) [28, 61] has emerged as a solution to further mitigate privacy concerns by only accessing the source model's outputs without any internal details. This approach ensures better privacy preservation compared to traditional UDA and SFDA methods. Recent methods such as DINE [28] and BETA [57] have made significant strides in this area. Nevertheless, they struggle with inconsistent label sets between domains.

Open-set and open-partial domain adaptation. Closed-set domain adaptation assumes identical label sets between source and target domains, focusing on minimizing distribution shifts using techniques like discrepancy minimization [31, 32] and adversarial training [8, 15]. However, these methods often struggle when label sets are not perfectly aligned. To address this issue, Partial Domain Adaptation (PDA) assumes that only the source domain contains private classes, with methods such as SAN [4] employing class-wise domain discriminators, and ETN [5] using progressive weighting schemes. Meanwhile, Open-Set Domain Adaptation (OSDA) handles scenarios where the target domain has private classes unknown to the source. Besides, Open-Partial Domain Adaptation (OPDA) addresses both domains having their own private classes. UAN [59] quantifies sample-level uncertainty using entropy and domain similarity, and Fu et al. [11] combines entropy, confidence, and consistency for better uncertainty measurement. To address the challenges faced by black-box domain adaptation, [9] combines OPDA with BDA to address both category shift and privacy concerns. It applies knowledge distillation to train the target model to emulate source predictor outputs, using a preset threshold to identify "unknown" categories. Though inspiring, it still faces significant limitations regarding pseudo-label quality and the detection of "unknown" categories. To address these issues, we propose the ADU framework, applying it to Open-Set BDA (OSBDA) and Open-Partial BDA (OPBDA) to mitigate the impact of noisy pseudo-labels and enhance adaptability to category shifts across varied target domains. This approach provides a robust solution to the limitations of existing methods.

Learning with noisy labels. Deep learning models often overfit on noisy labels, leading to poor generalization [60]. To address this, various approaches have been proposed, including noise-robust losses [23, 44], noise-transition matrix estimation [14], clean sample selection [53], and loss reweighting [34]. However, these methods often require noise-free validation sets or make assumptions about the noise distribution, which are impractical in BDA settings. These methods [29, 34] differ by not assuming any specific noise distribution and leveraging noisy scores from source training classes. Recently, NEL [1] introduced a novel approach by integrating a Negative Learning loss with a pseudo-label refinement framework that leverages ensembling techniques. Negative Learning [23] is an indirect learning method that employs complementary labels to address noise issues effectively. In our work, we use Negative Learning to refine high-quality pseudo-labels without ensembling, reducing computational cost and making our approach more flexible and robust for OSBDA and OPBDA.

3. Methodology

In this paper, we are provided with a target domain $\mathcal{D}_t = \{x_t^i\}_{i=1}^{N_t}$ with N_t unlabeled samples where $x_t^i \in \mathcal{X}_t$, and a black-box predictor f_s trained by a source domain $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ with N_s labeled samples where $x_s^i \in \mathcal{X}_s$. We use L_s and L_t to denote the label spaces of the source domain and target domain respectively. In general, model f consists of a feature extractor $\boldsymbol{G}$ and a fully connected layer-based classifier $\boldsymbol{C}$. We have no access to the source domain data $\mathcal{D}_s$ and the parameters of the source model f_s. Only a black-box predictor trained on the source domain, i.e., an API, is available. The objective is to leverage the predictions of the API of the source domain to learn a mapping model f_t which can label the target samples with either one of the L_s labels or the "unknown" label. The overall workflow is shown in Fig. 2.

3.1. Selective amplification knowledge distillation

Knowledge distillation (KD) [17] has been widely applied to address the black-box domain adaptation problem [9, 28, 57], as it enables the transfer of knowledge from one model (teacher) to another (student) by guiding the target model (student) to emulate the predictions of the source model (teacher), This approach is particularly suitable for BDA scenarios, where only the predictions of the source model are accessible. To better leverage the information available from the source domain's API, we use a knowledge distillation loss with both the source model's probabilities and hard pseudo-labels. This can be formulated as:

$$\mathcal{L}_{KD} = \mathbb{E}_{x_t \sim \mathcal{X}_t}[CE(\tilde{\boldsymbol{y}}_t, p_t) + CE(p_s, p_t)], \quad (1)$$

where $CE(\cdot, \cdot)$ denotes the cross entropy function, and

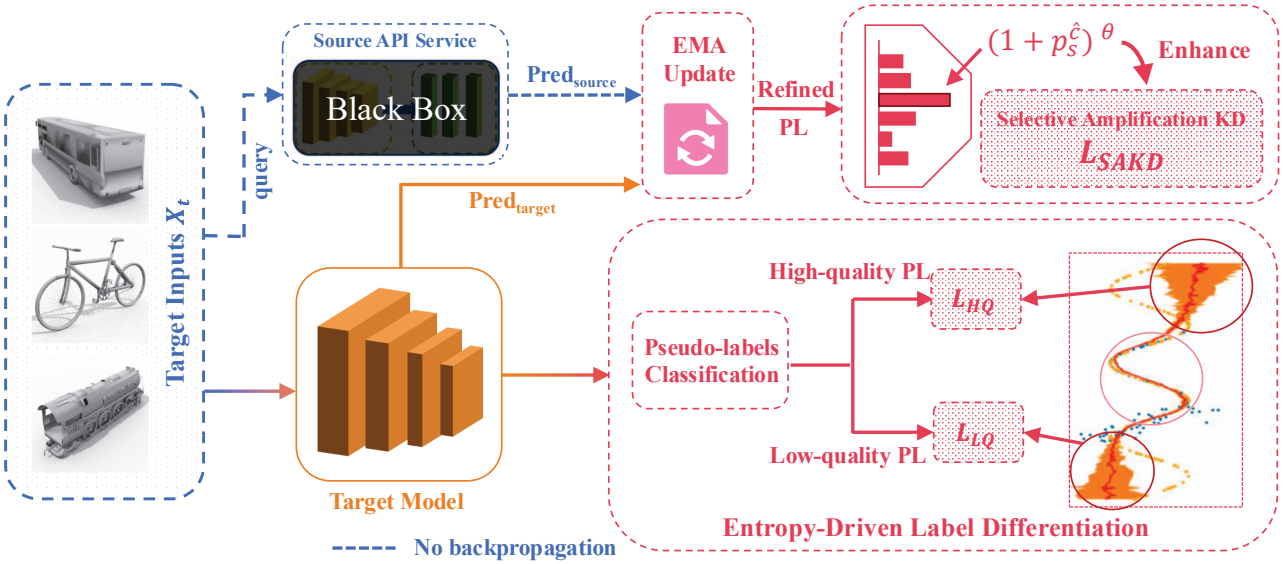


Figure 2. An overview of the proposed ADU framework. We utilize the black-box source predictor solely as an API service, obtaining only the source predictions from it. "PL" in the figure means pseudo-labels.

$\tilde{\boldsymbol{y}}_t$ is a one-hot pseudo-label derived from $f_s(x_t)$. In addition, we use p_s and p_t to replace $f_s(x_t)$ and $f_t(x_t)$ for simplity. However, due to domain and category shifts, the predictions from the source model are inevitably noisy. Consequently, Eq. (1) processes information from these predictions equally, which can adversely affect the performance of the target model. In order to solve the issue, we propose Selective Amplification Knowledge Distillation (SAKD), a method that enhances knowledge distillation by leveraging the confidence of pseudo-labels produced by the source model.

Firstly, we simplify Eq. (1) to derive the following formulation:

$$\begin{aligned}\mathcal{L}_{KD} &= -\mathbb{E}_{x_t\sim\mathcal{X}_t}(\log p_t^{\hat{c}} + \sum_{c=1}^{|L_s|} p_s^c \log p_t^c) \\ &= -\mathbb{E}_{x_t\sim\mathcal{X}_t}[(1+p_s^{\hat{c}})\log p_t^{\hat{c}} + \sum_{\substack{c=1\\c\neq\hat{c}}}^{|L_s|} p_s^c \log p_t^c],\end{aligned} \tag{2}$$

where $\hat{c}$ represents the label predicted by the source model, defined as $\hat{c} = \arg\max_c p_s^c$. Subsequently, we formulate the SAKD loss by incorporating a modulating parameter $\theta \geq 1$ into the term $\left(1+p_s^{\hat{c}}\right)$, which only pertains to $\hat{c}$:

$$\mathcal{L}_{SAKD} = -\mathbb{E}_{x_t\sim\mathcal{X}_t}[(1+p_s^{\hat{c}})^{\theta}\log p_t^{\hat{c}} + \sum_{\substack{c=1\\c\neq\hat{c}}}^{|L_s|} p_s^c \log p_t^c], \tag{3}$$

Discussion of SAKD loss: For simplicity, we consider the case with a single sample, where the SAKD loss simplifies to:

$$\mathcal{L}_{\text{SAKD}} = -[(1+p_s^{\hat{c}})^{\theta}\log p_t^{\hat{c}} + \sum_{c=1,c\neq\hat{c}}^{|L_s|} p_s^c \log p_t^c] \tag{4}$$

Next, we apply the generalized binomial theorem to expand $\left(1+p_s^{\hat{c}}\right)^{\theta}$ as follows:

$$\begin{aligned}\mathcal{L}_{\text{SAKD}} &= -\left[\sum_{k=0}^{\infty}\binom{\theta}{k}\left(p_s^{\hat{c}}\right)^k \log p_t^{\hat{c}} + \sum_{c=1,c\neq\hat{c}}^{|L_s|} p_s^c \log p_t^c\right] \\ &= \mathcal{L}_{\text{KD}} - \left[\sum_{k=1}^{\infty}\binom{\theta}{k}\left(p_s^{\hat{c}}\right)^k - p_s^{\hat{c}}\right]\log p_t^{\hat{c}} \\ &\approx \mathcal{L}_{\text{KD}} - \left[(\theta-1)\,p_s^{\hat{c}} + \frac{\theta(\theta-1)}{2}\left(p_s^{\hat{c}}\right)^2\right]\log p_t^{\hat{c}}\end{aligned} \tag{5}$$

In Eq. (5), the first term $\mathcal{L}_{\text{KD}}$ represents the original KD loss in Eq. (2), while the second term introduces an additional term, which is positive and solely depends on the target class $\hat{c}$. We demonstrate that this additional term enables the SAKD loss to capture more information from pseudo-labels with high confidence, meaning those with higher values of $p_s^{\hat{c}}$, thereby reducing the impact of noisy pseudo-labels. A comprehensive proof of this claim is provided in the supplemental material.

3.2. Entropy-driven label differentiation

As stated above, the outputs from the source model are highly likely to be inaccurate and noisy due to the domain shift [3] and category shift. Even if we propose a promising solution in Eq. (3), we still face a tough challenge to detect the "unknown" categories, which means we should widen the difference between "unknown" and others. [59] shows entropy is an effective tool to detect "unknown" in domain adaptation. Entropy quantifies the prediction uncertainty, and smaller entropy represents a more certain prediction. In order to effectively address the influence brought by the category shift, we implement an automatic threshold determined by average entropy. The prediction process can be formulated as follows:

$$y_t = \begin{cases} \arg\max_c p_t^c & H(p_t) < w \\ \text{unknown} & H(p_t) \geq w, \end{cases} \tag{6}$$

where $H(p_t)$ and w are computed as:

$$H(p_t) = -\sum_{c=1}^{|L_s|} p_t^c \log p_t^c, \tag{7}$$

$$w = \mathbb{E}_{x_t \sim \mathcal{X}_t} H(p_t). \tag{8}$$

Taking the average as a threshold eliminates the requirement of per-dataset hyper-parameter tuning and makes our selection process highly adaptive. As we employ entropy as a threshold to detect "unknown" categories, we still face a challenge to widen the gap between "unknown" and others. In order to address the issue, we propose Entropy-Driven Label Differentiation (EDLD) to make the "unknow" distinguishable and enhance the quality of pseudo-labels with high certainty. We use entropy to calculate the uncertainty level of pseudo-labels. Higher entropy always shows more uncertain predictions. We define the EDLD loss by dividing pseudo-labels into high-quality (HQ) and low-quality (LQ) by their entropy, the loss is defined as follows:

$$\mathcal{L}_{EDLD} = \mathbb{E}_{x_t \sim \mathcal{X}_t} \left[\begin{cases} \mathcal{L}_{HQ}(p_t) & \text{if } H(p_t) < w \\ \mathcal{L}_{LQ}(p_t) & \text{if } H(p_t) \geq w \end{cases} \right], \tag{9}$$

$$\mathcal{L}_{HQ} = H(p_t) + \mathcal{L}_{NL}(p_t, \bar{y}_t), \qquad \mathcal{L}_{LQ} = -H(p_t), \tag{10}$$

In the EDLD module, we not only use the entropy loss to widen the entropy gap between high-quality and low-quality pseudo-labels but also use a negative learning loss [21] to refine the high-quality pseudo-labels, the negative loss is the following:

$$\mathcal{L}_{NL}(p_t, \bar{y}_t) = -\sum_{c=1}^{|L_s|} \bar{\boldsymbol{y}}_t^c \log\left(1 - p_t^c\right), \tag{11}$$

where is $\bar{y}_t$ a complementary label $\bar{y}_t \in \{1, ..., |L_s|\} \setminus \{y_t\}$ chosen randomly from the set of labels, and $\bar{\boldsymbol{y}}_t$ is one-hot label derived from $\bar{y}_t$. Eq. (11) enables the probability value of the complementary label to be optimized as zero, resulting in an increase in the probability values of other classes, which can effectively refine the high-quality pseudo-labels.

3.3. Adaptive refinement of pseudo labels

To further mitigate the impact of noise in the pseudo-labels generated by the source model, we employ an exponential moving average (EMA) of the target predictions. This allows for a gradual and controlled update of the pseudo-labels supplied by the source model at each iteration. The update process is defined as follows:

$$p_s \leftarrow \gamma p_s + (1 - \gamma) f_t(x_t), \quad \forall x_t \in \mathcal{X}_t, \tag{12}$$

where γ is a smoothing factor that determines the extent to which the pseudo-labels should adapt to the most recent predictions from the target model. A higher value of γ places more weight on the existing pseudo-labels, while a lower value allows quicker adaptation to new information.

This strategy refines the pseudo-labels iteratively, balancing consistency with adaptability. By adjusting the pseudo-labels in a controlled manner, the model can better handle noise and gradually align the source model's outputs with the distribution of the target data. The EMA strategy ensures that updates are not overly reactive to fluctuations, thus enhancing the robustness of the model and improving performance in scenarios with diverse target domains.

3.4. The overall objective

Integrating these objectives introduced in Eqs. (3, 9) together, we obtain the final loss function as follows:

$$\mathcal{L} = \mathcal{L}_{SAKD} + \lambda \mathcal{L}_{EDLD}, \tag{13}$$

where λ is a hyper-parameter empirically set to 1.0, controlling the importance of L_{SAKD} and L_{EDLD} during distillation.

4. Experiments

4.1. Setup

Datasets. To assess the effectiveness of our approach, we conduct experiments using the **Office31** [40], **OfficeHome** [50], **VisDA** [38], **DomainNet** [39] datasets. **Office31** is a popular benchmark for UDA, consisting of three domains (**A**mazon, **W**ebcam, **D**slr) in 31 categories. **OfficeHome** is a more challenging benchmark for its distant domain shifts, which consists of four domains (**Ar**t, **Cl**ipart, **Pr**oduct, **Re**al World) in 65 categories. **VisDA** is a large-scale benchmark containing 2 different 12-class domains, with a source domain with 152k synthetic images and a target domain with

Table 1. H-score (%) comparison in OPBDA scenario on the OfficeHome dataset.

Method	Ar→Cl	Ar→Pr	Ar→Re	Cl→Ar	Cl→Pr	Cl→Re	Pr→Ar	Pr→Cl	Pr→Re	Re→Ar	Re→Cl	Re→Pr	Avg.
No Adapt.	55.8	67.3	72.8	64.2	62.3	70.5	65.7	52.1	71.7	66.1	56.7	69.2	64.5
DINE [28]	45.3	46.1	54.6	51.0	45.3	52.4	49.9	44.5	52.1	52.4	46.7	45.7	48.8
BETA [57]	45.9	47.4	54.8	49.3	45.1	50.1	49.3	45.5	53.5	51.5	45.8	48.8	48.9
SEAL [54]	40.6	46.8	47.8	44.5	42.7	45.2	47.3	40.0	47.1	45.5	46.7	46.6	45.1
UB^{2}DA [9]	60.9	69.6	76.3	**74.4**	69.2	76.5	74.5	60.3	76.2	**74.1**	62.0	71.1	70.4
ADU	**61.2**	**72.7**	**77.9**	70.3	**72.5**	**77.3**	**75.9**	**62.0**	**84.7**	73.2	**64.1**	**74.9**	**72.2**

Table 2. H-score (%) comparison in OPBDA scenario on the Office31, VisDA, and DomainNet datasets, respectively.

Method	Office31							VisDA	DomainNet						
	A→D	A→W	D→A	D→W	W→A	W→D	Avg.	S→R	P→R	P→S	R→P	R→S	S→P	S→R	Avg.
No Adapt.	79.9	71.9	80.1	91.5	78.7	89.8	82.0	37.7	52.8	35.0	43.6	32.5	35.7	51.8	41.9
DINE [28]	50.3	51.4	56.6	63.0	54.0	60.1	55.9	43.5	48.4	39.5	43.4	38.1	37.6	45.6	42.1
BETA [57]	52.4	54.0	51.6	61.2	53.4	57.6	55.0	45.5	49.2	40.3	43.1	38.2	38.0	48.4	42.9
SEAL [54]	73.8	70.3	51.1	55.6	47.0	57.0	59.1	45.6	52.7	39.9	43.6	38.4	39.2	49.7	43.9
UB^{2}DA [9]	80.9	78.2	**92.6**	92.6	**89.4**	87.9	86.9	45.2	57.1	47.2	**54.8**	44.0	41.4	51.5	49.3
ADU	**87.5**	**85.2**	87.0	**94.4**	83.8	**90.5**	**88.1**	**48.7**	**59.8**	**47.8**	52.5	**46.6**	**42.7**	**56.4**	**51.0**

55k real images from Microsoft COCO. **DomainNet** is the largest DA dataset with about 0.6 million images. Like [11, 24], we conduct experiments on three subsets of it, i.e., **P**ainting, **R**eal, and **S**ketch. Following existing works [11, 24, 59], we separate the label set into three parts: common ($|L_s \cap L_t|$), source-private ($|L_s - L_t|$) and target-private ($|L_t - L_s|$). The classes are separated according to their alphabetical order. We evaluate ADU in OPBDA using the four datasets, and in OSBDA using the first three datasets.

Evaluation protocols. Considering the trade-off between the accuracy of known and unknown classes is important in evaluating OSDA and OPDA methods. We evaluate methods using **H-score** [11]. H-score is the harmonic mean of the accuracy on common classes (Acc_c) and accuracy on the "unknow" classes (Acc_u) and is defined as:

$$h = 2 \cdot \frac{Acc_c \cdot Acc_u}{Acc_c + Acc_u} \tag{14}$$

So, this metric is designed to provide a more comprehensive evaluation by ensuring that improvements in one area do not come at the expense of the other. It can measure both accuracies well.

Implementation details. All experiments are implemented in Pytorch [37]. For fair comparisons to previous methods, we use the same backbone of ResNet50 [16] pre-trained on ImageNet [10] as the feature extractor in all experiments. For the source model, we fine-tune the model on source examples optimizing with cross-entropy loss function and then treat it like a black-box by only requiring the input-output interfaces of this model in our experiments. We use SGD optimizer with a learning rate of 0.01, a momentum of 0.9 with a weight decay of 5e-4 and a batch size of 128. Concerning the parameters in ADU, We set $\theta = 1.1$, $\gamma = 0.6$ and $\lambda = 1.0$ for all datasets and tasks. Additionally, following [23], we set the ratio of $L_{NL}(p_t, \bar{y}_t)$ to $H(p_t)$ as 0.01:1 in Eq. (10).

Baselines. We compare the proposed ADU with (i) BDA: **DINE** [28], **BETA** [57], **SEAL** [54] (ii) OPBDA: **UB^{2}DA** [9]. These methods represent the state-of-the-art in their respective settings. Notably, owing to black-box DA lacking the capability to identify "unknown" categories, we apply the average entropy as the threshold to them, similar to our setting. The term "No Adapt." refers to the baseline scenario where the source model is used directly for target label prediction, without any form of adaptation.

4.2. Results

Results for OPBDA. We first perform experiments under the most challenging scenario, namely OPBDA, in which both the source and target domains contain private categories. The results for the OfficeHome dataset are presented in Table 1, while those for the Office31, VisDA, and DomainNet datasets are shown in Table 2. As illustrated in these tables, our proposed ADU method achieves a new state-of-the-art, surpassing all existing methods across the four datasets. Notably, ADU consistently improves the H-score compared to the "No Adapt." baseline in each experimental setting, with a significant increase of 11.0% on the VisDA dataset. This improvement demonstrates that our method effectively mitigates the influence of noise from

Table 3. H-score (%) comparison in OSBDA scenario on the OfficeHome dataset.

Method	Ar→Cl	Ar→Pr	Ar→Re	Cl→Ar	Cl→Pr	Cl→Re	Pr→Ar	Pr→Cl	Pr→Re	Re→Ar	Re→Cl	Re→Pr	Avg.
No Adapt.	59.6	68.1	75.7	67.1	66.7	70.4	63.8	54.9	55.7	71.4	58.6	70.5	65.2
DINE [28]	47.0	45.8	52.2	49.7	47.1	50.0	48.2	43.8	50.4	52.6	46.0	46.8	48.3
BETA [57]	46.6	48.3	54.9	47.7	48.4	50.5	49.1	42.9	51.8	50.2	45.2	48.9	48.7
SEAL [54]	43.3	46.5	47.1	43.7	45.9	45.4	45.3	40.8	45.6	43.5	41.3	46.5	44.6
UB^{2}DA [9]	65.5	70.4	75.5	67.8	69.3	74.4	**71.2**	56.7	75.0	70.7	63.3	69.8	69.1
ADU	**66.0**	**70.5**	**77.4**	**72.2**	**70.1**	**75.0**	69.0	**63.6**	**76.1**	**73.4**	**64.1**	**75.2**	**71.1**

Table 4. H-score (%) comparison in OSBDA scenario on the Office31 and VisDA datasets, respectively.

Method	Office31							VisDA
	A→D	A→W	D→A	D→W	W→A	W→D	Avg.	S→R
No Adapt.	81.4	80.8	85.3	88.1	78.0	88.0	83.6	44.6
DINE [28]	60.5	54.6	56.8	69.0	56.3	61.5	59.8	43.1
BETA [57]	48.3	53.0	54.3	60.6	54.4	57.3	54.7	48.3
SEAL [54]	50.4	43.7	53.4	44.8	54.0	50.8	49.5	40.6
UB^{2}DA [9]	85.7	**87.4**	**91.0**	89.2	85.1	84.1	87.1	48.1
ADU	**86.9**	84.9	89.7	**91.3**	**86.3**	**89.2**	**88.1**	**50.8**

source model predictions on the target model and accurately identifies unknown categories within the target data. An examination of Tables 1 and 2 reveals that methods such as DINE [28], BETA [57], and SEAL [54] perform poorly compared to our approach and UB^{2}DA [9], with performance even falling below the "No Adapt." baseline on the Office31 and OfficeHome datasets. This underperformance is likely due to the lack of design tailored specifically for the OPBDA scenario in these methods, which hinders their ability to effectively differentiate unknown categories from other classes. These results underscore the importance of our ADU approach, which is specifically designed for OPBDA. When compared to UB^{2}DA [9], ADU achieves higher H-scores on the Office31, OfficeHome, VisDA, and DomainNet datasets, with improvements of 1.2%, 1.8%, 3.5%, and 1.7%, respectively. These gains further highlight the effectiveness of our proposed approach.

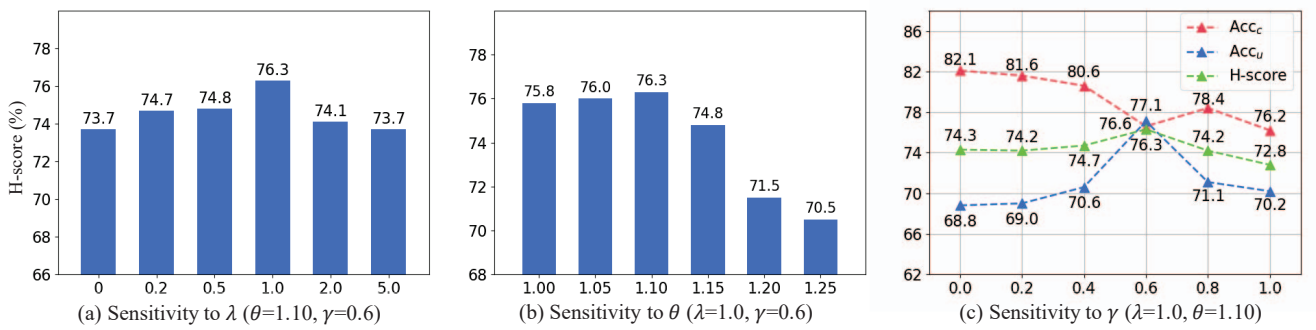
Results for OSBDA. We subsequently conduct experiments under OSBDA scenarios, where only the target domain includes categories absent from the source domain. The results for the OfficeHome dataset are provided in Table 3, while those for the Office31 and VisDA datasets are presented in Table 4. As shown in these tables, our proposed ADU method achieves performance that surpasses the current state-of-the-art. Specifically, ADU consistently outperforms the "No Adapt." baseline in terms of H-score across all experimental settings. Notably, for the Pr→Re scenario, it achieves an improvement of 20.4%. This substantial enhancement demonstrates that our method effectively reduces the influence of noise from source model predictions on the target model, allowing for accurate identification of unknown categories within the target data. Compared to UB^{2}DA [9], ADU achieves higher H-scores on the Office31, OfficeHome and Visda datasets, with improvements of 1.0%, 2.0%, and 2.7%, respectively. These results further validate the effectiveness of our proposed approach.

Table 5. Ablation Study. H-score (%) of different variants in OPBDA scenarios. $\mathcal{L}^1_{HQ}$, $\mathcal{L}^2_{HQ}$, $\mathcal{L}_{LQ}$ refer to the objectives corresponding to the negative loss in $\mathcal{L}_{HQ}$, entropy loss in $\mathcal{L}_{HQ}$, and loss associated with low-quality labels, respectively.

$\mathcal{L}^1_{HQ}$	$\mathcal{L}^2_{HQ}$	$\mathcal{L}_{LQ}$	Office31		OfficeHome				Avg.
			A → W	D → W	Ar → Cl	Cl → Re	Pr → Ar	Re → Cl	
-	-	-	82.9	92.6	61.2	72.9	71.5	62.0	73.7
✓	-	-	84.3	93.5	61.1	73.0	72.0	62.7	74.6
-	✓	-	83.5	92.7	61.7	74.3	72.1	62.5	74.5
-	-	✓	84.1	93.8	61.4	74.1	72.6	62.6	74.8
✓	✓	-	84.3	94.1	62.7	73.7	71.7	61.3	74.6
✓	-	✓	83.1	94.2	**63.3**	74.7	72.3	62.6	75.0
-	✓	✓	85.1	94.0	62.6	74.6	72.1	62.8	75.2
✓	✓	✓	**85.2**	**94.4**	61.2	**77.3**	**75.9**	**64.1**	**76.3**

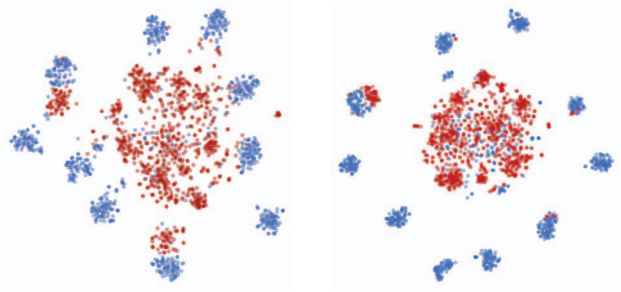
4.3. Analysis

Ablation study. To comprehensively assess the individual contribution of the components comprising our method, we conduct extensive ablation studies on two tasks from the Office31 dataset and four tasks from the OfficeHome dataset in OPBDA scenarios. The results are summarized in Table 5. Here, $\mathcal{L}^1_{HQ}$, $\mathcal{L}^2_{HQ}$, $\mathcal{L}_{LQ}$ refer to the objectives corresponding to the negative loss in $\mathcal{L}_{HQ}$, entropy loss in $\mathcal{L}_{HQ}$, and loss associated with low-quality labels, respectively. More detailed results can be found in supplementary material. It is important to emphasize that in all ablation experiments, we consistently employ the SAKD loss, which is a critical component of the ADU framework. Without it, the model would be unable to transfer knowledge from the source domain to the target model effectively. From the ablation study results, we can draw the following conclusions: (i) The introduction of any component alongside the SAKD loss leads to performance improvements, underscoring the vital role of the EDLD module. (ii) The full EDLD loss, which includes the negative loss term, yields better performance compared to its version without the negative loss,



(a) Sensitivity to λ (θ=1.10, γ=0.6) (b) Sensitivity to θ (λ=1.0, γ=0.6) (c) Sensitivity to γ (λ=1.0, θ=1.10)

Figure 3. Parameters sensitivity analysis for six tasks. (a-b) plot the H-score with different values of λ, θ; (c) plot the H-score, Acc$_c$, Acc$_u$ with different values of γ. The default values of these hyperparameters are set to $\lambda = 1.0$, $\theta = 1.1$, and $\gamma = 0.6$.



(a) No Adapt. (b) ADU

Figure 4. t-SNE feature visualization of target representations in D→A OPBDA task. Blue dots represent target "known" examples ($L_s \cap L_t$) while red dots are "unknown" examples ($L_s - L_t$).

demonstrating the effectiveness of incorporating this term. (iii) The integration of all components results in the highest H-scores, providing clear evidence of the synergy and efficacy of the combined modules.

Feature visualization. Fig. 4 displays the visualization of the target feature with t-SNE [49], providing a clear representation of the feature distribution. As expected, ADU achieves excellent alignment between the source and target domain features. Taking a closer look at the visualization, it is evident that ADU excels in distinguishing the "unknown" categories from the other classes. This improvement aligns well with the intended function of the EDLD module, which is designed to enhance the separation of "unknown" categories from known categories. This result further highlights the effectiveness of ADU in handling the challenges posed by unknown classes in black-box domain adaptation tasks.

Parameters sensitivity analysis. To better assess the impact of different hyperparameters, we conduct a detailed sensitivity analysis. We investigate the sensitivity of the parameters λ, θ, and γ by performing experiments on two tasks from the Office31 dataset and four tasks from the OfficeHome dataset in OPBDA scenarios, as shown in Fig. 3. The parameter λ is varied over the range [0.0, 0.2, 0.5, 1.0, 2.0, 5.0], θ spans [1.00, 1.05, 1.10, 1.15, 1.20, 1.25], and γ is explored within the range [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]. It is evident that the results are stable around the selected values of $\lambda = 1.0$, $\theta = 1.1$, and $\gamma = 0.6$. Additionally, as shown in Fig. 3(b), we examine the effect of varying θ in Eq. (3). The results around the chosen parameter $\theta = 1.1$ remain stable, and we also observe that increasing θ slightly from 1.0 leads to an improvement in the H-score, thereby highlighting the effectiveness of Eq. (3). Finally, we analyze the impact of γ. As shown in Fig. 3(c), there is an inverse relationship between Acc$_c$ and Acc$_u$. However, when $\gamma = 0.6$, the two metrics reach a relatively balanced state, and at this point, the H-score achieves an optimal result.

5. Conclusion

In this paper, we introduce the ADU model, a framework specifically designed to tackle Black-box Domain Adaptation with unknown classes in the target domain. ADU integrates two key innovations: Selective Amplification Knowledge Distillation (SAKD) and Entropy-Driven Label Differentiation (EDLD). SAKD enhances model accuracy by selectively amplifying high-confidence pseudo-labels, thereby effectively mitigating the influence of noisy pseudo-labels. Meanwhile, EDLD improves the recognition of unknown categories through an entropy-driven threshold, expanding the difference between unknown categories and others and bolstering the robustness of the method across a range of diverse target domains. Experiments across four benchmark datasets demonstrate that ADU outperforms existing state-of-the-art approaches, highlighting its exceptional adaptability and efficacy, setting a new benchmark for future research in the field.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant T2125006 and 42401415) and Jiangsu Innovation Capacity Building Program (Project BM2022028).

References

[1] Waqar Ahmed, Pietro Morerio, and Vittorio Murino. Cleaning noisy labels by negative ensemble learning for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1616–1625, 2022. 3

[2] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82: 102617, 2022. 3

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 5

[4] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, 2018. 3

[5] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2985–2994, 2019. 3

[6] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2090–2099, 2019. 1

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 1

[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2020. 3

[9] Bin Deng, Yabin Zhang, Hui Tang, Changxing Ding, and Kui Jia. On universal black-box domain adaptation. *arXiv preprint arXiv:2104.04665*, 2021. 1, 3, 6, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 567–583. Springer, 2020. 3, 6

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 1

[14] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International conference on learning representations*, 2022. 3

[15] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2477–2486, 2019. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[18] Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *International journal of computer vision*, 109:28–41, 2014. 2

[19] Yunzhong Hou and Liang Zheng. Source free domain adaptation with image translation. *arXiv preprint arXiv:2008.07514*, 2020. 1, 3

[20] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019. 1

[21] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 101–110, 2019. 5

[22] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021. 3

[23] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9442–9451, 2021. 3, 6

[24] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9757–9766, 2021. 6

[25] Qingmei Li, Yibin Wen, Juepeng Zheng, Yuxiang Zhang, and Haohuan Fu. Hyunida: Breaking label set constraints for universal domain adaptation in cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[26] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. Domain invariant and class discriminative feature learning for visual domain adaptation. *IEEE transactions on image processing*, 27(9):4260–4273, 2018. 2

[27] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020. 1, 3

[28] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8003–8013, 2022. 1, 3, 6, 7

[29] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7640–7650, 2023. 3

[30] Long Liu, Lechao Yang, and Bin Zhu. Sparse feature space representation: A unified framework for semi-supervised and domain adaptation learning. *Knowledge-Based Systems*, 156:43–61, 2018. 3

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 3

[33] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 1

[34] Bekhzod Olimov, Jeonghong Kim, and Anand Paul. Dcbt-net: Training deep convolutional neural networks with extremely noisy labels. *IEEE Access*, 8:220482–220495, 2020. 3

[35] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3764–3773, 2020. 1

[36] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017. 1

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

[38] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 5

[39] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 5

[40] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010. 5

[41] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 9000–9009, 2021. 1

[42] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 153–168, 2018. 1

[43] Chen Shen and Yuhong Guo. Unsupervised heterogeneous domain adaptation with sparse feature transformation. In *Asian conference on machine learning*, pages 375–390. PMLR, 2018. 2

[44] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 2022. 3

[45] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1

[46] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 1

[47] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 1

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1

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

[50] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 5

[51] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5495–5504, 2018. 3

[52] Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, 2017. 2

[53] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13726–13735, 2020. 3

[54] Mingxuan Xia, Junbo Zhao, Gengyu Lyu, Zenan Huang, Tianlei Hu, Gang Chen, and Haobo Wang. A separation and alignment framework for black-box domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16005–16013, 2024. 6, 7

[55] Rui Xia, Zhenchun Pan, and Feng Xu. Instance weighting for domain adaptation via trading off sample selection bias and variance. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden*, pages 13–19, 2018. 2

[56] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020. 1

[57] Jianfei Yang, Xiangyu Peng, Kai Wang, Zheng Zhu, Jiashi Feng, Lihua Xie, and Yang You. Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors. *arXiv preprint arXiv:2205.14467*, 2022. 1, 3, 6, 7

[58] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 1(2):5, 2020. 1, 3

[59] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019. 1, 3, 5, 6

[60] 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. 1, 3

[61] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. *arXiv preprint arXiv:2101.02839*, 2021. 1, 3

[62] Siqi Zhang, Lu Zhang, and Zhiyong Liu. Refined pseudo labeling for source-free domain adaptive object detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3