Unsupervised Accuracy Estimation of Deep Visual Models using Domain-Adaptive Adversarial Perturbation without Source Samples

JoonHo Lee[1], Jae Oh Woo[2], Hankyu Moon[2] and Kwonho Lee[1]
[1]Samsung SDS, [2]Samsung SDS America
{joonholee,jaeoh.w,hankyu.m,kwonho81.lee}@samsung.com

Abstract

Deploying deep visual models can lead to performance drops due to the discrepancies between source and target distributions. Several approaches leverage labeled source data to estimate target domain accuracy, but accessing labeled source data is often prohibitively difficult due to data confidentiality or resource limitations on serving devices. Our work proposes a new framework to estimate model accuracy on unlabeled target data without access to source data. We investigate the feasibility of using pseudo-labels for accuracy estimation and evolve this idea into adopting recent advances in source-free domain adaptation algorithms. Our approach measures the disagreement rate between the source hypothesis and the target pseudo-labeling function, adapted from the source hypothesis. We mitigate the impact of erroneous pseudo-labels that may arise due to a high ideal joint hypothesis risk by employing adaptive adversarial perturbation on the input of the target model. Our proposed source-free framework effectively addresses the challenging distribution shift scenarios and outperforms existing methods requiring source data and labels for training.

1. Introduction

When deep learning models are deployed for target applications, it is common to encounter a degradation in the accuracy of the models. This degradation is typically caused by a distributional discrepancy between the *source* domain on which models were trained and the *target* domain where they are being applied. To ensure reliable deployment, it is essential to continually monitor the performance of the deployed model on target data. However, since target labels are usually not immediately available, this necessity poses a challenging problem, unsupervised accuracy estimation (UAE) for models.

A few approaches have been developed to predict the target domain accuracy on unlabeled target data assuming that labeled source data can be freely accessed [9, 10, 13]. Nakkiran and Bansal [28] have observed that the disagreement rate between two separately trained source models exhibiting a small test error of ϵ, is remarkably similar to ϵ across a range of models. Subsequent studies have expanded this finding [5, 18]. They discovered that the expected disagreement rate of multiple pairs of models or from iterative self-training with the ensemble can effectively estimate model accuracy when tested on previously unseen data. Despite the noteworthy outcomes, these methods have a crucial limitation: they are solely applicable to in-distribution test data as indicated in [18]. In the most realistic scenarios, source and target distributions are supposed to have discrepancies for various reasons. Insufficient training data is a common cause of generalization issues, which is particularly relevant in scenarios where knowledge from synthetic simulation data is applied in real-world applications, such as autonomous driving and robotic manipulations. Also, distribution shifts can occur when input is distorted during sensing and pre-processing.

Two recent approaches [5, 6] have adopted unsupervised domain adaptation (UDA) methods, specifically, DANN [11] to tackle out-of-distribution UAE. These methods train supplementary models that learn domain-invariant representations for both the source and the target distributions to identify the maximum discrepancy between the adapted hypotheses and the source hypothesis [6] or to employ an ensemble of the adapted hypotheses for iterative self-training [5]. However, focusing on learning domain-invariant representations through DANN can lead to suboptimal outcomes when significant label distribution mismatches occur, increasing the lower bound of the ideal joint hypothesis risk [35]. As a consequence, the estimation performance of these methods is negatively affected.

Furthermore, all existing approaches require access to source samples, except for a few naive techniques like AC [13]. However, obtaining labeled source data, which are used to train the model of interest, is often infeasible in practice due to concerns over the confidentiality of sensitive data or the computing and storage constraints on the serving

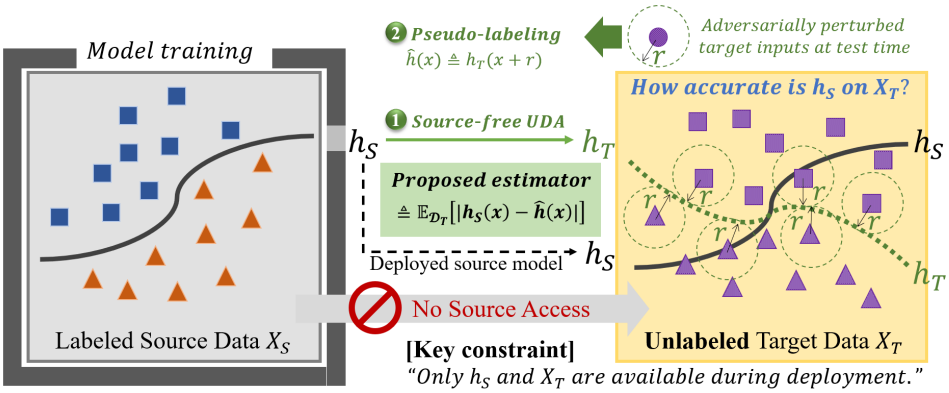


Figure 1. Schematic illustration of the proposed SF-DAP (Source-Free Domain-Adaptive Pseudo-labeling) framework. SF-DAP employs source-free UDA and introduces target-adaptive VAPs to tackle the UAE problem without accessing source samples.

device. Some methods, such as GDE [18], construct multiple source models using source samples in advance, thus eliminating the need for source access. However, training multiple models beforehand to estimate the accuracy of a deployed model is not practical as it is a common practice to train a single optimal model for deployment purposes. To overcome these challenges, we propose a novel framework for estimating the accuracy of models on unlabeled target data without requiring access to source samples. Specifically, we derive a target labeling function that reduces the target risk from a pre-trained source model without using the source data. We then estimate the target accuracy of the models by computing the disagreement rate between the source and the target models under the target distribution.

Initially, we explore the viability of employing straightforward pseudo-labeling strategies [1, 3, 22, 26, 32] that rely solely on target data. While a vanilla UDA method such as DANN requires access to both labeled source and unlabeled target data, recent advances in *source-free* UDA methods [19, 22, 23, 24, 26, 34] remove the need for source data access. Certain source-free UDA methods [22, 25, 26] freeze the head classifier of the source model and aim to train *target-specific* feature generators. These feature generators are trained to learn target features that align with the source distribution of the frozen head classifier under the target distribution. These methods exhibit comparable performance to the state-of-the-art vanilla UDA and align with our goal of developing a *source-free UAE* approach based on pseudo-labeling. To achieve this goal, we introduce the SF-DAP (Source-Free Domain-Adaptive Pseudo-labeling) framework, which incorporates source-free UDA algorithms into the source-free UAE. As the adapted model by UDA algorithms may not always approximate an ideal target labeling function, computing disagreement in a naive manner can lead to a less accurate estimation. To tackle this issue, we develop a systematic method that leverages perturbations, particularly virtual adversarial perturbations (VAPs) [27], to target data during inference. Our domain-adaptive VAPs account for predictive uncertainty and domain discrepancy, thereby mitigating the effects of distribution shifts. Extensive experimental results in various challenging scenarios demonstrate that our proposed method outperforms existing approaches without requiring source data. Our key contributions are summarized as follows:

- We propose SF-DAP, a source-free UAE framework that employs source-free UDA for a viable pseudo-labeling function under target distribution and combines target-adaptive VAPs with it. As far as we know, ours is the first source-free UAE approach that demonstrates comparable performance to source-based ones.
- We illustrate the effectiveness of our proposed framework through extensive experiments on various challenging cross-domain scenarios. Our approach consistently outperforms existing methods, even without labeled source data.
- We present Pseudo-labeling Assignment by Feature Alignment (PAFA) algorithm for source-free UDA, which extends existing methods by introducing modification that improves its efficacy. Empirical results indicate that PAFA matches well with our framework.

2. Preliminaries

2.1. Notation

We use $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ to denote the input, output, and feature representation space, respectively. For simplicity of exposition, we consider binary *hypothesis* $h : \mathcal{X} \to \{0, 1\}$ and h is a composition of *feature generator* $g : \mathcal{X} \mapsto \mathcal{Z}$ and *classifier* $f : \mathcal{Z} \to \{0, 1\}$ i.e., $h = f \circ g$. We define a *domain* as $\langle \mathcal{D}, h^* \rangle$. $\mathcal{D}$ denotes a distribution on input $\mathcal{X}$ and a labeling function h^* is defined by $h^* : \mathcal{X} \to [0, 1]$ (h^* can have a fractional value). The *error* of a hypothesis h w.r.t. the labeling function h^* under $\mathcal{D}$ is defined as $\varepsilon_{\mathcal{D}}(h, h^*) :=$

$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big[|h(\mathbf{x}) - h^*(\mathbf{x})|\big] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\big[\mathbb{1}(h(\mathbf{x}) \neq h^*(\mathbf{x}))\big]$ and reduces to a disagreement probability such that $\varepsilon_{\mathcal{D}}(h, h^*) = \Pr_{\mathbf{x}\sim\mathcal{D}}(h(\mathbf{x}) \neq h^*(\mathbf{x}))$.

$\varepsilon_{\mathcal{D}}(h)$ denotes the *risk* of h that is the error of h w.r.t. the true labeling function under $\mathcal{D}$, i.e., in case h^* is the true labeling function, $\varepsilon_{\mathcal{D}}(h) = \varepsilon_{\mathcal{D}}(h, h^*)$. To distinguish *source* and *target*, we attach subscripts S and T, respectively. *e.g.*, $\mathcal{D}_S$ and $\mathcal{D}_T$ for the source and the target domain. For simplicity, we use $\varepsilon_S := \varepsilon_{\mathcal{D}_S}$ and $\varepsilon_T := \varepsilon_{\mathcal{D}_T}$.

2.2. Unsupervised Accuracy Estimation

One of the early UAE approaches [10] draws on the negative correlation between the distribution discrepancy and model accuracy and builds a regression model between these quantities for UAE. The correlation between the rotation estimation and the classification tasks is observed and used to build a simple regression-based UAE in the same manner, [9]. AC and DoC [13] approach the problem from the prediction confidence. Similarly, ATC [12] utilizes confidence measures after calibration to identify the source data's confidence threshold that matches the source accuracy, which is applied to calculate the target accuracy[14]. The initial UAE approach based on disagreement from random subsets [28] was extended by GDE [18] to random model ensembles. With a flavor of the UDA approach, Proxy Risk [6] explored the maximization of proxy risks, RI, and RM [5] employed the iterative ensemble as well.

Source-Free UAE. We consider the UAE problem where we have m unlabeled target data $\{\mathbf{x}_j\}_{j=1}^m \in \mathcal{X}^m$ and a model (hypothesis) h trained by n labeled source samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$. In contrast to prior methods, we adopt a more practical and widely applicable assumption that the source samples are unavailable, as illustrated in Fig. 1. The goal of *source-free* UAE is to find a function that correctly estimates accuracy, or equivalently, risk of the source model on unlabeled target data under $\mathcal{D}_T$ (denoted by $\varepsilon_T(h_S)$) without source samples. We consider the case where both domains have the same set of labels.

2.3. Unsupervised Domain Adaptation

The goal of UDA is to find a hypothesis h that correctly predicts the label y_j of a new target sample x_j by learning from the labeled source and unlabeled target data. The UDA problem considers that the learning algorithm has access to a set of n labeled points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n \overset{\text{i.i.d.}}{\sim} \langle \mathcal{D}_S, h_S^* \rangle$ and a set of unlabeled points $\{\mathbf{x}_j\}_{j=1}^m \in \mathcal{X}^m \overset{\text{i.i.d.}}{\sim} \langle \mathcal{D}_T, h_T^* \rangle$. Similarly to UAE, we assume both domains share the same label set. DANN [11] has received attention among earlier UDA works due to its approach to learning a domain-invariant representation with theoretical elegance. DANN is also employed by some existing UAE methods.

Source-Free UDA. Source-free UDAs consider more practical limitations of source sample unavailability. Unlike vanilla UDA, source-free UDA addresses the same problem without access to source samples, which aligns well with the nature of source-free UAE. We pay close attention to an earlier work SHOT [26] which proposed a source hypothesis transfer strategy and achieved comparable performance to vanilla UDAs. ISFDA [25] utilize SHOT to tackle class-imbalanced scenarios. FAUST [22] also fixes the head classifier but produces improved performance by enforcing two consistency losses from multiple perturbed views of an input. They typically learn a target-specific feature embedding $z = g(x)$ that can be aligned with the source distribution contained in the frozen (head) classifier f.

2.4. Analysis on Existing DANN-Based Approaches

We examine the theoretical framework associated with the recent approaches [5, 6] that learn domain-invariant representation using DANN [11]. Ben-David *et al.* [2] introduced $\mathcal{H}\triangle\mathcal{H}$-divergence which is determined by the discrepancy between source and target distributions of the hypothesis class $\mathcal{H}$. Given domains $\mathcal{D}_S$ and $\mathcal{D}_T$ over $\mathcal{X}$, the $\mathcal{H}\triangle\mathcal{H}$-divergence, $d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$, is defined by $\sup_{h,h'\in\mathcal{H}} |\varepsilon_S(h, h') - \varepsilon_T(h, h')|$. This divergence leads to

Theorem 1 (Ben-David *et al.*)**.** *For any hypothesis $h \in \mathcal{H}$, let $\lambda = \min_{h'\in\mathcal{H}} \varepsilon_S(h') + \varepsilon_T(h')$ is the risk of ideal joint hypothesis. Then we have*

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda. \tag{1}$$

Inequality 1 implies that no classifier can accurately predict both domains when λ is large, making it impossible to find a high-quality target hypothesis through UDA. Later, Zhao *et al.* [35] presented an information-theoretic lower bound for the risk of ideal joint hypothesis as follows:

$$\varepsilon_S + \varepsilon_T \geq \frac{1}{2}(d_{JS}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{JS}(\mathcal{D}_S^Z, \mathcal{D}_T^Z))^2, \tag{2}$$

where d_{JS}, $\mathcal{D}^Y$, and $\mathcal{D}^Z$ denote Jensen-Shannon distance, label distribution, and feature distribution, respectively.

By analyzing the inequality 1 and the inequality 2, we can identify three drawbacks of current approaches that use DANN. Firstly, $\mathcal{H}\triangle\mathcal{H}$-divergence in the inequality 1 cannot be accurately estimated from finite samples of arbitrary distributions, which limits its practical usefulness. Secondly, as indicated in the inequality 2, relying solely on learning domain-invariant representation can increase the lower bound of the risk of ideal joint hypothesis, which results in unsuccessful UDA and negatively impacts estimation performance. Thirdly, distance terms in the inequality 1 and the

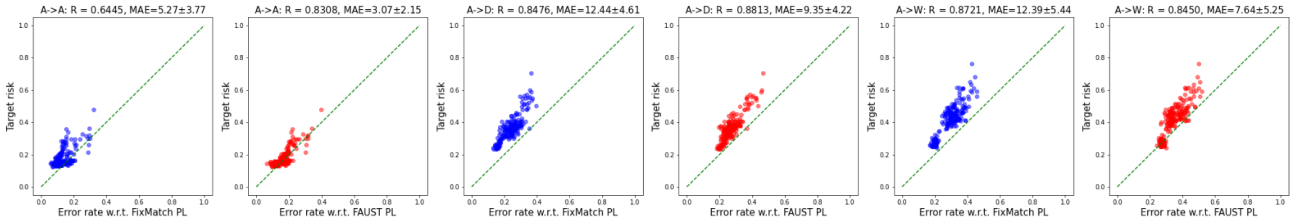


Figure 2. (Best viewed in color) Investigation on Office-31. The scatter plots show the error rate w.r.t. pseudo-labels (x-axis) vs the true target risk (y-axis) of 192 different models trained on Amazon images. The green dashed line denotes the true target risk estimates. Each pair is tested on Amazon (test), DSLR, and Webcam, in that order. Blue and red dots are produced based on [32] and [22], respectively.

inequality 2 require access to source samples, which does not align with the objective of a more practical source-free UAE.

3. Straightforward Pseudo-Labeling Approach

To achieve the goal of a source-free UAE, our focus is on determining how to bound ε_T using terms that can be computed within the target domain. From the definition of risk and triangular inequality, for any hypothesis $h'_S \in \mathcal{H}$, the target risk ε_T of the source hypothesis h_S can be bounded as $\varepsilon_T(h_S) \leq \varepsilon_T(h_S, h'_S) + \varepsilon_T(h'_S)$. Substituting h'_S with a pseudo-labeling function h^{pl}_T on $\mathcal{D}_T$ leads to:

Proposition 1. *Let h^{pl}_T be any pseudo-labeling function under $\mathcal{D}_T$. Then, for given h_S, we have*

$$\varepsilon_T(h_S) \leq \varepsilon_T(h_S, h^{pl}_T) + \varepsilon_T(h^{pl}_T). \tag{3}$$

Proof. By definition, $\varepsilon_T(h_S) := \mathbb{E}_{D_T}[|h_S - h^*_T|]$ where h^*_T denotes true labeling function under $\mathcal{D}_T$. Then, $\mathbb{E}_{D_T}[|h_S - h^*_T|] = \mathbb{E}_{D_T}[|h_S - h^{pl}_T + h^{pl}_T - h^*_T|] \leq \mathbb{E}_{D_T}[|h_S - h^{pl}_T|] + E_{D_T}[|h^{pl}_T - h^*_T|]$ from triangular inequality. Hence, $\varepsilon_T(h_S) \leq \varepsilon_T(h_S, h^{pl}_T) + \varepsilon_T(h^{pl}_T)$. □

Proposition 1 suggests that we can estimate the target accuracy of h_S by identifying a suitable pseudo-labeling function for the target samples. This approach transforms our source-free UAE problem into the task of finding an effective pseudo-labeling function under $\mathcal{D}_T$ that reduces the estimation error. It is worth noting that Proposition 1 is different from the inequality 1 or the inequality 2 in that the right-hand side can be computed only using the unlabeled target data in Proposition 1.

Based on Proposition 1, we investigate how well the disagreement between the source model's outputs h_S and pseudo-labels on target data h^{pl}_T matches true risk through experiments. Among the previously proposed pseudo-labeling strategies [1, 3, 22, 32], we select (1) FixMatch [32] approach that computes an artificial label by direct inference of the source model h_S and (2) weighted feature prototype-based approach such as [1, 3, 22]. For the first approach, we obtain the model's predicted class distribution q given a weakly perturbed view of an input, x^w, then use $h^{pl} := \arg\max(q(x^w))$ as a pseudo-label, which is compared with the model's output for a strongly perturbed view of the input, $q(x^s)$, as suggested by [32]. We also follow the approach of [22] that obtains the feature prototype of each class k by $c^k := \sum_{i \in \mathcal{B}} q^k(x_i) g(x_i)$ where x_i is drawn from the empirical target data (in each mini-batch $\mathcal{B}$) and $q^k(x_i)$ is the prediction probability of x_i to class k. Then, the soft pseudo-label of a target sample is defined by

$$h^{pl} := \text{softmax}(C^T g(x)) \tag{4}$$

where columns of the matrix C are the prototypes.

Experiment. We investigate on Office-31 dataset which contains 4,652 ImageNet[8]-like images across 31 objects collected from three different domains: Amazon (A), DSLR (D), and Webcam (W). We split the Amazon set into a development set (90%) and a holdout set (10%), then train 192 different ResNet50 source models by sharing the development set but different training options such as learning rate, input augmentation, optimizer, and label smoothing. More details are described in Appendix A.Webcam, DSLR, and Amazon holdout images are assigned as target data on which we evaluate the pseudo-label based estimators' performance. The pseudo-labeling from weighted feature prototypes seems more noticeable as presented using red dots in Fig. 2. For the in-distribution setting (A→A), the mean absolute error (MAE) of the estimator is 3.07±2.15. MAEs in the out-of-distribution settings (A→D, A→W) are no more than 9.35. Their Pearson's correlation coefficients are all above 0.83, which is close to the ideal correlation coefficient, 1.0.

4. Proposed Method

The observations presented in Sec 3 corroborate our notion that the accuracy estimation performance can be enhanced by devising a pseudo-labeling function that more closely approximates the true labeling function under the target distribution. This concept logically leads to the application of source-free UDA to generate an appropriate

pseudo-labeling function since source-free UDA aims to adapt the source model to enhance prediction accuracy on unlabeled target data without access to source data.

Let h_T be a target model adapted to target distribution from the source model h_S via a source-free UDA algorithm and therefore h_T can be regarded as a pseudo-labeling function that has relatively less target risk. By replacing h_T^{PL} with h_T, for any (adapted) target hypothesis $h_T \in \mathcal{H}$, Proposition 1 comes to $\varepsilon_T(h_S) \leq \varepsilon_T(h_S, h_T) + \varepsilon_T(h_T)$. This inequality indicates that the target risk of the source model $\varepsilon_T(h_S)$ becomes closer to the disagreement between source and target models under $\mathcal{D}_T$ as the source-free UDA algorithm improves, i.e. the target risk of the adapted model $\varepsilon_T(h_T)$ becomes less.

However, UDA cannot fully approximate a true labeling function on $\mathcal{D}_T$ in general due to limitations imposed by the classification problem itself as acknowledged by [6]. To tackle this limitation, we propose a target-adaptive pseudo-labeling technique that combines the pseudo-labeling function of each target sample with the inference over its ϵ-neighbors that maximally disagrees with its own output. Then its disagreement with h_S naturally estimates the desired target risk. Based on these ideas, we introduce a novel source-free UAE framework called **SF-DAP** (Source-Free Domain-Adaptive Pseudo-labeling), which comprises source-free UDA and target-adaptive estimation.

4.1. Source-Free UDA Selection

For the selection of the source-free UDA algorithm, the empirical results in Sec 3 allow us to consider two factors: (1) preservation of the ample information inherent in the classifier f delivered from the source domain, which can confer an advantage in a source-free scenario, and (2) alignment of the target feature representation with the source distribution, as a proxy for successful adaptation, since pseudo-labeling in the feature space has demonstrated more favorable performance.

SHOT [26] freezes the head classifier f and focuses on training the feature generator g, which satisfies the aforementioned criteria. Its simpler variant, SHOT-IM, trains the feature generator g using an information maximization objective that combines entropy minimization loss and mean entropy maximization loss. However, SHOT requires augmentation of the source network prior to training (such as a bottleneck layer and two normalization layers), which does not align with our scenario of estimating the performance of an already trained source model. Therefore, we introduce a modified version of SHOT in this work.

Instead of augmenting the network, we adopt a self-training approach that incorporates two input views into the training objective of SHOT-IM. This approach is motivated by previous works [1, 22, 32] and aims to alleviate the negative impact of not meeting network augmentation requirements. Specifically, we generate pseudo-labels based on the nearest feature prototype classifier defined in Eq. 4 using the weakly perturbed view, while the strongly perturbed view is utilized for forward pass prediction. We refer to the proposed source-free UDA training objective as PAFA (Pseudo-label Assignment by Feature Alignment) and formulate it as follows:

$$\min_g \; \mathcal{H}(h(x^w)) + D_{\mathrm{KL}}\left(\bar{h}(x^w) \middle\| \frac{1}{K}\mathbf{1}_K\right) - \log K + \alpha\, \mathcal{H}(h^{pl}(x^w), h(x^s)). \tag{5}$$

Here, x^w and x^s denote weakly and strongly perturbed views of the input x, which are generated by standard flip-crop-rotation augmentation and by RandAugment [7], respectively. The function $\mathcal{H}(\cdot)$ represents the Shannon entropy, while $\mathcal{H}(\cdot,\cdot)$ denotes cross-entropy. The term $\bar{h}(x)$ refers to the mean embedding of K-dimensional prediction outputs under the target distribution, and $\mathbf{1}_K$ is a K-dimensional vector with all ones. The positive constant α is a scaling factor for the self-training gradient. We note that the $\frac{1}{K}\mathbf{1}_K$ term assumes that the target label distribution is uniform as SHOT [26] did. This assumption is reasonable since we lack any prior knowledge about the target label distribution. Therefore, we also assume a uniform label distribution as a non-informative prior [17]. Nevertheless, our proposed method, PAFA, performs well even with non-uniform target label distributions as demonstrated in Sec 5. The UDA performance of PAFA is summarized in Sec 6. Additionally, we evaluate some other source-free UDA methods within the proposed framework and show there are no significant difference between their results (see Appendix C.3).

4.2. Enhancing the Quality of Estimation

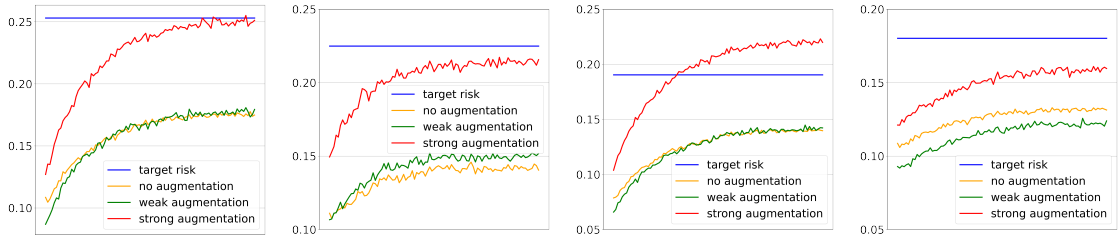
An adapted model obtained through UDA, as presented in Sec 4.1, can serve as a target labeling function in constructing the estimator using the following formulation:

$$\mathrm{EST}_{\mathrm{naive}} \coloneqq \mathbb{E}_{x\sim\mathcal{D}_T}[\mathbb{1}(h_S(x) \neq h_T(x))]. \tag{6}$$

Though the adapted model has reduced target risk when compared to the source model, it may not accurately approximate the true labeling function under a severe distribution shift. Consequently, computing the disagreement between the source model h_S and the adapted target model h_T in a straightforward manner like Eq. 6 often result in less accurate estimation. Therefore, we need to explore methods to enhance the quality of estimation.

4.2.1 Random Perturbation (RND)

While evaluating the effectiveness of $\mathrm{EST}_{\mathrm{naive}}$ using CIFAR-10 [20] and CIFAR-10-C [15] benchmarks, we ob-



(a) Fog (b) Contrast (c) Gaussian blur (d) Saturate

Figure 3. (Best viewed in color) The accuracy trends from different test-time perturbations as UDA progresses on CIFAR-10-C tasks. The plot consists of the UDA training iteration (x-axis) vs. the target risk (y-axis). The blue line denotes the true target risk of h_S, and the other lines denote the estimated target risk of h_S using the disagreement rate $\mathbb{E}_{x\sim\mathcal{D}_T}[|h_S(x) - h_T(x)|]$ where the input data x are augmented in different fashions. In each red, green, and orange line, inputs are strongly augmented, weakly augmented, and not augmented, respectively.

served that applying a test-time perturbation to target samples before computing disagreement can improve the quality of estimation. This observation is depicted in Fig. 3, where we can see that the estimation quality often increases with the strength of the perturbation. Our finding suggests that perturbing the input to h_T can reduce the estimation error, particularly when stronger perturbations (indicated by red lines in Fig. 3) are applied, resulting in improved estimation. We can express this estimator as follows:

$$\text{EST}_{\text{rnd}} \coloneqq \mathbb{E}_{x\sim\mathcal{D}_T}[\mathbb{1}(h_S(x) \neq h_T(x^r))], \tag{7}$$

where x^r denotes a randomly perturbed view of the input. Though generating multiple perturbed views and ensembling inferences of them can further improve estimation performance (see Appendix C.5), we have opted to present only a single perturbation for the sake of simplicity. As depicted in Fig. 3c, random perturbations often lead to improved performance but may sometimes underestimates.

4.2.2 Adversarial Perturbation (ADV)

Sec 4.2.1 suggests that adding random perturbations to the inputs of the target model can frequently enhance the estimation performance, but this is not always the case. To ensure appropriate randomness during estimation, we consider the virtual adversarial perturbation (VAP) [27], which can adapt to the target distribution. In addition to enforcing local smoothness of the conditional label distribution given unseen target inputs [27], VAP promotes the discovery of ϵ-neighbors for each target sample that maximally disagrees with its own output. The disagreement between the pseudo-label function and the source model output naturally estimates the target risk. As an illustrative example: if a target sample resides far away from the decision boundaries, its pseudo-label will mostly agree with its own. However, if the sample is close to the decision boundaries, its adversarially estimated pseudo-label may indicate a different class label. Hence, we define the accuracy estimator as follows:

$$\text{EST}_{\text{adv}} \coloneqq \mathbb{E}_{x\sim\mathcal{D}_T}\left[\mathbb{1}(h_S(x) \neq h_T(x + r_{vadv}))\right]. \tag{8}$$

The virtual adversarial perturbation r_{vadv} is computed as

$$r_{vadv} = \underset{\delta \ s.t. \ \|\delta\|_2\leq\epsilon}{\arg\max} \ D_{\text{KL}}\Big(h_T(x) \parallel h_T(x+\delta)\Big), \tag{9}$$

where ϵ is a hyperparameter that determines the magnitude of the adversarial perturbation. We note that VAP is computed during inference through the frozen target model. SF-DAP (ADV) comprises PAFA and EST_{adv}.

4.2.3 Adaptive Adversarial Perturbation (AAP)

By employing Eq. 8, we achieved estimation performance comparable to that of existing methods that require access to source samples, as presented in Sec 5. Nonetheless, we identified certain concerns, including instances where the estimation errors were unacceptably large in specific scenarios. The proposed method is able to address this issue by adjusting the sole hyperparameter ϵ. This section outlines how to quantify the factors influencing the magnitude of ϵ and concludes with an improved formulation.

Predictive Uncertainty. According to the cluster assumption [4], data points with low predictive uncertainty are typically located far from the class decision boundary, resulting in more accurate model predictions. Conversely, data points with high predictive uncertainty often reside in the low-density regions of the feature space, near or overlapping the class decision boundary, leading to less accurate model predictions. Therefore, to enhance estimation performance, we suggest reducing perturbation in data points with low uncertainty while increasing perturbation in those with high uncertainty. This can be achieved by scaling the perturbation magnitude based on the predictive uncertainty measurement. We define the uncertainty factor C_{unc} as the standard deviation (std) of the class probabilities predicted by the target model that ranges within 0.5. To evaluate uncertainty, we apply Monte-Carlo dropout inference sampling n-times (see Appendix C.4 for more details):

$$C_{unc} = \text{std}\left\{q_i^{(k)}\right\}_{i=1}^{n}, \tag{10}$$

Table 1. Overall benchmark results on various UAE tasks. Average mean absolute errors (MAEs, %) of each group and overall averages are reported. **Bold** numbers indicate the superior results and ***bold-italic*** do the next best results. Full results are presented in Appendix A.6.

			source access approach					source-free approach (ours)	
datasets	settings	subset description	DoC [13]	Proxy [6]	RI [5]	RM [5]	GDE [18]	SF-DAP (ADV)	SF-DAP (AAP)
Digits	6	MNIST, USPS, SVHN	14.62±0.80	9.63±1.44	9.50±1.51	11.33±1.36	27.80±0.75	***2.18±0.56***	**2.15±0.58**
Office-31	6	Amazon, DSLR, Webcam	5.00±1.26	4.64±1.23	8.69±1.62	***2.73±1.28***	9.79±1.10	4.72±1.13	**2.51±1.10**
Office-Home	12	Art, Clipart, Product, Real-World	25.62±0.50	8.75±1.41	12.13±1.58	**3.66±1.25**	37.52±0.51	8.73±0.65	***4.98±0.68***
VisDA	1	Synthetic images, MSCOCO	15.72±4.35	8.90±1.71	7.50±4.70	***4.41±2.52***	29.31±3.73	***4.41±1.10***	**1.73±0.93**
CIFAR-10	19	CIFAR-10 and CIFAR-10-C subsets	18.91±1.56	10.47±0.83	3.20±0.94	**2.17±1.18**	2.99±1.24	***2.34±1.47***	3.06±1.46
CIFAR-100	19	CIFAR-100 and CIFAR-100-C subsets	48.64±1.99	26.34±2.46	5.84±1.67	**1.90±0.84**	6.33±1.75	6.81±1.59	***3.28±1.35***
micro average (of all 63 settings)			27.37±1.50	14.27±1.67	6.89±1.46	**3.33±1.14**	14.00±1.29	5.15±1.29	**3.33±1.20**
macro average (of the above 6 averages)			21.42±1.32	11.45±1.23	7.81±1.42	***4.37±1.18***	18.96±1.23	4.86±1.04	**2.95±1.01**

where $q_i^{(k)}$ is the probability of the k-th class in the predicted class distribution q obtained from the i-th Monte-Carlo dropout sample. The class index k is determined by $\arg\max_{j\in\{1,\cdots,K\}} \left(\frac{1}{n}\sum_{i=1}^{n} q_i\right)^{(j)}$.

Domain Divergence Though the C_{unc} incorporates the cluster assumption into individual data points, bias may exist due to inaccurate predictions stemming from the distribution gap between the source and the target. To tackle this bias, we consider the divergence between both distributions. We cannot access the source dataset in the source-free scenarios, so we compute the divergence from the individual data point perspective instead of focusing on datasets. Hence, we define the divergence factor C_{div} using Jensen-Shannon Divergence (D_{JS}) between $h_S(x_t)$ and $h_T(x_t)$:

$$C_{div} = D_{\text{JS}}(h_S(x_t) \| h_T(x_t)). \tag{11}$$

By selecting base-2 logarithm, the D_{JS} value is bounded between 0 and 1, being able to provide normalized adjustment to alleviate bias. Adding this factor to C_{unc} yields an *adjusted predictive uncertainty* factor C_{adj_unc}.

Data Volume Density The density of data points in a dataset is often a crucial factor in determining the magnitude of the VAP. For instance, images that are synthetically contrast reduced have an increased color space density, while natural images typically have a much lower density than synthetic images. We can enhance the estimation performance by adjusting the magnitude of VAP in proportion to the relative density of the data points. To this end, we consider the standard deviations calculated for each axis of the three-dimensional RGB color space that can be easily computed from the image dataset. For the inaccessible source dataset, these values are still commonly available in the definition of data transform used for image normalization during training. The natural adjustment approach scales the perturbation magnitude by the relative density of the target data to that of the source data. For consistent application to both black-and-white and color images, we define the density scaling factor C_{den} as follows:

$$C_{den} = \frac{1}{3}\left(\frac{std_t^R}{std_s^R} + \frac{std_t^G}{std_s^G} + \frac{std_t^B}{std_s^B}\right), \tag{12}$$

where std_s^R and std_t^G denote the standard deviation of the source dataset in the red axis and that of the target dataset in the green axis, respectively (B denotes blue).

Class Complexity We also consider the number of classes (K) that affects the scale of the VAP. KL divergence tends to increase as K increases, as the volume of the probability space grows exponentially with K. While the behavior depends on specific distributions, our empirical analysis has shown that compensating for the increase of the KL divergence in Eq. 8 by $\log K$ improves the accuracy. Based on this observation, we introduce the class complexity factor C_{cls}:

$$C_{cls} = \log\left(\text{the number of categories}\right). \tag{13}$$

Proposed Accuracy Estimator. Considering the factors mentioned above, we propose EST_{aap} that employs adaptive adversarial perturbations (AAPs). EST_{aap} is, in fact, a modification of EST_{adv} in Eq. 8, with the only difference being the definition of hyperparameter ϵ as:

$$\epsilon \coloneqq \epsilon_0 C_{cls} C_{den} C_{adj\_unc}. \tag{14}$$

Here, the base constant ϵ_0 is typically set to 1.0 as suggested by [27]. According to Eq. 14, ϵ is re-scaled by the adjusted predictive uncertainty (C_{adj_unc}) that adaptively controls the perturbation on individual data points. Then the updated ϵ is multiplied by the class complexity (C_{cls}) of the target dataset and by the relative data point density of the target dataset compared to the source dataset (C_{den}). Thus, PAFA and EST_{aap} collaboratively compose SF-DAP (AAP).

5. Experiment

Setup. We evaluate and report absolute estimation error to measure the performance of UAE. For natural distribution shift, we experiment on Digits [16, 21, 29], Office-31

Table 2. Ablation study on Office-31. Uniformly scaling ϵ for all data instances within the dataset can hurt performance (Configuration 0). With instance-wise scaling (C_{adj_unc}), incorporation of scaling factors gradually improves accuracy estimation overall.

Method	VAP magnitude (ϵ)	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
SF-DAP (ADV)	ϵ_0(=1.0)	1.85±1.22	4.43±2.10	9.05±1.78	2.97±0.52	9.56±1.77	0.46±0.24	4.72±1.13
Configuration 0	$\epsilon_0 C_{den} C_{cls}$	9.39±4.01	5.64±5.20	5.24±3.02	7.64±2.75	**2.75±2.53**	3.90±1.47	5.76±1.78
Configuration 1	$\epsilon_0 C_{adj\_unc}$	3.33±1.12	1.85±1.14	10.47±1.92	**0.34±0.28**	11.98±1.80	0.22±0.14	4.70±1.03
Configuration 2	$\epsilon_0 C_{den} C_{adj\_unc}$	4.70±1.13	2.15±1.39	8.20±1.93	0.57±0.27	10.82±1.58	**0.16±0.15**	4.43±1.04
Configuration 3	$\epsilon_0 C_{cls} C_{adj\_unc}$	1.53±0.76	2.00±1.09	4.84±1.82	2.63±0.71	6.03±1.99	1.25±0.56	3.05±1.07
SF-DAP (AAP)	$\epsilon_0 C_{cls} C_{den} C_{adj\_unc}$	**0.96±0.87**	**1.62±0.98**	**3.28±1.97**	3.08±0.89	5.23±2.03	0.88±0.49	**2.51±1.10**

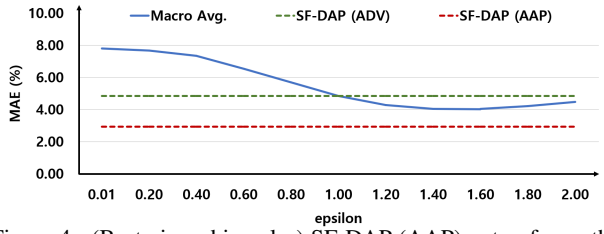


Figure 4. (Best viewed in color) SF-DAP (AAP) outperforms the optimal result (at $\epsilon \approx 1.5$) achieved by uniform ϵ scaling.

[31], Office-Home [33], and VisDA [30] datasets referring to the UDA literature. We note that a few challenging scenarios are included, such as MNIST→SVHN and VisDA sim-to-real. For synthetic distribution shift setting, we use CIFAR-10 and CIFAR-100 [20], paired with CIFAR-10-C and CIFAR-100-C[15], respectively. The CIFAR-10-C and CIFAR-100-C datasets contain 19 distinct types of corruptions applied to CIFAR-10 and CIFAR-100 images, respectively. We employed an identical backbone for each scenario across all methods for a fair comparison. To be specific, we use a LeNet variant for Digits, ResNet18 for CIFAR-10 and CIFAR-100, ResNet50 for Office-31 and Office-Home, and ResNet101 for VisDA. More detailed experimental setups, including datasets, network architectures, and training options, are described in Appendix A.

Evaluation. As our baselines for comparison, we consider Difference of Confidence (DoC) [13], Proxy Risk (Proxy) [6], Random Initialization (RI) [5], Representation Matching (RM) [5], and Generalization Disagreement Equality (GDE) [18]. We re-implement all baselines and compare them under the same experimental setup. We repeat experiments on ten independently-trained source models and report the mean and standard deviation of the computed MAE. Evaluation protocols of the baselines and their details are described in Appendix A.5 and Appendix B, respectively.

Results. As shown in Table 1, our proposed framework outperforms existing methods in 63 cross-domain scenarios across six benchmark groups, even without the need to access source samples. The proposed SF-DAP (AAP) demonstrates comparable or superior results to the state-of-the-art

Table 3. Comparison of PAFA with SHOT-IM [26] and FAUST [22] on the six UDA benchmark tasks. *Source-only* denotes the target accuracy of the source model without applying UDA. (S: SVHN, M: MNIST, U:USPS, Syn: SYNSIG, G:GTSRB)

Method	S→M	M→S	M→U	U→M	Syn→G	VisDA	Avg.
Source only	70.0	47.8	77.4	84.9	77.1	56.6	69.0
PAFA	**99.5**	83.5	97.8	**98.5**	99.4	83.8	93.8
SHOT-IM	98.5	11.0	97.8	97.6	97.2	82.0	80.7
FAUST	99.2	**88.0**	**98.9**	97.6	**99.7**	**84.0**	94.6

methods like RM and RI. Despite its simpler approach, SF-DAP (ADV) also exhibits comparable or often better performance than previous approaches that require source sample access. While some existing methods such as GDE [18], RI, and RM [5] show strong results on synthetic distribution shift scenarios, they have limitations under natural distribution shift settings. We emphasize that our proposed framework requires only unlabeled target data, whereas all these baseline methods cannot be applied if additional access to source samples is prohibited. All detailed results of 63 scenarios are presented in Appendix A.6.

6. Analysis

Ablation. We conduct an ablation study on the Office-31 dataset by incrementally introducing scaling factors of SF-DAP (AAP), starting from SF-DAP (ADV). Configuration 0 in Table 2 indicates that including the adjusted uncertainty factor (C_{adj_unc}) is critical without prior knowledge of ϵ_0. Given C_{adj_unc}, both the class complexity (C_{cls}) and the data volume density (C_{den}) factors contribute substantially to enhancing the estimation performance. As a consequence, SF-DAP (AAP) surpass other configurations. Additional ablation studies can be found in Appendix C.1.

MAE on a Smooth ϵ-Axis. We investigate the MAE trajectory over a smooth ϵ-axis as depicted in Fig. 4, with the blue line representing the macro average MAE across all 63 scenarios. Fig. 4 suggests that an optimal estimation can be realized at $\epsilon \approx 1.5$ under a uniform VAP scaling constraint applied to all target datasets. Though SF-DAP (ADV) employing $\epsilon = 1.0$ (green dashed line) might be sub-optimal, SF-DAP (AAP) – illustrated by the red dashed line – further diminishes MAE beyond the optimal MAE at $\epsilon \approx 1.5$. This reduction is achieved by introducing C_{adj_unc} which

enables adaptive VAP scaling per target instance, contrasting with the fixed scaling of C_{cls} and C_{den}.

Applying AAP Directly to the Source Model. In the application of AAP to the source model (without UDA) for Digits and Office-31 datasets, a notable 3.0% increase is observed in average MAE compared to our SF-DAP (AAP) that incorporates UDA. This result underlines the crucial role of the source-free UDA element within the proposed SF-DAP framework.

Estimation Time. We compare the runtime of various methods that require additional training. Our method shows a competitive running time to other methods. In particular, the proposed SF-DAP framework runs faster than RM and Proxy Risk, which also perform UDA during estimation (see Appendix C.2 for details).

PAFA Performance in UDA. To contribute to the proposed SF-DAP framework as a source-free UDA algorithm, it is important for PAFA to produce a comparable performance to the existing state-of-the-art methods like SHOT-IM [26] and FAUST [22]. We evaluated PAFA's performance on six commonly used UDA benchmarks and found that it achieved comparable results to SHOT-IM and FAUST as shown in Table 3. Notably, when adapting from an MNIST (source) model to the SVHN (target) dataset, PAFA achieved a target risk of less than 20%, which is the state-of-the-art result in UDA literature.

UAE Performance during UDA. We track the performance trend of accuracy estimation as the proposed source-free UDA, PAFA, progresses to each target domain. Our proposed framework, SF-DAP, starts producing accurate estimates surprisingly early and remains steady throughout the rest of the UDA iterations as illustrated in Appendix C.6.

7. Conclusion

In this work, we proposed a novel framework that estimates model accuracy under unseen target distributions without access to source data. Our approach utilized source-free UDA algorithms to produce a viable pseudo-labeling function and employ virtual adversarial perturbations to enhance the accuracy estimation. We also developed a simple yet effective source-free UDA algorithm, PAFA, which achieved comparable performance to existing methods on six popular UDA benchmarks, while avoiding the need for network augmentation of the source model. To further improve the estimation quality, we introduced adaptive adversarial perturbations based on adjusted predictive uncertainty and domain discrepancy information. Our experimental evaluation demonstrated that both AAP and ADV configurations in our proposed framework outperformed existing methods. To the best of our knowledge, this is the first study to address source-free scenarios in the UAE literature.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *ICCV*, pages 8443–8452, October 2021. 2, 4, 5

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. 3

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2, 4

[4] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 57–64. PMLR, 06–08 Jan 2005. 6

[5] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, volume 34, pages 14980–14992. Curran Associates, Inc., 2021. 1, 3, 7, 8

[6] Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. *ICML*, 2020. 1, 3, 5, 7, 8

[7] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020. 5

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. A large-scale hierarchical image database. In *CVPR*, 2009. 4

[9] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *ICML*. 1, 3

[10] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *CVPR*, pages 15069–15078, June 2021. 1, 3

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016. 1, 3

[12] Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *ICLR*, 2022. 3

[13] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *ICCV*, pages 1134–1144, October 2021. 1, 3, 7, 8

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. 3

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 5, 8

[16] J. J. Hull. A database for handwritten text recognition research. *PAMI*, pages 550–554, 1994. 7

[17] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946. 5

[18] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *ICLR*, 2022. 1, 2, 3, 7, 8

[19] 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. 2

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 5, 8

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient based learning applied to document recognition. volume 86(11), pages 2278–2324, 1998. 7

[22] JoonHo Lee and Gyemin Lee. Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation. *Neural Networks*, 161:682–692, 2023. 2, 3, 4, 5, 8, 9

[23] Jingjing Li, Zhekai Du, Lei Zhu, Zhengming Ding, Ke Lu, and Heng Tao Shen. Divergence-agnostic unsupervised domain adaptation by adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8196–8211, 2022. 2

[24] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, June 2020. 2

[25] Xinhao Li, Jingjing Li, Lei Zhu, Guoqing Wang, and Zi Huang. Imbalanced source-free domain adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 3330–3339, New York, NY, USA, 2021. Association for Computing Machinery. 2, 3

[26] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pages 6028–6039, July 13–18 2020. 2, 3, 5, 8, 9

[27] T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019. 2, 6, 7

[28] Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *CoRR*, abs/2009.08092, 2020. 1, 3

[29] Y. Netzer, T.Wang, A. Coates, A. Bissacco, B.Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. volume 2011, page 5, 2011. 7

[30] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko. Visda: The visual domain adaptation challenge. In *CVPR*, 2016. 8

[31] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domain. In *ECCV*, 2010. 8

[32] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *NeurIPS*, volume 33, pages 596–608. Curran Associates, Inc., 2020. 2, 4, 5

[33] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 8

[34] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *ICCV*, pages 9010–9019, October 2021. 2

[35] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 7523–7532. PMLR, 09–15 Jun 2019. 1, 3