Bridging Theory and Practice in Source-Free Domain Adaptation via Adversarial Proxy Perturbation

Dexuan Zhang[1✉], Qianyu Zhou[1✉], Thomas Westfechtel[1,2], Yusuke Mukuta[1], and Tatsuya Harada[1,3]

[1] The University of Tokyo
[2] Tohoku University
[3] RIKEN
{dexuan.zhang,q-zhou,thomas,mukuta,harada}@mi.t.u-tokyo.ac.jp

Abstract. Domain adaptation addresses the distributional shifts between source and target domains. Given increasing data privacy constraints, Source-Free Domain Adaptation (SFDA) has attracted growing interest, as it precludes access to raw source data during adaptation. While existing SFDA methods predominantly rely on source-like data generation or heuristic pseudo-labeling, their connection to target error generalization is often unclear. In this work, we first present a theoretical analysis by establishing an upper bound on the target error that overcomes explicit dependence on the classical joint-error term. Guided by this analysis, we propose an optimization principle based on Adversarial Proxy Perturbation (APP), serving as a practical bridge between the derived bound and empirical training. We further extend the framework to multi-source scenarios and enable the integration of zero-shot priors from vision-language models (e.g., CLIP) through mutual distillation. Extensive experiments on standard benchmarks demonstrate consistent improvements over state-of-the-art methods in both unimodal and multimodal settings, validating the proposed theoretical framework in practice.

Keywords: Source-Free Domain Adaptation · Generalization Analysis · Adversarial Perturbation · Multi-Source Integration

1 Introduction

In many real-world scenarios, models trained on well-annotated source domains often fail to generalize to new target environments due to distribution shifts. To address this challenge, Unsupervised Domain Adaptation (UDA) [2] has achieved notable success across diverse machine learning and computer vision tasks, including image recognition [34,51,55], semantic segmentation [73], and object detection [29]. Representative approaches—such as adversarial learning [12, 35, 50, 65, 69, 72],

✉ Corresponding authors.

self-supervised learning [4,20,37,58], and self-training [59,76]—have demonstrated substantial effectiveness in reducing cross-domain discrepancies. However, as machine learning systems are increasingly deployed in privacy-sensitive environments, access to raw source data is often restricted. Regulations such as the General Data Protection Regulation (GDPR) directly challenge conventional UDA methods that assume full source data availability during adaptation. These constraints motivate Source-Free Domain Adaptation (SFDA), where only a pre-trained source model—rather than source data—is available for target adaptation. SFDA preserves privacy and reduces communication overhead, making it particularly relevant in applications such as medical imaging and surveillance.

Despite promising empirical progress, much of the existing SFDA literature [4, 20, 26, 31, 32, 52, 53, 62, 68] relies primarily on heuristic objectives. Meanwhile, classical UDA theories [1], which typically assume access to both source and target data for learning domain-invariant representations, are not directly applicable in the source-free setting. This gap raises a fundamental question: how can we analyze and guide SFDA from a target-error generalization perspective without access to source data?

In this work, we first present a theoretical analysis tailored to SFDA by deriving an upper bound on the target error that overcomes explicit dependence on the classical joint-error term. The bound is established under a finite-sample realizability assumption and provides a principled objective for optimization. Guided by this analysis, we propose Adversarial Proxy Perturbation (APP), which operationalizes the bound through adversarial optimization and enables adaptive refinement of target hypothesis beyond fixed source classifiers [31]. Furthermore, we extend the framework to multi-source scenarios and show that it naturally accommodates zero-shot priors from vision-language models (e.g., CLIP) via mutual distillation, enabling knowledge integration across heterogeneous pre-trained models. Our primary contributions are summarized as follows:

- ***Learning Theory*:** We establish a target error upper bound for SFDA under a finite realizability assumption. Unlike classical bounds relying on small joint-error assumptions, our analysis decouples target realizability from source–target label conflict, thus making it more suitable for large domain shifts.
- ***Algorithmic Bridging*:** We propose Adversarial Proxy Perturbation (APP), which operationalizes the derived bound through a proxy distribution and yields a tractable adversarial optimization for refining the target hypothesis.
- ***Multi-source Extension*:** We generalize the framwork to incorporate multiple source models, including Vision-Language Models (VLMs), and establish a co-training scheme via mutual distillation.
- ***Empirical Validation*:** Extensive experiments on standard benchmarks demonstrate consistent improvements over state-of-the-art methods in both unimodal and multimodal settings.

2 Related Works

As described in [67], research in Source-Free Domain Adaptation (SFDA) is generally categorized into model-centric and data-centric methods.

Model-centric methods typically involve fine-tuning a pre-trained encoder while keeping the source classifier fixed, often utilizing self-supervision. Early advances, such as Source Hypothesis Transfer (SHOT) [31], demonstrated that a frozen classifier from the source model can serve as a potent hypothesis for adaptation. In this framework, the target-specific encoder is optimized via information maximization and self-supervised pseudo-labeling. Building on this foundation, subsequent methods introduced mechanisms to mitigate pseudo-label noise, capture neighborhood consistency, or integrate external supervision. Specifically, these approaches utilize pre-trained source models to generate auxiliary signals for feature alignment, including multi-hypothesis classifiers [24, 60], prototypes [18, 46, 47], memory-banks [4, 22], clusters [26, 61–63, 70], and source distribution estimations [9]. Recently, researchers have explored multimodal foundation models to transcend the limitations of source-only knowledge. For example, the DIFO [52] framework distills task-specific information from VLMs (e.g., CLIP [48]), alternating between customizing the multimodal model via prompt learning and transferring its knowledge to the target model. While these approaches successfully align target features with fixed classifier weights, they generally suffer from two issues: *i)* the instability of self-supervision; as observed in training trends, the model may achieve high accuracy in early epochs but eventually converges to a lower performance due to the degrading quality of pseudo-labels; and *ii)* architectural constraints; since alignment relies on the similarity between target features and the source classifier, the framework often necessitates the use of normalized vectors.

Data-centric methods explicitly align a "pseudo-source" domain with the target domain [6, 11, 19, 28, 54], treating SFDA as a specialized case of UDA. This alignment is frequently achieved by reconstructing the source domain through generative models. However, such generative pipelines are often computationally intensive and struggle to synthesize highly structured domains. Furthermore, recovering source-like imagery may inadvertently violate the core privacy preservation principles of the SFDA setting.

Theoretical Foundations in recent work [75] revisited the target risk upper bound, defined by source error on reliable pseudo-labeled target data and intra-domain divergence. However, the target data splitting remains heuristic, and the theory follows the classical framework of [1], which is often insufficient for large domain shift scenarios due to an unbounded joint-error term. Similarly, while [41] investigated the insights of existing SFDA algorithms through discriminability and diversity losses, the resulting theorem introduced intractable constraints, leaving a significant gap between theory and practice. Motivated by these limitations, we propose a simple yet effective learning theory for SFDA that bypasses the intractable joint-error, which is known to degrade adaptation performance in large domain shift tasks [65, 66, 71]. Our theory is designed to be smoothly optimized

in practical settings, directly bridging the gap between theoretical bounds and actual algorithmic implementation.

3 Learning Theory for SFDA

In this section, we establish a target error upper bound based on the deviation between true and approximated labeling functions. We address the intractability of the true function by introducing a finite realizability assumption that holds across varying degrees of domain shift. Finally, we clarify the gap between the theory and practice and discuss constraints on the hypothesis space to mitigate potential failures when updating the target classifier in the source-free setting.

3.1 Theoretical Analysis and Insights

We formulate SFDA as a multi-class classification task on a target domain T. The learning algorithm is provided with a model $f_S^* \in \mathcal{H} : \mathcal{X} \to \mathcal{K}$ pre-trained on a source dataset $\hat{S} = \{(x_i^s, y_i^s)\}_{i=1}^n$ sampled *i.i.d.* from source distribution S, where $x_i^s \in \mathcal{X} \subseteq \mathbb{R}^D$ and $y_i^s \in \mathcal{Y} = \{1, \ldots, K\}$. We are also given an unlabeled target dataset $\hat{T} = \{x_i^t\}_{i=1}^m$ sampled *i.i.d.* from the target distribution T. Let $\mathcal{K} = \{c \in \mathbb{R}^K : \sum_{y \in \mathcal{Y}} c[y] = 1, c[y] \in [0, 1]\}$ denote the output space, facilitating multi-class labeling via the softmax function.

Lemma 1 (Target Error Upper Bound[4]). *Let $f_S, f_T : \mathcal{X} \to \mathcal{K}$ be the true labeling functions for the source and target domains, respectively, outputting one-hot vectors. Let $\epsilon : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ denote a distance metric, and let $\epsilon_D(f, f') := \mathbb{E}_{x \sim D} \epsilon(f(x), f'(x))$ measure the disagreement between f, f' over distribution D. Let $f_S^* = \arg\min_{f \in \mathcal{H}} \epsilon_{\hat{S}}(f_S, f)$ be the pre-trained source model. For $\forall\ f \in \mathcal{H}$, the minimal target error is bounded by:*

$$\min_{h \in \mathcal{H}} \epsilon_T(f_T, h) \leq \epsilon_T(f_T, f_S^*) \leq \epsilon_T(f_S^*, f) + \epsilon_T(f_T, f), \tag{1}$$

implying that the target error is bounded by $\epsilon_T(f_T, f)$, the deviation between any f and the true labeling function f_T. To render this upper bound optimizable in the absence of true target labels, we introduce a finite realizability assumption to circumvent the intractability of f_T.

Assumption 1 (Finite Realizability [4]). *$\exists f_\rho \in \mathcal{H}_T \subseteq \mathcal{H}$ and $\rho \gtrapprox 0$ s.t. $\epsilon_{\hat{T}}(f_T, f_\rho) \leq \rho$ given a finite sample $\hat{T}$ and an appropriate hypothesis space $\mathcal{H}$.*

By setting $f = f_\rho$ (ρ-approximated labeling function) in Lemma 1 and leveraging Assumption 1, we derive a generalized upper bound based on empirical Rademacher complexity that avoids the intractable true labeling function f_T.

[4] See proofs and justifications for the theorems and assumption in Suppl. 7 & 8.

Definition 1 (Vector-Valued Rademacher Complexity [40]). *Given a class of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}^K$, finite samples $\hat{D} = \{x_i\}_{i=1}^m$ drawn i.i.d. from a distribution D, and independent uniform random variables $\sigma_{ik} \in \{-1, +1\}$, the vector-valued Rademacher complexity of $\mathcal{F}$ is defined as*:

$$\hat{\Re}_{\hat{D}}^K(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K \sigma_{ik} f(x_i)[k] \right]. \tag{2}$$

Theorem 1 (Generalization Error[4]). *Given an ϵ satisfying L-Lipschitz w.r.t. the second argument and bounded functions $x \mapsto \epsilon(f_S^*(x), f(x)) + \epsilon(f_T(x), f(x)) \leq M : f \in \mathcal{H}$, for $\forall\ \delta \in (0, 1]$, with probability at least $1 - \delta$:*

$$\min_{h \in \mathcal{H}} \epsilon_T(f_T, h) \leq \sup_{f_T' \in \mathcal{H}_T} \underbrace{\epsilon_{\hat{T}}(f_S^*, f_T')}_{L_{adv}} + 4\sqrt{2} L \hat{\Re}_{\hat{T}}^K(\mathcal{H}) + 3M \sqrt{\frac{\log(2/\delta)}{2m}} + \rho. \tag{3}$$

Minimizing the RHS of Eq. (3)—by parameterizing hypotheses via a shared encoder $\mathcal{G} : \mathcal{X} \to \mathcal{Z}$ and seeking a representation space $\mathcal{Z}$ that minimizes L_{adv}—effectively constrains $\epsilon_T(f_T, f_S^*)$ toward the minimum target error. Unlike [75] relying on the classical UDA theory [1] with its unbounded joint-error, our finite realizability assumption remains robust even under large domain shifts (Suppl. 8), ensuring the tractability of the upper bound during optimization.

3.2 Gap between Theory and Practice

Theorem 1 suggests that $\mathcal{H}_T$ containing the best empirical target approximator $f_T^* = \arg\min_{f \in \mathcal{H}} \epsilon_{\hat{T}}(f_T, f)$ provides the strongest generalization if the supremum is not achieved at f_T^* . However, an improperly constrained hypothesis space may yield a loose bound—either through an inflating approximation error ρ or excessive maximization that drives the target hypothesis f_T' significantly away from the pre-trained f_S^*. Such divergence risks semantic drift from inherited source knowledge, which can be catastrophic particularly for large-scale VLMs. While a conventional remedy involves incorporating source error terms into the optimization objective, this is precluded in the source-free setting. To bridge this gap between theory and practice, we propose a distributional proxy to approximate this constrained optimization. By recovering source-aligned feature representations, we transform the theoretical bound into a tractable task that can be solved via adversarial perturbation, as detailed in the following section.

4 Methodology

Domain Adaptation (DA) seeks domain-invariant representations. In SFDA, the absence of source data prevents direct distribution alignment. Methods such as SHOT [31] preserve source prototypes via rigid, weight-normalized classifiers, enabling feature-to-classifier alignment. However, such architectural constraints are restrictive for arbitrary pre-trained models. To alleviate these limitations and

bridge the gap identified in Theorem 1, we propose a flexible framework that approximates source prototypes without imposing structural constraints on the classifier, ensuring applicability across diverse architectures.

4.1 Adversarial Proxy Perturbation

Definition 2 (Distributional Proxy Set). *To model the missing source features, we define a distributional proxy set* $\mathcal{Q} = \{Q(z, y) = \pi_y Q(z|y)\}_{y\in\mathcal{Y}}$ *over* $\mathcal{Z} \times \mathcal{Y}$. *In the absence of prior label knowledge, we further set a uniform class prior* $\pi_y = 1/K$. *Each conditional distribution satisfies* $Q(z|y) \in \mathcal{B}_\xi(\delta_{\mu^y})$, *the Wasserstein-1 ball of radius* ξ *centered at the class prototype* μ^y :

$$\mathcal{B}_\xi(\delta_{\mu^y}) = \{Q(z|y) \in \mathcal{P}(\mathcal{Z}) : W_1(Q(z|y), \delta_{\mu^y}) \leq \xi\}, \tag{4}$$

and for a Dirac measure δ_{μ^y}, $W_1(Q(z|y), \delta_{\mu^y}) = \mathbb{E}_{z\sim Q(z|y)}\|z - \mu^y\|$.

Unlike parametric models (e.g., Gaussians), this non-parametric formulation provides robustness against arbitrary distributional shifts within a transport budget ξ. In Theorem 1, the adversarial loss L_{adv} prefers a supremum over the space $\mathcal{H}_T \subseteq \mathcal{H}$ containing f_T^*. Given the absence of target labels, predictive consistency on the source data serves as an effective surrogate constraint [50,65,69]. To compensate for missing source data, we replace the source distribution with its proxy counterpart Q, yielding an upper bound on predictive consistency.

Corollary 1 (Predictive Consistency[4]). *Let* $Q^y := Q(z|y)$ *and* $S^y := S(x|y)$ *denote the conditional distributions over the feature space* $\mathcal{Z}$ *and input space* $\mathcal{X}$, *respectively. Given a decomposed source model* $f_S^* \circ \mathcal{G}^* : \mathcal{X} \to \mathcal{K}$ *and pre-trained encoder* $\mathcal{G}^* : \mathcal{X} \to \mathcal{Z}$, *the following holds for* $\forall Q \in \mathcal{Q}$ *and* $\forall\ f'_T \in \mathcal{H}' : \mathcal{Z} \to \mathcal{K}$ *provided* Q^y *and* S^y *share the same label distribution* π_y :

$$\epsilon_S(f_S^*, f'_T; \mathcal{G}^*) \leq \underbrace{\epsilon_Q(f_S^*, f'_T)}_{L_{ce}} + \sum_{y\in\mathcal{Y}} \pi_y \underbrace{\sup_{f,f'\in\mathcal{H}'} |\epsilon_{Q^y}(f, f') - \epsilon_{S^y}(f, f'; \mathcal{G}^*)|}_{d_{\mathcal{H}'\Delta\mathcal{H}'}(Q^y, S^y; \mathcal{G}^*)}, \tag{5}$$

$d_{\mathcal{H}'\Delta\mathcal{H}'}(Q^y, S^y; \mathcal{G}^*)$ quantifies the divergence between Q^y and features from S^y. According to [12], the surrogate for $d_{\mathcal{H}'\Delta\mathcal{H}'}(Q^y, S^y; \mathcal{G}^*)$ can be expressed by the following if we consider f_S^* as a domain discriminator that distinguishes Q^y from features of S^y for $\forall y$ by maximizing:

$$2\left(1 + \mathop{\mathbb{E}}_{z\sim Q^y} \log(1 - f_S^*(z)[y]) + \mathop{\mathbb{E}}_{x\sim S^y} \log(f_S^*(\mathcal{G}^*(x))[y])\right). \tag{6}$$

However, optimizing f_S^* as a discriminator is problematic without source data to maintain its classification utility. Therefore, by noticing that $\mathbb{E}_{x\sim S^y} \log(f_S^*(\mathcal{G}^*(x))[y])$ is nearly zero for a well-trained source model, we instead optimize over Q^y:

$$\sup_{Q^y\in\mathcal{B}_\xi(\delta_{\mu^y})} \mathop{\mathbb{E}}_{z\sim Q^y} \log(1-f_S^*(z)[y]) = \inf_{\beta\geq 0} \left\{\beta\xi + \sup_{z\in\mathcal{Z}}[\log(1 - f_S^*(z)[y]) - \beta\|z - \mu^y\|]\right\}, \tag{7}$$

which represents the Distributionally Robust Optimization (DRO) objective and its corresponding Kantorovich dual form with Lagrange multiplier β [23]. The absence of source data necessitates estimating proxy centroids μ^y via the only available target data. Specifically, we employ an encoder $\mathcal{G} : \mathcal{X} \to \mathcal{Z}$ (initialized as $\mathcal{G}^*$) to extract target representations and compute class-wise prototypes via soft assignments induced by f_S^*. To avoid costly full-updates at every iteration, we compute batch-wise prototypes $\mu^y_{\hat{B}}$ and update the global prototype μ^y with Exponential Moving Average (EMA):

$$\mu^y \leftarrow \alpha\mu^y + (1-\alpha)\mu^y_{\hat{B}}, \quad \mu^y_{\hat{B}} := \frac{\sum_{x\in\hat{B}} f_S^*(\mathcal{G}(x))[y] \cdot \mathcal{G}(x)}{\sum_{x\in\hat{B}} f_S^*(\mathcal{G}(x))[y]}. \tag{8}$$

While the exact dual form is theoretically complete, it is often computationally prohibitive for non-convex models because it requires a nested optimization over β and a global search for the worst-case distribution support. To arrive at a more tractable objective, we follow the approach in [42] by substituting $z := \mu^y + r^y$, where r^y represents an adversarial perturbation. Since the Wasserstein-1 ball is centered at a Dirac measure, its worst-case distribution concentrates on a single adversarial point [27]. Hence, we can simplify the minimax problem in Eq. (7) into a constrained optimization over a perturbed prototype:

$$\begin{cases} \min_{\mathcal{G}} \underbrace{\sum_{y\in\mathcal{Y}} \pi_y \log(1 - f_S^*(\mu^y + r^y_{adv})[y])}_{L_{dis}} \\ \text{subject to:} \quad r^y_{adv} = \arg\max_{\|r^y\|\le\xi} \log(1 - f_S^*(\mu^y + r^y)[y]), \end{cases} \tag{9}$$

yielding an objective analogous to GANs [15], where minimizing L_{dis} with respect to $\mathcal{G}$ encourages reduction of a surrogate Jensen–Shannon divergence between Q^y and S^y. Since a closed-form solution for r^y_{adv} is generally unavailable, we approximate the inner maximization. Rather than optimizing the dual variable β, which can introduce instability during initialization, we employ projected gradient descent [39]. For a first-order approximation, the worst perturbation is:

$$r^y_{adv} \approx \xi \frac{g^y}{\|g^y\|}, \quad \text{with } g^y = \nabla_z \log(1 - f_S^*(z)[y])\big|_{z=\mu^y}. \tag{10}$$

Consequently, the worst-case proxy distribution collapses to $Q^y = \delta_{\mu^y + r^y_{adv}}$ for each class. We therefore define L_{ce} in Eq. (5) and the constrained hypothesis space $\mathcal{H}'_T = \{f'_T \in \mathcal{H}' : L_{ce} \le \tau\}$ at these adversarial anchors.

4.2 Surrogate Objectvie

The derivation of Lemma 1 relies on the triangle inequality of the $0-1$ loss ϵ, which is non-differentiable and thus unsuitable for gradient-based optimization. To address this, we define a surrogate loss to enable smooth optimization.

Definition 3 (Cross-Entropy Discrepancy). *Let $f, f' : \mathcal{X} \to \mathcal{K}$ denote multi-class labeling functions, where $f(x)[y]$ indicates the probability of x labeled as y.*

Given induced labeling function $l_f(x) = \arg\max_{y \in \mathcal{Y}} f(x)[y]$*, the Cross-Entropy Discrepancy (CED) between* f, f' *over a distribution* D *is defined by*:

$$\mathrm{ced}_D(f, f') = -\mathbb{E}_{x \sim D} f(x)[l_f(x)] \log f'(x)[l_f(x)]. \tag{11}$$

CED is non-negative and locally Lipschitz continuous with respect to its second argument, ensuring the model complexity remaining tractable as $\hat{\mathfrak{R}}^K_{\hat{T}}(\mathcal{H})$ during optimization. Crucially, it serves as an upper bound for the $0-1$ loss such that $\epsilon_D(f, f') \leq \frac{K}{\log 2} \mathrm{ced}_D(f, f')$ while ensuring stable adaptation. Consequently, we formulate L_{ce} and L_{adv} using CED to preserve the validity of the inequalities required in Theorem 1 and Corollary 1 (up to a constant scaling factor). Moreover, this specific design ensures that L_{adv} aligns the proxy distribution Q with target domain T via minimal balanced Jensen-Shannon divergence (Suppl. 9).

Remark 1. For training stability and avoid vanishing or exploding gradients typical in adversarial optimization, we follow the modification proposed in [15]. Specifically, we optimize $\mathbb{E}_{x \sim D} f(x)[y] \log(1 - f'(x)[y])$ as a surrogate for L_{adv}.

4.3 Algorithm

In this section, we incorporate self-supervised regularizations L_{ssl} to refine the constrained space $\mathcal{H}'_T = \{f'_T \in \mathcal{H}' : L_{ce} + \gamma L_{ssl} \leq \tau\}$ since the proxy distribution primarily maintains consistency with source predictions, rather than directly ensuring reliable target domain classification.

Weighted Pseudo-Labeling. For $x \in \hat{T}$ and its random augmentation x' both encoded by $\mathcal{G}$, we minimize a CED-like objective between f^*_S and f'_T, which favors high-confidence predictions, thereby mitigating the impact of noisy pseudo-labels:

$$L_{pse} = -\frac{1}{m} \sum\nolimits_{x \in \hat{T}} f^*_S(\mathcal{G}(x))[l_{f^*_S \circ \mathcal{G}}(x)] \log f'_T(\mathcal{G}(x'))[l_{f^*_S \circ \mathcal{G}}(x)]. \tag{12}$$

Regularized Entropy Minimization. As introduced in [14], we add a class balance regularization by encouraging a uniform estimated label distribution $\hat{p}[y] := \sum_{x \in \hat{T}} f'_T(\mathcal{G}(x))[y]/m$ to circumvent the trivial solution where all unlabeled data have the same one-hot encoding resulting from entropy minimization [16] for a more diversity-promoting solution:

$$L_{ent} = -\frac{1}{m} \sum\nolimits_{x \in \hat{T}, y \in \mathcal{Y}} f'_T(\mathcal{G}(x))[y] \log f'_T(\mathcal{G}(x))[y] + \sum\nolimits_{y \in \mathcal{Y}} \hat{p}[y] \log \hat{p}[y]. \tag{13}$$

Overall Optimization. Following the requirements of Theorem 1, our algorithm (Algorithm 1) minimizes the proxy-based consistency losses (L_{ce}, L_{dis}) alongside the self-supervised constraints $L_{ssl} = L_{pse} + L_{ent}$ to fulfill the constraint on $\mathcal{H}'_T$. By noticing that $\min_{h \in \mathcal{H}} \epsilon_T(f_T, h) \leq \min_{\mathcal{G}} \sup_{f'_T \in \mathcal{H}'_T} \epsilon_{\hat{T}}(f^*_S \circ \mathcal{G}, f'_T \circ \mathcal{G}) + Const.$ given Assumption 1 over $\mathcal{Z}$ and a bounded parametric family $\mathcal{G}$ with controlled capacity, we optimize the adversarial loss L_{adv} through a min-max game over $\mathcal{G}$ and f'_T using a trade-off parameter λ (acting as the Lagrange multiplier that implicitly determines τ):

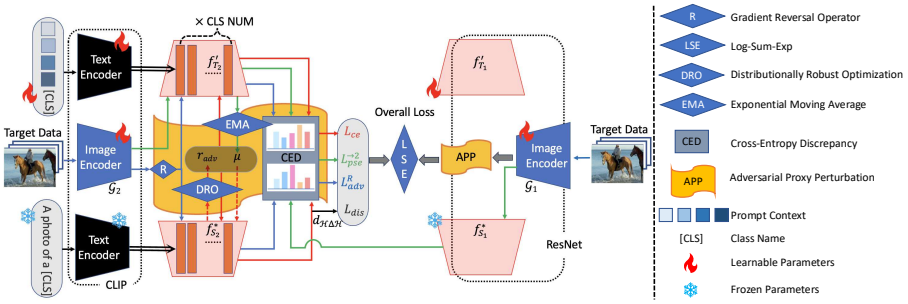


Fig. 1: Pipeline of APP^M learning integrated model based on a pre-trained ResNet model $f^*_{S_1} \circ \mathcal{G}^*_1$ and a VLM such as CLIP $f^*_{S_2} \circ \mathcal{G}^*_2$, where the classifier of CLIP $f^*_{S_2}$ is given by the encoded prompts as "a photo of a [CLS]" (learnable parameters $\mathcal{G}_1, \mathcal{G}_2, f'_{T_1}, f'_{T_2}$ are initialized by $\mathcal{G}^*_1, \mathcal{G}^*_2, f^*_{S_1}, f^*_{S_2}$; L_{ent} & $L^{\to 1}_{pse}$ are omitted for simplicity).

$$\begin{cases} \min_{f'_T \in \mathcal{H}'} L_{ce} + \gamma L_{ssl} - \lambda L_{adv}, \\ \min_{\mathcal{G}} L_{dis} + L_{ce} + \gamma L_{ssl} + \lambda L_{adv}. \end{cases} \tag{14}$$

4.4 Co-training with Vision-Language Models

Vision-Language Models (VLMs), such as CLIP [48] and ALIGN [21], have demonstrated remarkable success by learning modality-invariant representations across vast datasets. Recent literature explores two primary paradigms for leveraging VLMs as external knowledge: zero-shot application [30], which exploits inherent generalization, and domain-specific tailoring via prompt tuning or adapters [5]. In the context of SFDA, DIFO [52] aligns target features to a customized VLM latent space via alternatively updating CLIP and pre-trained model based on predictive consistency minimization. Building on this, we generalize Theorem 1 to integrate knowledge from multiple models through a co-training framework with mutual distillation (see Fig. 1).

Theorem 2 (Multi-Source Knowledge Integration [4]). *Let $f^*_{S_i} \in \mathcal{H}_i \subseteq \mathcal{H} : \mathcal{X} \to \mathcal{K}$ denote i-th source model. Suppose for $\forall i$, $\exists f_\rho \in \mathcal{H}_{T_i} \subseteq \mathcal{H}_i$ s.t. $\epsilon_{\hat{T}}(f_T, f_\rho) \le \rho$. Given an ϵ satisfying L-Lipschitz w.r.t. the second argument and bounded functions for all N sources $x \mapsto \epsilon(f^*_{S_i}(x), f(x)) + \epsilon(f_T(x), f(x)) \le M$: $f \in \mathcal{H}_i$, for $\forall\ \delta \in (0, 1]$, with probability at least $1 - \delta$:*

$$\min_{h \in \mathcal{H}} \epsilon_T(f_T, h) \le \sup_{\{f'_{T_i} \in \mathcal{H}_{T_i}\}_{i=1}^N} \log \sum_{i=1}^N e^{\epsilon_{\hat{T}}(f^*_{S_i}, f'_{T_i})} + 4\sqrt{2}L \max_i \hat{\mathfrak{R}}^K_{\hat{T}}(\mathcal{H}_i) + \mathcal{O}(\sqrt{\frac{M^2}{m} \log \frac{N}{\delta}}) + \rho. \tag{15}$$

Algorithm 1 APP

Input: source model $f_S^* \circ \mathcal{G}^* : \mathcal{X} \to \mathcal{K}$, unlabeled target data $\hat{T}$
Parameter: encoder $\mathcal{G} : \mathcal{X} \to \mathcal{Z}$, target hypothesis $f_T' \in \mathcal{H}' : \mathcal{Z} \to \mathcal{K}$; $w = (\mathcal{G}, f_T')$
Hyper-parameter:coefficients $\lambda, \gamma, \alpha, \xi$; learning rate η
Initialization: $f_T' \leftarrow f_S^*, \mathcal{G} \leftarrow \mathcal{G}^*, \mu^y \leftarrow \mu_{\hat{T}}^y$
for *iteration* $= 1, 2, \ldots$ **do**
  randomly sample a mini-batch $\hat{B}$ from $\hat{T}$
  compute the perturbation loss L_{dis} and hypothesis constraints L_{ce}, L_{ssl} given $\hat{B}$
  compute the adversarial loss L_{adv}^R given the gradient reversal operator $R(\cdot)$:
    $L_{adv}^R = \text{ced}_{\hat{B}}(f_S^* \circ R \circ \mathcal{G}, f_T' \circ R \circ \mathcal{G})$
  minimize overall objective w.r.t. all parameters w:
    $w \leftarrow w + \eta \Delta w$, $\Delta w = -\frac{\partial (L_{dis} + L_{ce} + \gamma L_{ssl} - \lambda L_{adv}^R)}{\partial w}$
end for

We adopt a prompt learning framework [74] where the learnable components include the image encoder $\mathcal{G}_2$ and text-encoded prompts f_{T_2}' each assigned for a specific class initialized as "a photo of a [CLS]" s.t. $f_{T_2}' \circ \mathcal{G}_2 \in \mathcal{H}_2$ ([CLS] denotes the classname). A significant challenge in SFDA is customizing these domain-generic VLMs for a specific target domain without reliable supervision. To address this, we leverage the "wisdom of the crowd" through predictive interaction. Specifically, we replace the standard weighted pseudo-labeling in Eq. (12) with mutual distillation among models, minimizing the cross-entropy discrepancy between $f_{\neq i} := \sum_{j:j\neq i} f_{S_j}^* \circ \mathcal{G}_j/(N-1)$ and $f_{T_i}' \circ \mathcal{G}_i$:

$$L_{pse}^{\to i} = -\frac{1}{m} \sum_{x \in \hat{T}} f_{\neq i}(x)[l_{f_{\neq i}}(x)] \log f_{T_i}'(\mathcal{G}_i(x'))[l_{f_{\neq i}}(x)], \tag{16}$$

such that we can co-train both models based on Theorem 2 to learn an integrated target model which leverages wider knowledge from multiple domains. This co-training strategy encourages the model to inherit specific domain expertise from the source model while maintaining the broad contextual knowledge of the VLM. This interactive paradigm avoids the pitfalls of over-reliance on a single, potentially biased model.

5 Evaluation

As for the evaluation setting, the trade-off parameter λ and self-supervised coefficient γ is set to 0.01 and 0.1 during the training procedure according to [65]. In addition, we empirically set EMA ratio α to 0.9 and adversarial perturbation coefficients ξ to 1.0 for ResNet while 0.1 for CLIP since its features are normalized (Suppl. 11 for sensitivity analysis and statistics). Source models are prepared based on the protocol established in [31] using label smoothing [43]. All experiments are conducted with the ImageNet [8] pre-trained ResNet-50 (for Office-Home & DomainNe)/ ResNet-101 [17] (for VisDA) and ViT-B/32 [10] as the encoder $\mathcal{G}$. For ResNet backbone, we deploy 2-layer neural network for classifiers, trained

Table 1: Closed-set SFDA on Office-Home & VisDA-C (%). SF and M means source-free and multimodal, respectively.

Method	Venue	SF	M	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.	VisDA-C
Source	–	–	–	43.7	67.0	73.9	49.9	60.1	62.5	51.7	40.9	72.6	64.2	46.3	78.1	59.2	49.2
SHOT [31]	ICML20	✓	✗	56.2	76.2	81.1	68.2	77.8	79.1	68.8	54.4	82.0	74.7	58.4	84.2	71.8	82.7
NRC [62]	NIPS21	✓	✗	57.7	80.3	82.0	68.1	79.8	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2	85.9
AaD [61]	NIPS22	✓	✗	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7	88.0
AdaCon [4]	CVPR22	✓	✗	47.2	75.1	75.5	60.7	73.3	73.2	60.2	45.2	76.6	65.6	48.3	79.1	65.0	86.8
CoWA [26]	ICML22	✓	✗	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5	86.9
ELR [64]	ICLR23	✓	✗	58.4	78.7	81.5	69.2	79.5	79.3	66.3	58.0	82.6	73.4	59.8	85.1	72.6	85.8
PLUE [32]	CVPR23	✓	✗	49.1	73.5	78.2	62.9	73.5	74.5	62.2	48.3	78.6	68.6	51.8	81.5	66.9	88.3
I-SFDA [41]	CVPR24	✓	✗	60.7	78.9	82.0	69.9	79.5	79.7	67.1	58.8	82.3	74.2	61.3	86.4	73.4	88.4
RGV [75]	CVPR25	✓	✗	61.2	80.9	82.7	69.3	81.2	81.4	68.1	58.8	83.4	74.6	62.4	85.7	74.1	89.1
DVD [57]	NIPS25	✓	✗	60.1	79.6	82.5	69.1	80.8	80.6	67.9	58.5	84.3	73.5	59.3	87.6	73.7	88.9
APP	–	✓	✗	61.6	79.8	82.9	71.4	81.2	81.8	70.1	58.1	84.3	75.0	64.0	84.9	74.6	89.5
DIFO [52]	CVPR24	✓	✓	70.6	90.6	88.8	82.5	90.6	88.8	80.9	70.1	88.9	83.4	70.5	91.2	83.1	90.3
DUET [25]	NIPS25	✓	✓	73.6	90.4	91.0	83.6	90.7	90.9	82.7	73.7	91.2	83.6	74.0	91.2	84.7	91.4
APP^M	–	✓	✓	77.2	91.8	90.1	84.8	92.1	90.2	84.8	78.1	90.4	84.6	77.4	92.4	**86.2**	**91.7**

Table 2: Closed-set SFDA on DomainNet (%). SF and M means source-free and multimodal, respectively. The slash denotes full avg./subset avg.

Method	Venue	SF	M	C→P	C→R	C→S	P→C	P→R	P→S	R→C	R→P	R→S	S→C	S→P	S→R	Avg.
Source	–	–	–	44.6	59.8	47.5	53.3	75.3	46.2	55.3	62.7	46.4	55.1	50.7	59.5	54.7
SHOT	ICML20	✓	✗	63.5	78.2	59.5	67.9	81.3	61.7	67.7	67.6	57.8	70.2	64.0	78.0	68.1
NRC	NIPS21	✓	✗	62.6	77.1	58.3	68.2	81.3	60.7	64.7	69.4	57.3	66.5	65.0	78.7	67.1
AdaCon	CVPR22	✓	✗	60.0	74.8	55.9	62.2	78.3	58.2	63.1	67.1	55.6	67.1	66.0	75.4	65.4
CoWA	ICML22	✓	✗	64.6	80.6	60.0	66.2	79.8	60.8	69.0	67.2	60.0	69.0	64.5	79.9	68.4
PLUE	CVPR23	✓	✗	59.8	74.0	56.0	61.6	78.5	57.9	61.6	65.9	53.8	67.3	64.3	76.0	64.7
TPDS [53]	IJCV23	✓	✗	62.9	77.1	59.9	65.1	79.0	61.5	66.4	67.0	58.2	68.4	64.0	75.3	67.1
RGV	CVPR25	✓	✗	–	–	68.0	71.9	83.2	–	77.6	–	67.3	–	71.2	–	–/73.2
APP	–	✓	✗	65.5	75.5	67.8	76.2	81.3	72.2	77.6	74.2	69.0	78.0	70.2	77.5	73.8/73.8
DIFO	CVPR24	✓	✓	76.6	87.2	74.9	80.0	87.4	75.6	80.8	77.3	75.5	80.5	76.7	87.3	80.0
DUET	NIPS25	✓	✓	80.1	89.6	79.0	82.4	89.8	79.2	82.8	80.6	78.8	83.0	80.3	89.6	82.9
APP^M	–	✓	✓	82.1	90.2	80.0	86.4	89.4	80.2	86.5	81.5	79.9	86.0	81.9	90.0	**84.5**

with Stochastic Gradient Descent optimizer with a 1e-3 learning rate annealed according to [13] and a momentum of 0.9. For CLIP backbone, the encoded prompts are used as classifiers, trained with AdamW [36] with a 1e-6 learning rate annealed by cosine decay.

Office-Home [56] is a widely-used medium-sized domain adaptation benchmark, which consists of 15,500 images from 65 categories and four distinct domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-World images.

DomainNet [44] is a challenging benchmark dataset for large-scale domain adaptation that has 345 classes and 6 domains. Following the protocol established in [49], we pick four domains (Real, Clipart, Painting, Sketch) with 126 classes.

VisDA-C [45] is a challenging large-scale benchmark that focuses on the 12-class synthetic-to-real object recognition task. The source domain contains 152k synthetic images generated by rendering 3D models while the target domain has 55k real object images sampled from Microsoft COCO.

Table 3: Comparison of Partial-set and Open-set SFDA methods.

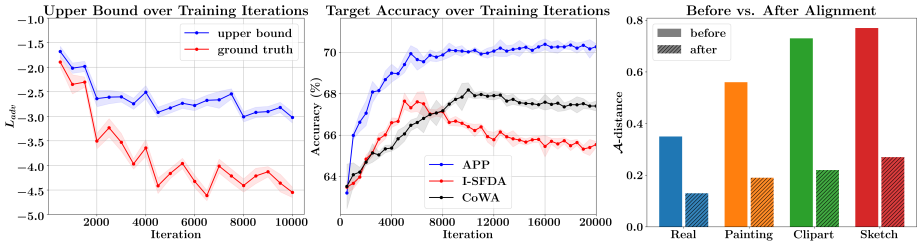
Partial-set SFDA	Venue	Avg.	Open-set SFDA	Venue	Avg.
Source	–	62.8	Source	–	46.6
SHOT	ICML20	79.3	SHOT	ICML20	72.8
HCL [20]	NIPS21	79.6	HCL	NIPS21	72.6
CoWa	ICML22	83.2	CoWa	ICML22	73.2
AaD	NIPS22	79.7	AaD	NIPS22	71.8
CRS [68]	CVPR23	80.6	CRS	CVPR23	73.2
DIFO	CVPR24	85.6	DIFO	CVPR24	75.9
APP^M	–	**88.3**	**APP^M**	–	**85.2**

Table 4: Ablation Studies on Office-Home & DomainNet Dataset.

Config.	Office-Home				DomainNet			
	Ar→ Cl		Ar→ Rw		P→ S		P→ R	
	Acc.	$\mathcal{A}$-dist.	Acc.	$\mathcal{A}$-dist.	Acc.	$\mathcal{A}$-dist.	Acc.	$\mathcal{A}$-dist.
Source	43.7	1.72	73.9	0.99	46.2	1.55	75.3	1.04
APP	61.6	0.63	82.9	0.36	72.2	0.56	81.3	0.41
w/o $L_{dis} + L_{ce}$	57.6	1.01	82.1	0.53	67.8	0.92	80.6	0.60
w/o L_{adv}	58.3	0.95	81.9	0.56	68.6	0.88	80.3	0.59
w/o L_{ssl}	60.2	0.70	81.4	0.49	70.5	0.62	79.7	0.53

5.1 Object Recognition (closed-set)

We evaluate APP across three diverse object recognition benchmarks: Office-Home, DomainNet, and VisDA-C, under the standard closed-set setting ($\mathcal{Y}_S = \mathcal{Y}_T$). In Tabs. 1 and 2, APP consistently outperforms existing baselines across all benchmarks, both in unimodal and multimodal configurations. Our empirical results suggest that the performance gain is particularly pronounced in scenarios characterized by significant domain shift. For instance, on the VisDA-C benchmark, the performance margins are more subtle. We attribute this to the fact that standard backbones (e.g., ResNet, CLIP) are pre-trained on ImageNet, which shares a high degree of structural and semantic similarity with the VisDA-C Real target domain, thereby providing a strong initial feature prior. Similarly, on the medium-scale Office-Home dataset, the advantage of APP is relatively moderated in transitions towards Product and Real-World domains. This is expected, as both domains consist of realistic imagery, resulting in a narrower distribution gap. Conversely, when the target domain shifts to more abstract representations such as Clipart or Art, APP demonstrates superior robustness, effectively reducing target error despite the increased domain discrepancy. Notably, on the Clipart target domain, APP^M surpasses the previous state-of-the-art, DIFO, by approximately 7%. This trend is further validated on the large-scale DomainNet dataset. Here, the improvements are even more substantial, as domains such as Clipart, Sketch, and Painting deviate significantly from natural image distributions. In these challenging contexts, the integration of multimodal information acts as a key catalyst, bridging the gap where visual-only features struggle; APP^M achieves a remarkable 84.5% average, a 4.5% improvement over the competitor. Finally, while graph-based methods [41, 61, 62] perform competitively on tasks with limited categories, their efficacy may diminish as the label space expands. This is because Euclidean-based neighbor selection loses discriminative power as the density of the feature space increases; with more clusters present, the probability of neighbors being sampled from overlapping class boundaries rises, leading to misaligned affinity graphs. In contrast, APP maintains robust performance by utilizing proxy-induced perturbations that effectively manage high-dimensional class boundaries regardless of category density.



(a) validity **(b)** convergence **(c)** domain shift

Fig. 2: Valid upper bound $\sup_{f'_T \in \mathcal{H}'_T} \epsilon_{\hat{T}}(f^*_S, f'_T; \mathcal{G})$ over the ground truth $\epsilon_{\hat{T}}(f^*_S, f^*_T; \mathcal{G})$ **(a)** that ensures consistent improvement (Pr→Ar) of APP on target accuracy throughout training **(b)**, and efficiency of APP on addressing large domain shift **(c)**.

5.2 Object Recognition (beyond closed-set)

To further demonstrate the robustness of APP against complex distributional discrepancies, we extend our evaluation: Partial-set DA (PDA), where $\mathcal{Y}_T \subset \mathcal{Y}_S$, and Open-set DA (ODA), where $\mathcal{Y}_S \subset \mathcal{Y}_T$. Following [3, 33], for the Office-Home dataset, the PDA task involves 25 target classes sub-sampled from 65 source classes. Conversely, the ODA task utilizes the same 25 classes for the source domain while the target domain encompasses all 65 classes, with the remaining 40 classes treated as unknown. As summarized in Tab. 3, APP significantly exceeds previous baselines, particularly in the ODA setting where it achieves a ∼10% gain, suggesting it is particularly effective at navigating both covariate shift (feature-level misalignment) and label shift (asymmetric category spaces). These findings reinforce our central claim on the distinct advantage in large domain shift scenarios, as the model maintains high discriminative power even when the target domain contains extraneous or missing categories (Suppl. 10).

5.3 Quantification of Domain Shift

To intuitively present domain shift, we quantify it with a well-established metric $\mathcal{A}$-distance [2] in Fig. 2c. For simplicity, we consider CLIP as the only source model pre-trained on a large scale realistic domain, where the text features of initial prompts "a photo of a [CLS]" can be viewed as source features. Target features can be obtained by encoded target images from DomainNet. We train an SVM [7] to distinguish source from target features and the $\mathcal{A}$-distance is proportional to the accuracy. The results demonstrate that APP is highly effective at reducing large domain shifts. This reduction is particularly pronounced when the target domain is unrealistic (e.g., Clipart and Sketch), confirming the algorithm's ability to bridge significant distributional gaps.

5.4 Ablation Study

We conduct an ablation study in Tab. 4 to evaluate the contribution of each component. While all loss functions consistently improve performance, their

impact varies with the degree of domain shift. In large domain shift tasks (e.g., Ar→Cl and P→S), the alignment-based losses L_{dis} and L_{adv} provide the most significant gains and lead to a notable reduction in $\mathcal{A}$-distance. This empirically suggests that these components effectively minimize the distributional discrepancy between domains.Conversely, for small domain shift tasks where the target is a realistic dataset (e.g., Ar→Rw and P→R), the initial $\mathcal{A}$-distance is lower, and the contribution of L_{ssl} becomes more prominent. In these cases, since the features are already relatively well-aligned, the focus shifts from global distribution matching to refining local cluster discriminability through self-supervised learning.

5.5 Validity and Stable Convergence

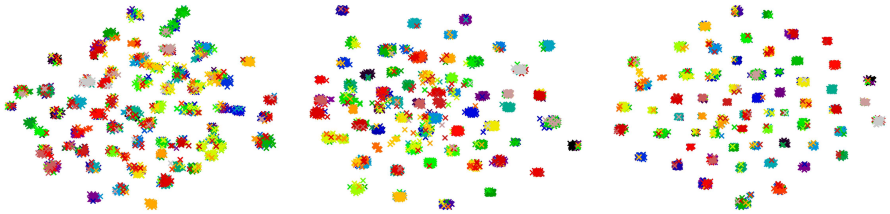
To avoid the looseness, we propose the distributional proxy and self-supervised regularization to create constrained hypothesis space $\mathcal{H}'_T$. While it remains difficult to ensure the decomposed empirical target approximator $f_T^* = \arg\min_{f \in \mathcal{H}'} \epsilon_{\hat{T}}(f_T, f \circ \mathcal{G})$ is covered in practice, the empirical validity of our theoretical proposal is confirmed in Fig. 2a, which shows that the APP objective remains a consistent upper bound of the ground truth throughout the entire training process. This validity results in superior training dynamics: APP offers remarkably stable convergence without suffering from performance decay after reaching an optimum (Fig. 2b) due to the lack of theoretical guarantees.

5.6 Feature Visualization

Prior works such as SHOT, often rely on a fixed, weight-normalized linear classifier to preserve source prototypes and guide feature alignment. While this architecture is widely adopted due to its efficiency [41, 52, 61, 62], it restricts the richness of the hypothesis space and imposes rigid architectural limitations on the pre-trained models. As illustrated in Fig. 3, existing approaches struggle to effectively group target features when these specific architectural constraints are absent. For instance, the t-SNE visualizations for AaD and I-SFDA show significant cluster overlap and less distinct category boundaries. In contrast, APP yields clearly separated and compact clusters. This superior performance stems from APP's ability to align target features with proxy centroids, ensuring high discriminative power without requiring specific classifier architectures.

6 Conclusion

We tackle SFDA by proposing a simple yet effective learning framework grounded in a finite realizability assumption, which remains robust under large domain shifts. To bridge the gap between theory and algorithm, we introduce a proxy distribution to approximate source prototypes through adversarial perturbation, enabling smooth optimization. Furthermore, we extend our formulation to multi-source scenarios and seamlessly integrate the zero-shot prediction capability of



(a) AaD **(b)** I-SFDA **(c)** APP

Fig. 3: Target features visualized with t-SNE [38] of Rw→Cl in Office-Home, where APP offers clearly separated clusters.

VLMs via co-training to enhance performance. Extensive experiments on multiple benchmarks, including open/partial-set tasks, demonstrate that our model consistently outperforms recent baselines across both unimodal and multimodal backbones, validating the effectiveness of our approach.

Acknowledgements

This research is partially supported by JSPS KAKENHI Grant Number JP26K21298, JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. Machine Learning **79**, 151–175 (2010)
2. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 19 (2006)
3. Cao, Z., You, K., Long, M., Wang, J., Yang, Q.: Learning to transfer examples for partial domain adaptation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2980–2989 (2019)
4. Chen, D., Wang, D., Darrell, T., Ebrahimi, S.: Contrastive test-time adaptation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 295–305 (2022)
5. Cho, J., Nam, G., Kim, S., Yang, H., Kwak, S.: Promptstyler: Prompt-driven style generation for source-free domain generalization. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15656–15666 (2023)
6. Chopra, S., Kothawade, S., Aynaou, H., Chadha, A.: Source-free domain adaptation with diffusion-guided source data generation. arXiv:2402.04929 (2024)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995)

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
9. Ding, N., Xu, Y., Tang, Y., Xu, C., Wang, Y., Tao, D.: Source-free domain adaptation via distribution estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7202–7212 (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations (ICLR) (2021)
11. Du, Y., Yang, H., Chen, M., Luo, H., Jiang, J., Xin, Y., Wang, C.: Generation, augmentation, and alignment: a pseudo-source domain based method for source-free domain adaptation. Machine Learning **113**, 3611–3631 (2023)
12. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). vol. 37, pp. 1180–1189 (2015)
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research (JMLR) **17**, 1–35 (2016)
14. Gomes, R., Krause, A., Perona, P.: Discriminative clustering by regularized information maximization. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 23 (2010)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 27 (2014)
16. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 17 (2004)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
18. Hegde, D., Patel, V.M.: Attentive prototypes for source-free unsupervised domain adaptive 3d object detection. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3054–3064 (2024)
19. Hou, Y., Zheng, L.: Source free domain adaptation with image translation. arXiv:2008.07514 (2021)
20. Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 3635–3649 (2021)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). vol. 139, pp. 4904–4916 (2021)
22. Karim, N., Mithun, N.C., Rajvanshi, A., Chiu, H.p., Samarasekera, S., Rahnavard, N.: C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24120–24131 (2023)
23. Kuhn, D., Shafiee, S., Wiesemann, W.: Distributionally robust optimization. arXiv:2411.02549 (2025)

24. Lao, Q., Jiang, X., Havaei, M.: Hypothesis disparity regularized mutual information maximization. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 35, pp. 8243–8251 (2021)
25. Lee, J., Park, J.H., Lee, G., Kim, B., Cha, M.H., Nam, H., Jeon, J., Lee, H., Cho, S.I.: Duet: Dual-perspective pseudo labeling and uncertainty-aware exploration & exploitation training for source-free domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 38, pp. 74927–74954 (2025)
26. Lee, J., Jung, D., Yim, J., Yoon, S.: Confidence score for source-free unsupervised domain adaptation. In: Proceedings of the 39th International Conference on Machine Learning (ICML). vol. 162, pp. 12365–12377 (2022)
27. Levine, A., Feizi, S.: Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS). vol. 108, pp. 3938–3947 (2020)
28. Li, R., Jiao, Q., Cao, W., Wong, H.S., Wu, S.: Model adaptation: Unsupervised domain adaptation without source data. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9638–9647 (2020)
29. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7571–7580 (2022)
30. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7061–7070 (2023)
31. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: Proceedings of the 37th International Conference on Machine Learning (ICML). vol. 119, pp. 6028–6039 (2020)
32. Litrico, M., Bue, A.D., Morerio, P.: Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12345–12354 (2023)
33. Liu, H., Long, M., Wang, J., Yang, Q.: Separate to adapt: Open set domain adaptation via progressive separation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2922–2931 (2019)
34. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). vol. 37, pp. 97–105 (2015)
35. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the 7th International Conference on Learning Representations (ICLR) (2019)
37. Luo, Y., Wang, Z., Huang, Z., Baktashmotlagh, M.: Progressive graph learning for open-set domain adaptation. In: Proceedings of the 37th International Conference on Machine Learning (ICML). vol. 119, pp. 6468–6478 (2020)
38. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (JMLR) **9**, 2579–2605 (2008)
39. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations (ICLR) (2018)

40. Maurer, A.: A vector-contraction inequality for rademacher complexities. In: Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT). pp. 3–17 (2016)
41. Mitsuzumi, Y., Kimura, A., Kashima, H.: Understanding and improving source-free domain adaptation from a theoretical perspective. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28515–28524 (2024)
42. Miyato, T., ichi Maeda, S., Koyama, M., Nakae, K., Ishii, S.: Distributional smoothing with virtual adversarial training. In: Proceedings of the 4th International Conference on Learning Representations (ICLR) (2016)
43. Müller, R., Kornblith, S., Hinton, G.: When does label smoothing help? In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
44. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1406–1415 (2019)
45. Peng, X., Usman, B., Kaushik, N., Wang, D., Hoffman, J., Saenko, K.: Visda: A synthetic-to-real benchmark for visual domain adaptation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2102–21025 (2018)
46. Qiu, Z., Zhang, Y., Lin, H., Niu, S., Liu, Y., Du, Q., Tan, M.: Source-free domain adaptation via avatar prototype generation and adaptation. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI). pp. 2921–2927 (2021)
47. Qu, S., Chen, G., Zhang, J., Li, Z., He, W., Tao, D.: Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Proceedings of the 17th European Conference on Computer Vision (ECCV). pp. 165–182 (2022)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). vol. 139, pp. 8748–8763 (2021)
49. Saito, K., Kim, D., Sclaroff, S., Darrell, T., Saenko, K.: Semi-supervised domain adaptation via minimax entropy. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8049–8057 (2019)
50. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3723–3732 (2018)
51. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8722–8732 (2020)
52. Tang, S., Su, W., Ye, M., Zhu, X.: Source-free domain adaptation with frozen multimodal foundation model. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23711–23720 (2024)
53. Tang, S., Zou, Y., Song, Z., Lyu, J., Chen, L., Ye, M., Zhong, S., Zhang, J.: Source-free domain adaptation via target prediction distribution searching. International Journal of Computer Vision (IJCV) **131**, 1–19 (2023)
54. Tian, J., Zhang, J., Li, W., Xu, D.: Vdm-da: Virtual domain modeling for source data-free domain adaptation. IEEE Transactions on Circuits and Systems for Video Technology **32**, 3749–3760 (2022)

55. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2962–2971 (2017)
56. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5385–5394 (2017)
57. Wang, J., Bae, W., Chen, J., Wang, W., Noh, J.: Vicinity-guided discriminative latent diffusion for privacy-preserving domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 38, pp. 146100–146131 (2025)
58. Wang, Y., Zhu, R., Ji, P., Li, S.: Open-set graph domain adaptation via separate domain alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 9142–9150 (2024)
59. Westfechtel, T., Yeh, H.W., Zhang, D., Harada, T.: Gradual source domain expansion for unsupervised domain adaptation. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1946–1955 (2024)
60. Xia, H., Zhao, H., Ding, Z.: Adaptive adversarial network for source-free domain adaptation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8990–8999 (2021)
61. Yang, S., Wang, Y., Wang, K., Jui, S., van de Weijer, J.: Attracting and dispersing: A simple approach for source-free domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 5802–5815 (2022)
62. Yang, S., wang, y., van de Weijer, J., Herranz, L., Jui, S.: Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 29393–29405 (2021)
63. Yang, S., Wang, Y., van de Weijer, J., Herranz, L., Jui, S.: Generalized source-free domain adaptation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8958–8967 (2021)
64. Yi, L., Xu, G., Xu, P., Li, J., Pu, R., Ling, C., McLeod, A.I., Wang, B.: When source-free domain adaptation meets learning with noisy labels. In: Proceedings of the 11th International Conference on Learning Representations (ICLR) (2023)
65. Zhang, D., Westfechtel, T., Harada, T.: Unsupervised domain adaptation via minimized joint error. Transactions on Machine Learning Research (TRML) (2023)
66. Zhang, D., Westfechtel, T., Harada, T.: Open-set domain adaptation via joint error based multi-class positive and unlabeled learning. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). pp. 105 –120 (2024)
67. Zhang, D., Westfechtel, T., Harada, T.: A theory of learning unified model via knowledge integration from label space varying domains. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10142–10152 (2025)
68. Zhang, Y., Wang, Z., He, W.: Class relationship embedded learning for source-free unsupervised domain adaptation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7619–7629 (2023)
69. Zhang, Y., Liu, T., Long, M., Jordan, M.I.: Bridging theory and algorithm for domain adaptation. In: Proceedings of the 36th International Conference on Machine Learning (ICML). vol. 97, pp. 7404–7413 (2019)
70. Zhang, Z., Chen, W., Cheng, H., Li, Z., Li, S., Lin, L., Li, G.: Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 5137–5149 (2022)
71. Zhao, H., des Combes, R.T., Zhang, K., Gordon, G.J.: On learning invariant representations for domain adaptation. In: Proceedings of the 36th International Conference on Machine Learning (ICML). vol. 97, pp. 7523–7532 (2019)

72. Zhao, H., Zhang, S., Wu, G., Costeira, J.P., Moura, J.M.F., Gordon, G.J.: Adversarial multiple source domain adaptation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 31 (2018)
73. Zhao, X., Mithun, N.C., Rajvanshi, A., Chiu, H.P., Samarasekera, S.: Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2388–2398 (2024)
74. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16795–16804 (2022)
75. Zhu, R., Hu, M., Zhuang, W., Lyu, L., Yu, X., Li, S.: Revisiting source-free domain adaptation: Insights into representativeness, generalization, and variety. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25688–25697 (2025)
76. Zou, Y., Yu, Z., Liu, X., Kumar, B.V.K.V., Wang, J.: Confidence regularized self-training. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5981–5990 (2019)