TOWARDS UNIFIED SOURCE-FREE DOMAIN ADAPTATION VIA CAUSALITY-INSPIRED DISCOVERY OF LATENT INVARIANT FACTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

In the pursuit of transferring a source model to a target domain without access to the source training data, Source-Free Domain Adaptation (SFDA) has been extensively explored across various scenarios, including Closed-set, Open-set, Partial-set, Open-partial-set, and Generalized settings. Existing methods, focusing on specific scenarios, not only address a limited subset of challenges but also necessitate prior knowledge of the target domain, significantly limiting their practical utility and deployability. In light of these considerations, we introduce a more practical yet challenging problem, termed *unified SFDA*, which comprehensively incorporates all specific scenarios in a unified manner. In this paper, we propose a novel approach *causality-inspired discovery of latent invariant factors for unified SFDA* (CausalDA). In contrast to previous alternatives that emphasize learning the statistical description of reality, we formulate CausalDA from a causality perspective. The objective is to uncover potential causality-induced connections between latent invariant variables and model decisions, enhancing the reliability and robustness of the learned model against domain shifts. To integrate extensive world knowledge, we leverage a pre-trained vision-language model such as CLIP. This aids in the formation and discovery of latent invariant factors in the absence of supervision in the variation of distribution and semantics, coupled with a newly designed self-supervised invariant information maximization with theoretical guarantees. Extensive experiments demonstrate that CausalDA can achieve new state-of-the-art results in distinct SFDA settings, as well as source-free out-of-distribution generalization.

1 INTRODUCTION

In practical scenarios, using a source model (pre-trained on the source domain) without access to the actual source data, known as Source-Free Domain Adaptation (SFDA) (Kim et al., 2021), has become increasingly feasible. This approach involves the challenges posed by strict data access controls.

At present, SFDA has been explored across various scenarios, as summarized in Tab. 1. The most constrained scenario, termed *Closed-set*, assumes identical categories of interest in both the source and target domains, aiming to address covariate shift, such as domain-specific appearance distributions (Li et al., 2020b). *Generalized SFDA* extends this by requiring the adapted model to perform well in both the source and target domains without forgetting the source domain (Yang et al., 2021b). Moving beyond the vanilla category identity constraint, *Open-set* SFDA (Panareda Busto & Gall, 2017) considers the target domain with additional categories, while *Partial-set* SFDA (Cao et al., 2019) takes the opposite direction. Both scenarios require handling the additional challenge of semantic shift. Synthesizing the two orthogonal

Table 1: Summary of different SFDA settings. C_s/C_t: The category set of the source/target domain, $C = C_s \cap C_t$; AF is short for Anti-Forgetting.

Setting	Covariate-shift	Semantic-shift	AF	Category
Closed-set	✓	✗	✗	$C_s = C_t$
Generalized	✓	✗	✓	$C_s = C_t$
Open-set	✓	✓	✗	$C_s \subset C_t$
Partial-set	✓	✓	✗	$C_s \supset C_t$
Open-partial-set	✓	✓	✗	$\lvert C\rvert/\lvert C_s \cup C_t\rvert \in (0, 1)$
Unified (Ours)	✓	✓	✓	Any

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dimensions leads to *Open-Partial-Set* (Universal) SFDA (Kundu et al., 2020), a generalized setting that explicitly accounts for unknown category relationships.

Typically, prevalent SFDA methods tend to concentrate on particular scenarios, tackling only a subset of challenges. For instance, SFDA-DE(Ding et al., 2022) focuses on Closed-set scenarios, SF-PGL(Luo et al., 2023) on Open-set scenarios, PSAT(Tang et al., 2024b) on Generalized settings, CRS (Zhang et al., 2023) on Partial-set scenarios, and LEAD (Qu et al., 2024) on Open-partial-set scenarios. Nevertheless, this approach significantly restricts their practical utility and deployability, given that we often possess limited prior knowledge about the target domain and minimal control or selection over its conditions.

To address the aforementioned limitation, this study introduces a more realistic yet challenging problem, referred to as *Unified Source-Free Domain Adaptation* (*Unified SFDA*), with the goal of comprehensively addressing all the specific scenarios mentioned in a unified manner. To achieve this, we propose a novel approach *Causality-inspired discovery of latent factors for unified SFDA (CausalDA)*, going beyond conventional statistical association learning of related variables by exploring the underlying causal relationships (Pearl, 2009). Specifically, we build an SFDA-specific Structural Causal Model (SCM) by actively mining causal factors to address the lack of real supervision. We further develop a proxy learning strategy to approximate causal refinement by discovering domain-invariant factors from both external and internal perspectives, leveraging a pre-trained large Vision-Language (ViL) model and the task-specific target model, respectively.

Our **contributions** are summarized as follows: (1) We introduce a unified SFDA problem that aligns with practical demands of real-world deployment, ranging from focus on unknown category relationships to source-knowledge preservation and out-of-distribution robustness. (2) We propose a novel CausalDA approach for unified SFDA. Instead of learning the statistical description of problem reality as conventional methods do, CausalDA is formulated under a causality perspective. (3) Extensive experiments demonstrate that CausalDA consistently achieves new state-of-the-art results on varying SFDA scenarios as well as source-free out-of-distribution generalization.

2 Related Work

Most of the prior SFDA methods consider the Closed-set setting where the source and target domains share the same classes, and the focus is on cross-domain distribution alignment. Existing methods exist in two categories. The first converts SFDA to Unsupervised Domain Adaptation (UDA) by constructing a pseudo-source domain (Li et al., 2020a), exploiting source prototype-guided data generation (Tian et al., 2021) or splitting source-like subset from target data (Du et al., 2024). The second follows self-supervised learning (Liang et al., 2020; Chen et al., 2022b), introducing self-guidance. Besides widely used pseudo-labels (Liang et al., 2020; Litrico et al., 2023), geometry information (Yang et al., 2021a; 2023; Tang et al., 2022), contrastive data (Zhang et al., 2023) and historical data (Tang et al., 2024b) have also been exploited.

Recently, more practical settings have been studied. For example, the Generalized setting shifts aim to mitigate forgetting the source domain (Yang et al., 2021b). A representative approach is combining continual learning techniques with cross-domain adaptation, e.g., the domain attention-based gradient regulation (Yang et al., 2021b) and source guidance constructed by historical information refining (Tang et al., 2024b). There exist works considering Open-set (Luo et al., 2023), Partial-set (Li & Chen, 2023) and Open-partial-set (Kundu et al., 2020) settings. For instance, to control the Open-set risk, a progressive balanced pseudo-labeling strategy (Luo et al., 2023), and Closed-set class prototypes are proposed (Vray et al., 2024). Estimating the target data distribution helps with negative transfer reduction from the partial classes shift (Lee et al., 2022). Feature decomposition adaptively identifies the target-private data (Qu et al., 2024). Alternatively, some works focus on enhancing the discriminating ability for out-of-distribution classes (Samadh et al., 2023). Two strategies are available: Expanding inter-class distance to prevent semantic confusion (Tang et al., 2021; Liang et al., 2020), and introducing external semantics, e.g., CLIP, to match distribution (Shu et al., 2022).

Most recently, multimodal large models such as CLIP and ShareGPT (Chen et al., 2024b) have been introduced to advance SFDA, substantially enhancing adaptation performance by leveraging generic knowledge in these models. Two strategies are adopted: (1) customizing and adapting this knowledge via mutual knowledge distillation (Tang et al., 2024c; Zhang et al., 2025) or multi-teacher

guidance (Chen et al., 2024a) or multimodal space alignment (Chen et al., 2026); and (2) further denoising ViL predictions (Tang et al., 2025).

Critically, a key aspects that make CausalDA conceptually distinct from the existing methods. That is, CausalDA is unique in its design philosophy: it tries to learn a causal representation, in contrast to the statistical knowledge that previous methods are designed to learn. As a consequence, the outreach and generalization can be further extended, leading to the first dedicated unified SFDA approach.

3 METHODOLOGY

Problem statement of unified SFDA. SFDA aims to transfer a model pre-trained on the source domain to a different but related target domain without labeling. Formally, let $\mathcal{X}_s = \{\boldsymbol{x}_i^s\}_{i=1}^{n_s}$ and $\mathcal{Y}_s = \{y_i^s\}_{i=1}^{n_s}$ be the source samples and their truth labels. Similarly, the unlabeled samples and the truth target labels are $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{n}$, respectively. The unified SFDA task is to learn a target model $f_{\theta_t} : \mathcal{X} \to \mathcal{Y}$ *without any prior knowledge of the target domain not the relationships between the source and target domains (see Tab. 1)*, given (i) the source model f_{θ_s} pre-trained over $\mathcal{X}_s$, $\mathcal{Y}_s$, and (ii) the unlabeled target data $\mathcal{X}$.

3.1 CAUSALITY IN TRANSFER LEARNING WITH LABELED SOURCE DOMAIN

In transfer learning with available source domain, such as DG and UDA, the causality relation hidden in the statistical dependence (between raw image $\boldsymbol{x}$ and available label $\boldsymbol{y}$) can be summarized to a Structural Causal Model (SCM). As depicted in Fig. 1(a), $\boldsymbol{x}$ is generated jointly by non-causal factors U and causal factors S, while label $\boldsymbol{y}$ is determined by S alone. Here, non-causal factors collectively represent the latent variables that generate the category-independent appearance of images, e.g., different domain styles and spurious dependence (Fig. 2); The causal factors present the ones determining the classification, e.g., the object shape.

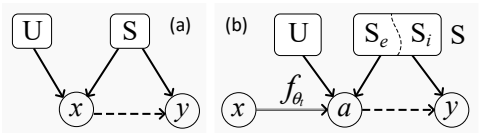


Figure 1: SCM differences. **(a)** SCM for transfer learning with source domain, where S and U represent causal factors (such as the shape/structure of a dog, see Fig. 2) and non-causal factors (Domain shift), forming the observation x. **(b)** Our SFDA-specific SCM in the logit $\boldsymbol{a}$ space, where the causal factors S are decomposed into the external S_e and internal S_i components.

Solving the transfer learning issue involves eliminating the domain shift by building a robust classification function, i.e., $P(\boldsymbol{y}|S)$. From the view of SCM, we need to disconnect U from $\boldsymbol{x}$. So existing works impose intervention upon $\boldsymbol{x}$, according to the intervention theory (Pearl, 2009). Formally, this scheme can be formulated as $P(\boldsymbol{y}|S) = P(\boldsymbol{y}|do(x), S)$, where $do(\cdot)$ means the do-operation of imposing intervention upon the input variables.

Typically, $P(\boldsymbol{y}|S)$ is addressed by: When we change the probability distribution of non-causal factors (intervention) whilst keeping the category (prediction consistency), the causality can be extracted. Often, domain-related data augmentation is adopted as an intervention, whilst the source labels ensure prediction consistency. However, both conditions are unavailable in SFDA, making it inapplicable.

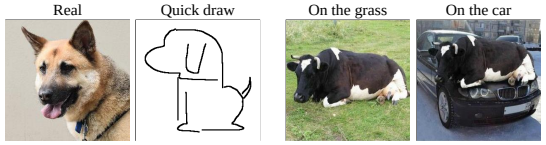


Figure 2: Domain shift illustration. Left: **Different appearance styles** of dog; Right: **Spurious dependence of background:** Cow in grass ground vs. on a car.

3.2 FORMULATING CAUSALITY IN UNIFIED SFDA

Instead of extracting causality in the raw input space $\boldsymbol{x}$, we shift our focus to the logit space $\boldsymbol{A}$: $\boldsymbol{a} = f_{\theta_t}(\boldsymbol{x}) \in \boldsymbol{A}$. The choice of this space for causality analysis is crucial for two main reasons: First and foremost, from a causal perspective, using a post-intervened space enhances causality capture by incorporating a probability distribution shift into the data (Mahajan et al., 2021; Wang et al., 2022; Pearl, 2009). In our case, the logit space is post-intervened due to the inherent distribution shift

between the in-training target model (initialized with the pre-trained source domain model) and the target domain. Thus, the target model's logits for the target domain data reflect this domain shift, making them a particular form of post-intervention in this context.

Furthermore, the logit space is both highly semantic and compact, making it more efficient to manipulate. Since logits encode rich semantic information about class relationships, operating in this space allows for better generalization and transferability across domains. Additionally, its compact representation reduces computational complexity, facilitating more efficient learning and adaptation. These properties collectively make the logit space a more effective choice for our approach. Tab. 2 presents an analogy between CausalDA and the existing causality-inspired learning approach.

Table 2: Analogy of causality-inspired learning paradigms: existing methods, e.g., (Chen & Bühlmann, 2021; Christiansen et al., 2021; Wang et al., 2022), versus CausalDA.

Approach	Intervention	Consistency
Existing methods	Data augmentation	Supervised
CausalDA (Ours)	Using logit	Unsupervised

Our SCM is illustrated in Fig. 1(b). It models both non-causal factors U and causal factors S within a latent space A. Notably, the causal component S is further decomposed into internal (S_i) and external (S_e) sub-components to facilitate a more complete and interpretable modeling of causality.

This internal-external causality decomposition is motivated by the recognition that identifying all relevant causal relationships solely from the source and target domains is often infeasible due to limited training data and domain-specific knowledge. To address this, we introduce a pretrained ViL model, such as CLIP, as an external knowledge source to aid in the discovery of causal relationships. This design leads to two types of causality: (i) **Internal causality**, which refers to causal factors that can be readily inferred from the source domain model and target domain data; and (ii) **External causality**, which denotes the causal relations that are nearly intractable to detect from target domain observations alone, thereby requiring an external semantic prior (e.g., a ViL model with broad, many-domain information) to be uncovered. Such a collaborative design enables a more holistic understanding of causal structures, as empirically validated in `Sec.4.3` and `Appendix-G.1`.

Unlike conventional intervention-driven methods, we focus on isolating S from the entangled U as

$$P(\boldsymbol{y}|S) = P(\boldsymbol{y}|dis(S|\{U,S\})),\ S=\{S_i,S_e\}, \tag{1}$$

where $dis(S|\{U,S\})$ means conducting discover operation on $\{U,S\}$ to decompose S. To make this process computationally tractable, we consider

Principle 1 ***(Peters et al., 2017)****: The conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms.*

Inspired by Principle 1, we assume conditional independence between external and internal causal factors to enable a divide-and-conquer learning approach. Under this perspective, we propose a factorized approximation as $P(\boldsymbol{y}|S) \approx P(\boldsymbol{y}|dis(S_e|\{U,S_e\})) \cdot P(\boldsymbol{y}|dis(S_i|\{U,S_i\}))$.

Our formulation of causality in SFDA typically transforms the problem from conventional U removal into passively capturing S. Although this provides a causal perspective for SFDA, directly mining S lacks a practical solution. To bridge this gap, we propose a proxy learning strategy of invariant factors to approximate the capture of S. The following section elaborates on our design.

3.3 CausalDA design

Overview. As depicted in Fig. 3, our discovery consists of `phase-1` for external invariant factors D_e (a proxy of S_e) and `phase-2` for internal ones D_i (a proxy of S_i), as detailed below.

3.3.1 External invariant factor discovery

To obtain external causality for a target task, we explore the potential of recent large multimodal ViL models such as CLIP (Radford et al., 2021). This is because they have accumulated a vast amount of knowledge and exhibited strong generalization abilities across a wide range of perception tasks.

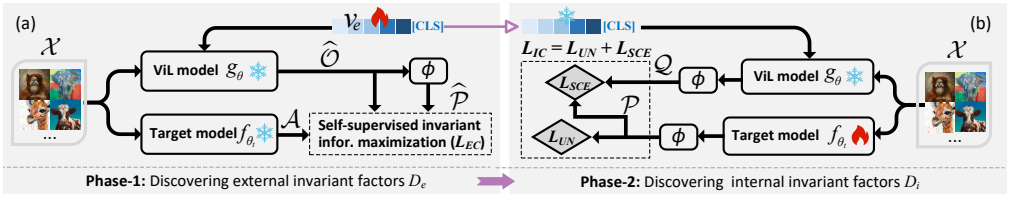


Figure 3: Overview of our CausalDA framework: (a) `Phase-1`: Discovering external invariant factors D_e in form of prompt context $\boldsymbol{v}_e$ from a frozen ViL model g_θ using our self-supervised invariant information maximization ($\mathcal{L}_{EC}$); (b) `Phase-2`: The target model f_{θ_t} discovers internal invariant factors D_i ($\mathcal{L}_{UN}$), while classification based on the updated $\boldsymbol{v}_e$-associated predictions from g_θ preserves the discovered D_e within f_{θ_t} ($\mathcal{L}_{SCE}$).

Specifically, we accomplish this idea by encoding D_e into the prompt context $\boldsymbol{v}_e$ of a ViL model g_θ efficiently. That is, our learning target is $\boldsymbol{v}_e$, which is the explicit expression of D_e. It enables effective integration of external knowledge into the learning process for our successive discovery. More concretely, we start by formulating the D_e discovery as maximizing the correlation between a random variable representing D_e and the prediction of the target model. Then, the original formulation is converted to a self-supervised invariant information maximization problem with a theoretical guarantee, followed by the deep learning instantiation.

With the target samples $\mathcal{X}$ and learnable $\boldsymbol{v}_e$, we obtain the prediction of g_θ as $\hat{\mathcal{P}} = \{\hat{\boldsymbol{p}}_i\}_{i=1}^n$. Meanwhile, we obtain the logits of target model f_{θ_t} for the same data as $\mathcal{A} = \{\boldsymbol{a}_i\}_{i=1}^n$. We introduce three *random variables*, V, Y, and Z, following the probability distributions of $\boldsymbol{v}_e$, $\hat{\mathcal{P}}$ and $\mathcal{A}$, respectively. In CausalDA, we further write $Z = \{Z_e, Z_i\}$ where the latent random variables Z_e and Z_i represent the external D_e and the internal D_i invariant factors, respectively.

Formally, we discover D_e by maximizing mutual information $\max_{\boldsymbol{v}_e} I(Z_e, V)$, where $I(\cdot,\cdot)$ is mutual information of two random variables. Consider that prompt V is specific to the ViL model g_θ, we further transform $\max_{\boldsymbol{v}_e} I(Z_e, V)$ as

$$\max_{\boldsymbol{v}_e} I(Z_e, Y), Y = \phi(g_\theta(V)), \tag{2}$$

where $\phi(\cdot)$ is softmax function; so that the two variables under correlation maximization reside in similar semantic spaces. Theoretically, we can prove that the transformation above is reasonable according to **Lemma** 1 below (see the proof in `Appendix-A.1`).

Lemma 1 *Given random variables $\{Z_1, X_1, Y_1\}$ where X_1, Y_1 satisfy $f_1 : X_1 \mapsto Y_1$. When f_1 is compressed, i.e., the output's dimension is smaller than the input's, then $I(Z_1, X_1) \geq I(Z_1, Y_1)$.*

Given ViL model g_θ maps a high-dimensional image to a low-dimensional category vector, and let $(Z_1, X_1, Y_1) = (Z_e, X, Y)$ and $f_1 = \phi(g_\theta(\cdot))$, then $I(Z_e, V) \geq I(Z_e, Y)$ by **Lemma** 1.

Self-supervised invariant information maximization. To optimize latent Z_e and Y, we reformulate the objective using observable variables. By introducing a random variable $Z' = g_\theta(V)$ for the ViL's logits $\hat{\mathcal{O}} = \{\hat{\boldsymbol{o}}_i\}_{i=1}^n$ between V and Y, we form a dual information maximization expression as:

$$\max_{\boldsymbol{v}_e} I(Z, Z') + I(Z', Y), \tag{3}$$

where Z is the logit variable of target model f_{θ_t} which is observed. In Eq. (3), the first term aligns the generic external knowledge of the ViL model (Z') with the target domain representations (Z), ensuring that the extracted Z_e is specifically tailored to the target domain. The second term is intractable, as Z' is a high-dimensional continuous neural representation and requires estimating the unknown joint distribution $p(Z', Y)$. Furthermore, the deterministic mapping $Y = \phi(Z')$ causes the mutual information to degenerate into an entropy term, introducing optimization instability. Hence, we optimize a tractable lower bound of $I(Z', Y)$ (the proof is given in `Appendix-A.2`).

Theorem 1 *Suppose that there are five random variables Z, V, Z', Y and Y'. Among them, Z represents the target domain knowledge; V, Z' and Y express the input prompt, intermediate features of the ViL model, and predictions, respectively; Y' depicts a pseudo-label that has a learnable*

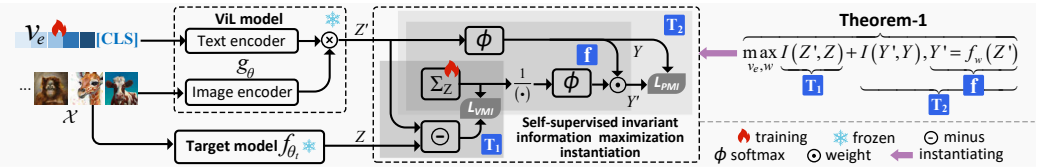


Figure 4: Realizing our self-supervised invariant information maximization w.r.t. **Theorem** 1. T_1 is estimated by Variational Mutual Information (VMI) with Gaussian distribution assumption $\mathcal{N}(Z, \Sigma_Z)$. T_2 is computed by Probabilistic Mutual Information (PMI), where the function f_w generates self-supervised guidance Y', employing the covariance matrix Σ_Z from T_1 as weighting parameters w.

functional relationship f_w (w is parameters) with Z', i.e., $Y' = f_w(Z')$. For the mutual information maximization objective, we have the following lower bound:

$$\max_{\boldsymbol{v}_e} I\left(Z, Z'\right) + I\left(Z', Y\right) \geq \max_{\boldsymbol{v}_e, w} \underbrace{I\left(Z, Z'\right)}_{\mathrm{T}_1} + \underbrace{I\left(Y', Y\right), Y' = f_w\left(Z'\right)}_{\mathrm{T}_2}. \tag{4}$$

We realize the proposed invariant information maximization according to the pipeline illustrated in Fig. 4. The instantiations of T_1 and T_2 are elaborated below.

As exactly computing mutual information of two vector variables is intractable, we maximize the variational lower bound for the term T_1 (see the proof in (Barber & Agakov, 2004)):

$$I\left(Z, Z'\right) = \mathbb{E}_{Z,Z'}\left[\log \frac{q(Z'|Z)}{p\left(Z'\right)}\right] + \mathrm{KL}\left(q(Z'|Z) \| p\left(Z'|Z\right)\right) \geq H\left(Z'\right) + \mathbb{E}_{Z,Z'}\left(\log q\left(Z'|Z\right)\right), \tag{5}$$

where $q(Z'|Z)$ is variational distribution approximating the real distribution $p(Z'|Z)$, $\mathrm{KL}(\cdot,\cdot)$ is KL-divergence function, and $H(Z')$ is a constant. In practice, we model $q(Z'|Z)$ as a Gaussian distribution $\mathcal{N}(Z, \Sigma_Z)$ with mean Z and diagonal covariance matrix Σ_Z. For maximizing T_1, we adopt the Variational Mutual Information (VMI) loss (see $\mathbf{T}_1$ in Fig. 4):

$$\mathcal{L}_{\mathrm{VMI}}\left(\mathcal{X}; \boldsymbol{v}_e, \Sigma_Z\right) = \mathbb{E}_{Z,Z'}\left(||Z' - Z||^2_{\Sigma_Z^{-1}} + \log|\Sigma_Z|\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{o}}_i - \boldsymbol{a}_i\right)^T \Sigma_Z^{-1}\left(\hat{\boldsymbol{o}}_i - \boldsymbol{a}_i\right) + \log|\Sigma_Z^i|. \tag{6}$$

Consider that the ViL's prediction Y and the pseudo-label Y' are probability distributions, we compute the T_2 term by vanilla mutual information (Ji et al., 2019): $I(Z, Z') = p(Z, Z')/[p(Z)p(Z')]$. The corresponding Probabilistic Mutual Information (PMI) loss function is formed as Eq. (7), where $\sum_{c=1}^{C} \boldsymbol{p}'_{i,c} = 1$; $f_w(\cdot)$ is a weighting-based function needed to be constructed.

$$\mathcal{L}_{\mathrm{PMI}}\left(\mathcal{X}; \boldsymbol{v}_e, \Sigma_Z\right) = \frac{1}{n}\sum_{i=1}^{n} I\left(\hat{\boldsymbol{p}}_i, \boldsymbol{p}'_i\right), \text{with}\, \hat{\boldsymbol{p}}_i = \phi(\hat{\boldsymbol{o}}_i),\ \ \boldsymbol{p}'_i = f_w(\hat{\boldsymbol{o}}_i), \tag{7}$$

The f_w construction lies in satisfying three properties: (i) The output is a probability distribution. (ii) It is inherently an information-distilling mapping that satisfies the Data Processing Inequality (DPI), ensuring that redundant non-causal features can be eliminated. (iii) It explicitly functions as a filter, actively discarding unstable domain noises to isolate and distill invariant semantics.

In practice, the softmax projection (ϕ) simultaneously satisfies the first two properties by mapping logits Z' into a probability simplex through a many-to-one, non-invertible transformation. This discards redundant logit-scale information and enforces the information loss required by DPI, making f_w an effective information-distilling mapping. For the third property, we couple the weighting parameters w with $I(Z', Z)$, which is formulated as a Σ_Z-weighted MSE (Eq. (6)) to capture domain sensitivity. High-variance dimensions are suppressed as domain-specific noise, while stable dimensions are preserved as invariant semantics. Thus, inverse-variance weighting filters domain artifacts and yields purified invariant guidance in Y'.

Inspired by the analysis above, we take $1/\Sigma_Z$ as the weighting parameters of f_w, thereby shifting the focus of predicting to the causal component S_e (see $\mathbf{f}$ in Fig. 4), formally,

$$\boldsymbol{p}'_i = f_{\left[w=\Sigma_Z^i\right]}\left(\hat{\boldsymbol{o}}_i\right) = \phi\left(\phi\left(\frac{1}{\Sigma_Z^i}\right) \odot \phi\left(\hat{\boldsymbol{o}}_i\right)\right), \tag{8}$$

where $\phi(\cdot)$ is softmax function. Combining Eq. (6), (7), and (8) together, our self-supervised objective for external causality discovery is formed as

$$\max_{\boldsymbol{v}_e, \Sigma_Z} \mathcal{L}_{\mathrm{EC}}\left(\mathcal{X}; \boldsymbol{v}_e, \Sigma_Z\right) = \mathcal{L}_{\mathrm{PMI}}\left(\mathcal{X}; \boldsymbol{v}_e, \Sigma_Z\right) + \alpha\mathcal{L}_{\mathrm{VMI}}\left(\mathcal{X}; \boldsymbol{v}_e, \Sigma_Z\right). \tag{9}$$

3.3.2 INTERNAL INVARIANT FACTOR DISCOVERY

In `phase-2`, we encode D_i in the target model, as it is latent in nature. Similar to capturing the external D_e (Eq. (2)), this encoding can be formulated by maximizing conditional mutual information:

$$\max_{\theta_t} I\left(Z_i, Z_\theta\right), \ \text{s.t.} \ \max_{\theta_t} I\left(Z_e^*, Z_\theta\right) \tag{10}$$

where random variable Z_θ represents target model parameters θ_t, whilst Z_i denotes random variable of internal invariant factors D_i; Z_e^* represents the currently discovered external invariant factor D_e^*. The primary term, $\max_{\theta_t} I\left(Z_i, Z_\theta\right)$, aims to capture the internal D_i. The condition, $\max_{\theta_t} I\left(Z_e^*, Z_\theta\right)$, ensures the target model simultaneously accommodates the discovered external factors D_e^*.

Crucially, theoretical proofs demonstrate that mutual information between inputs and outputs is maximized when outputs are both highly confident and exhibit balanced class distributions (Krause et al., 2010). Inspired by this finding, we maximize $I\left(Z_i, Z_\theta\right)$ in Eq. (10) by minimizing:

$$\mathcal{L}_{\text{UN}}(\mathcal{P}; \theta_t) = -\sum_{i=1}^{n} \boldsymbol{p}_i \log \boldsymbol{p}_i + \tau \sum_{c=1}^{C} \text{KL}\left(\varrho_c \| \frac{1}{C}\right), \ \text{with} \ \boldsymbol{p}_i \in \mathcal{P} = \phi(f_{\theta_t}(\mathcal{X})). \tag{11}$$

In this equation, τ is a trade-off parameter; the first term minimizes the information entropy of each prediction $\boldsymbol{p}_i = [p_{i,c}]_{c=1}^{C}$. The KL divergence term regulates the predictive balance across all C categories via pulling the empirical distribution $\varrho_c = \frac{1}{n}\sum_{i=1}^{n} p_{i,c}$ towards a uniform distribution.

Additionally, we convert factors D_e^* from its prompt format $\boldsymbol{v}_e$ to pseudo-labels by evaluating the ViL model g_θ: $\mathcal{Q} = \phi(g_\theta(\mathcal{X}, \boldsymbol{v}_e))$. This conversion simplifies model optimization, as it enables the use of the conventional softmax cross-entropy loss, $\mathcal{L}_{\text{SCE}}\left(\mathcal{P}, \mathcal{Q}; \theta_t\right)$. Combining these elements, we implement the optimization in Eq. (10) using the following objective, where σ serves as a trade-off parameter. More details for model training are presented in `Appendix-B`.

$$\mathcal{L}_{\text{IC}} = \min_{\theta_t} \mathcal{L}_{\text{UN}}(\mathcal{P}; \theta_t) + \sigma \mathcal{L}_{\text{SCE}}\left(\mathcal{P}, \mathcal{Q}; \theta_t\right) \tag{12}$$

4 EXPERIMENTS

Datasets and implementation details are detailed in `Appendix-C` and `-D`, respectively.

4.1 MAIN RESULTS ON MULTIPLE SFDA SETTINGS

Competitors. To evaluate CausalDA, we select 21 state-of-the-art methods in four groups. *(i) The first* is the source model and CLIP that lay the basis for CausalDA. *(ii) The second* includes 13 state-of-the-art SFDA methods: SHOT (Liang et al., 2020), GKD (Tang et al., 2021), NRC (Yang et al., 2021a), AaD (Yang et al., 2022), AdaCon (Chen et al., 2022a), CoWA (Lee et al., 2022), PLUE (Litrico et al., 2023), TPDS (Tang et al., 2024a), HCL (Huang et al., 2021), CRS (Zhang et al., 2023), MMGA (Chen et al., 2026), DIFO (Tang et al., 2024c), and ProDe (Tang et al., 2025). *(iii) The third* includes two Generalized SFDA methods: GDA (Yang et al., 2021b) and PSAT (Tang et al., 2024b). *(iv) The fourth* includes four prompt learning methods: CoOP (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), TPT (Shu et al., 2022), and ProGrad (Zhu et al., 2023).

For comprehensive comparisons, we implement two variants, CausalDA-C-RN and CausalDA-C-B32, based on CLIP image encoders with ResNet and ViT-B/32 (Han et al., 2022), respectively. Also, we upgrade some typical comparisons with ViT-B/32 (see Networks in `Appendix-D` for details).

Quantitative comparison. In the *Closed-set SFDA setting* (see Tab. 3), both CausalDA-C-RN and CausalDA-C-B32 achieve the best average accuracy, with margin gains over previous methods. The full results and further analysis are provided in `Appendix-F.1`. In the one-step adaptation task of *Generalized SFDA setting* (see Tab. 4, the evaluation protocol is introduced in `Appendix-E.1`), among the "-B32" methods, CausalDA-C-B32 (without any anti-forgetting design) defeats the best specially designed model PSAT-B32 (with an anti-forgetting mechanism) with a lead of **0.3**% in harmonic mean accuracy. CausalDA-C-RN outperforms other alternatives with the same backbone. The continual adapting results are provided in `Appendix-F.2`. In the *Open-set & Partial-set SFDA setting* (see Tab. 5), CausalDA-C-B32 achieves the highest average accuracy, outperforming all competing methods by margins of **5.9**% and **10.7**% over AaD-B32 in the Partial-set setting and

Table 3: Closed-set results (%) on **OfficeHome** (OH), VisDA, and **DomainNet-126** (DN126). SF, ViL means: source-free, ViL model-required.

Method	Venue	SF	ViL	OH	VisDA	DN126
Source	–	–	–	59.2	49.2	54.7
CLIP-RN	ICML21	–	✓	72.1	72.1	72.7
CLIP-B32	ICML21	–	✓	79.2	79.2	79.2
SHOT	ICML20	✓	✗	71.8	82.9	68.1
NRC	NeurIPS21	✓	✗	72.3	85.9	67.5
AdaCon	CVPR22	✓	✗	65.0	86.8	65.4
CoWA	ICML22	✓	✗	72.4	86.9	68.6
PLUE	CVPR23	✓	✗	66.9	88.3	64.7
TPDS	IJCV24	✓	✗	73.5	87.6	67.1
SHOT-B32	ICML20	✓	✓	79.5	85.2	76.6
NRC-B32	NeurIPS21	✓	✓	79.5	88.2	76.7
AaD-B32	NeurIPS22	✓	✓	79.5	82.6	76.9
TPDS-B32	IJCV24	✓	✓	79.7	78.0	77.0
MMGA-B32	PR26	✓	✓	80.2	89.2	77.4
CausalDA-C-RN	–	✓	✓	77.6	89.3	78.0
CausalDA-C-B32	–	✓	✓	**84.3**	**90.3**	**82.2**

Table 4: Generalized SFDA results (%) on **Office-Home**. S, T, H are the results on the source and target domains, and harmonic mean accuracy, respectively; WAD is With Anti-forgetting Design.

Method	Venue	WAD	Avg.		
			S	T	H
SHOT	ICML20	✗	84.2	71.8	77.1
NRC	NeurIPS21	✗	91.3	72.3	80.4
CoWA	ICML22	✗	91.8	72.4	80.6
PLUE	CVPR23	✗	96.3	66.9	78.4
TPDS	IJCV24	✗	83.8	73.5	78.0
GDA	ICCV21	✓	80.0	70.2	74.4
PSAT-RN	TMM24	✓	85.3	72.6	78.4
SHOT-B32	ICML20	✗	81.9	79.5	80.7
NRC-B32	NeurIPS21	✗	81.1	78.6	79.8
TPDS-B32	IJCV24	✗	88.9	79.7	84.1
PSAT-B32	TMM24	✓	89.5	80.7	84.9
MMGA-B32	PR26	✗	86.9	80.2	83.4
CausalDA-C-RN	–	✗	85.0	77.6	80.7
CausalDA-C-B32	–	✗	86.1	84.3	**85.2**

Table 5: Partial-/Open-set (%) on **Office-Home**.

Partial-set SFDA			Open-set SFDA		
Method	Venue	Avg.	Method	Venue	Avg.
SHOT	ICML20	79.3	SHOT	ICML20	72.8
HCL	NeurIPS21	79.6	HCL	NeurIPS21	72.6
CoWA	ICML22	83.2	CoWA	ICML22	73.2
AaD	NeurIPS22	79.7	AaD	NeurIPS22	71.8
CRS	CVPR23	80.6	CRS	CVPR23	73.2
SHOT-B32	ICML20	80.6	SHOT-B32	ICML20	73.3
AaD-B32	NeurIPS22	81.2	AaD-B32	NeurIPS22	72.4
CausalDA-C-RN	–	82.9	**CausalDA**-C-RN	–	79.6
CausalDA-C-B32	–	**87.1**	**CausalDA**-C-B32	–	**84.0**

Table 6: SF-OODG results (%) on **ImageNet variants**. L and ViL means label-required and ViL model-required, respectively.

Method	Venue	L	ViL	ImageNet→X				
				IN-V2	IN-K	IN-A	IN-R	Avg.
Source	–	–	–	62.7	22.2	0.7	35.1	30.2
CLIP-RN	ICML21	–	✓	51.5	33.3	21.8	56.1	40.7
CLIP-B32	ICML21	–	✓	54.8	40.8	29.5	66.2	47.8
CoOP-RN	IJCV21	✓	✓	55.4	34.7	23.1	56.6	42.4
CoOP-B32	IJCV21	✓	✓	58.2	41.5	31.3	65.8	49.2
CoCoOp-RN	CVPR22	✓	✓	55.7	34.5	23.3	57.7	42.8
CoCoOp-B32	CVPR22	✓	✓	56.6	40.7	30.3	64.1	47.9
TPT-RN	NeurIPS22	✗	✓	54.7	35.1	26.7	59.1	43.9
ProGrad-RN	ICCV24	✓	✓	54.7	34.4	23.1	56.8	42.2
ProGrad-B32	ICCV24	✓	✓	57.4	41.7	**31.9**	66.5	49.4
SHOT	ICML20	✗	✗	62.5	38.4	2.1	42.7	36.4
NRC	NeurIPS21	✗	✗	61.5	0.9	1.3	28.9	23.1
AdaCon	CVPR22	✗	✗	50.9	18.5	2.0	38.6	27.5
CoWA	ICML22	✗	✗	62.7	37.8	1.7	46.3	37.1
PLUE	CVPR23	✗	✗	53.0	18.5	2.0	38.6	28.0
TPDS	IJCV24	✗	✗	62.9	35.0	2.9	48.4	37.3
SHOT-B32	ICML20	✗	✓	**69.4**	43.6	1.6	58.2	43.2
NRC-B32	NeurIPS21	✗	✓	66.2	18.6	15.4	38.0	34.6
TPDS-B32	IJCV24	✗	✓	69.0	38.9	0.6	58.7	41.8
MMGA-B32	PR26	✗	✓	65.8	39.1	18.6	63.5	46.7
CausalDA-C-RN	–	✗	✓	64.4	42.6	22.3	61.5	47.7
CausalDA-C-B32	–	✗	✓	64.7	**48.4**	30.6	**71.0**	**53.7**

SHOT-B32 in the Open-set setting, respectively. Similarly, it significantly outperforms SHOT-B32 and AaD-B32 under both Partial-set and Open-set settings, indicating the efficacy of the proposed design. As for the *Source-free out-of-distribution generalization setting*, termed SF-OODG, (see Tab. 6), in average accuracy, even compared to CLIP and other CLIP-based methods, CausalDA maintains a clear advantage. For example, CausalDA-C-NR improves upon CLIP-C-RN by **7.0**% and surpasses the second-best CLIP-based method, TPT-RN, by **3.8**%. Among the "-B32" models, CausalDA-C-B32 excels with a margin of at least **10.5**% over previous methods. These results indicate that CausalDA is competitive on all main SFDA settings.

4.2 Unification Study

Evaluation on domain fitness and domain generalization. In this strategic evaluation (all experiments use the strong variant CausalDA-C-B32), we compare CausalDA against statistical DIFO and ProDe both equipped with the same ViL model. Our analysis focuses on two critical modeling aspects, examined across various SFDA settings: *Domain fitness*: Evaluated under Closed-set conditions; *Domain generalization*: Assessed across Generalized, Partial-set, Open-set, and SF-OODG settings.

Based on this evaluation presented in Tab. 7, we draw the following key observations: (1) Domain fitness (Closed-set SFDA): CausalDA is indeed outperformed by both DIFO and ProDe in the Closed-set setting. This is attributed to the inherent statistical learning nature of these alternatives, which tends to overfit their models towards the specific target domain. (2) Under the Generalized SFDA setting, CausalDA demonstrates superior performance in H-accuracy, also exhibiting the least drop in source domain accuracy (see S-accuracy). This is because the captured causality mitigates the risk of forgetting the source domain. (3) Except for the Closed-set scenario, CausalDA consistently stands out with strong performance on the other four settings. Collectively, these results validate that our approach strikes the optimal balance between domain fitness and generalization, thereby laying a solid foundation to serve as a unified SFDA solution.

Table 7: Domain fitness, generalization comparison. OH/DN126: **OfficeHome/DomainNet-126**. S/T: Source/Target domain; H: Harmonic mean accuracy. H_g: Average all settings except closed-set.

Method	Venue	Closed-set				Generalized			Open-set	Partial-set	SF-OODG					
		OH	VisDA	DN126	*Avg.*	S	T	*H*	T	T	IN-V2	IN-K	IN-A	IN-R	*Avg.*	H_g
Source	–	59.2	49.2	54.7	54.4	98.1	59.2	73.8	46.6	62.8	62.7	22.2	0.7	35.1	30.2	53.4
DIFO	CVPR24	83.1	90.3	80.0	84.5	78.0	83.1	80.5	75.9	85.6	59.6	37.7	25.2	66.4	47.2	72.3
ProDe	ICLR25	**84.5**	**91.0**	**85.0**	**86.8**	85.1	84.5	84.8	82.6	84.2	62.1	45.8	24.7	69.7	50.6	75.5
CausalDA	–	84.3	90.3	82.2	85.6	86.1	84.3	**85.3**	**84.0**	**87.1**	**64.7**	**48.4**	**30.6**	**71.0**	**53.7**	**77.5**

Table 8: Unification comparison results (%). H_{all}: Average over all settings; H_{wrg}: The largest percentage shortfall of a method from the per-setting best across all settings; H_{loso}: Omit one setting at a time and average the scores over the rest.

Method	$H_{wrg}(\downarrow)$	$H_{loso}(\uparrow)$					$H_{all}(\uparrow)$
		w/o Closed-set	w/o Generalized setting	w/o Open-set	w/o Partial-set	w/o SF-OODG	
DIFO	12.1	72.3	73.3	74.4	72.0	81.6	74.7
ProDe	5.77	75.5	76.0	76.6	76.2	84.6	77.7
CausalDA	**1.38**	**77.5**	**77.6**	**77.9**	**77.1**	**85.5**	**79.1**

Unification evaluation. To underscore CausalDA's advantage over DIFO and ProDe, we perform a unification evaluation using the metrics detailed in Appendix-E.2. The comparison in Tab. 8 shows that CausalDA attains the highest score in all scenarios. This indicates that CausalDA delivers robust, stable performance across diverse settings. Moreover, it exhibits no apparent weakness in any single scenario (cf. H_{wrg}), and its all-round superiority is not driven by any individual setting (cf. H_{loso}). Together, this evidence supports our claim of being towards a universally adaptable framework.

4.3 Indirect Validation of Causality Based on Invariance Analysis

Validating the discovery of causal factors is significant yet non-trivial in causality learning, especially with latent factors, as in this study. To that end, we adopt a general principle that there exists an intimate connection between invariance and causality useful for generalization (Schölkopf et al., 2021; Arjovsky et al., 2019). That being said, if a learned model enables varying environments generalization, it constitutes *indirect evidence* that the model has captured causal predictors.

Concretely, we simulate a spectrum of viewing environments by imposing varying levels of Gaussian noise with gradually increased kernel sizes from 8 to 20. We compare the competitors without (SHOT) and with (DIFO and ProDe) ViL knowledge (all experiments use the strong variant CausalDA-C-B32). As shown in Fig. 5, all competitors degrade clearly as the noise increases, whilst CausalDA levels off. These findings suggest the method captures invariant predictors that are compatible with causal explanations, though they do not prove causal recovery rigorously. To further examine if external and internal factors both are effective, we further derive a variant of CausalDA, named as CausalDA^{in}, without the external invariant factor discovery. Specifically, in CausalDA^{in}, the pseudo-labels are directly derived from CLIP's zero-shot predictions. This is consistent with external factors contributing to improved invariance, suggesting a possible complementarity with internal counterparts. Further model analyses are provided in Appendix-G.

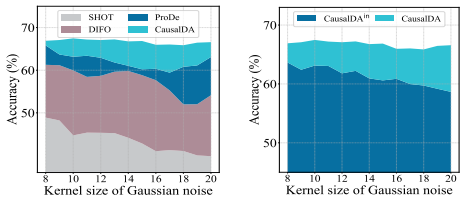


Figure 5: Invariance analysis across varying noise on the Ar→Cl task of **Office-home**.

5 Conclusion

In this paper, we introduce a more challenging Unified SFDA setting that unifies diverse scenarios within a single framework. To tackle this, we propose CausalDA, a causality-inspired approach that differs from previous statistical dependence-based SFDA methods. Specifically, beyond label-dependent non-causal removal, we develop an SFDA-specific causal structure to mine internal and external causal factors. We then approximate this process by discovering invariant factors through internal-external complementary learning. Extensive experiments across SFDA settings and SF-OODG validate the superiority of CausalDA (its limitation is discussed in Appendix-H).

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing and presentation of introduction in this paper. We have not used generative AI tools for generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, and interpret results. We have reviewed all AI-assisted contents in the introduction. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is conducted in accordance with the ICLR Code of Ethics. Our studies focus on the design of malicious client detection in federated learning, especially when the global model initialized with pretrained backbones. All experiments are carried out on publicly available datasets (e.g., *Domain Adaptation Group*: Office-Home, VisDA, and DomainNet-126; *ImageNet Variants Group*: ImageNet-V2, ImageNet-A, ImageNet-R, and ImageNet-Sketch) under their respective licenses, with no involvement of human subjects or sensitive personal data. In principle, this work has no ethical issues since it does not reveal any new security risks and is purely defensive. We report our methods and findings transparently to foster reproducibility and responsible use.

REPRODUCIBILITY STATEMENT

We have provided detailed information on datasets, models, training settings, and computational facilities in our main manuscript and appendix. The code and data will be made available after the publication of this paper.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *rXiv preprint arXiv:2006.10726*, 2019.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Proc. Adv. Neural Inform. Process. Syst.*, 16(320):201, 2004.

Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2985–2994, 2019.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 295–305, 2022a.

Dongjie Chen, Kartik Patwari, Zhengfeng Lai, Sen-ching Cheung, and Chen-Nee Chuah. Empowering source-free domain adaptation with mllm-driven curriculum learning. *arXiv preprint arXiv:2405.18376*, 2024a.

Lijuan Chen, Yunxiang Bai, Ying Hu, Qiong Wang, and Xiaozhi Qi. Source-free domain adaptation via multimodal space-guided alignment. *Pattern Recognition*, 173:112827, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4v: Improving large multi-modal models with better captions. In *Proc. Eur. Conf. Comput. Vis.*, pp. 370–387. Springer, 2024b.

Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, and Yueting Zhuang. Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *Proc. IEEE Int. Intell. Rob. Syst.*, pp. 10185–10192. IEEE, 2022b.

Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *J. Mach. Learn. Resear.*, 22(1):11856–11935, 2021.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10): 6614–6630, 2021.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2006.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7212–7222, 2022.

Yuntao Du, Haiyang Yang, Mingcai Chen, Juan Jiang, Hongtao Luo, and Chongjun Wang. Generation, augmentation, and alignment: A pseudo-source domain based method for source-free domain adaptation. *Mach. Learn.*, 113:3611–3631, 2024.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):87–110, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 15262–15271, 2021b.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Proc. Adv. Neural Inform. Process. Syst.*, 34:3635–3649, 2021.

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9865–9874, 2019.

Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Trans. on Artif. Intell.*, 2(6):508–518, 2021.

Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. *Proc. Adv. Neural Inform. Process. Syst.*, 23, 2010.

Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4544–4553, 2020.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *Int. Conf. Mach. Learn.*, pp. 12365–12377. PMLR, 2022.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9638–9647, 2020a.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9641–9650, 2020b.

Weikai Li and Songcan Chen. Partial domain adaptation without domain alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8787–8797, 2023.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the Int. Conf. Mach. Learn.*, pp. 6028–6039, 2020.

Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for test-time adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Yadan Luo, Zijian Wang, Zhuoxiao Chen, Zi Huang, and Mahsa Baktashmotlagh. Source-free progressive graph learning for open-set domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *Int. Conf. Mach. Learn.*, pp. 7313–7324. PMLR, 2021.

Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 754–763, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. VisDA: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1406–1415, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Sanqing Qu, Tianpei Zou, Lianghua He, Florian Röhrbein, Alois Knoll, Guang Chen, and Changjun Jiang. Lead: Learning decomposition for source-free universal domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 23334–23343, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Int. Conf. Mach. Learn.*, pp. 5389–5400. PMLR, 2019.

Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Proc. Adv. Neural Inform. Process. Syst.*, volume 36, pp. 80396–80413, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Proc. Adv. Neural Inform. Process. Syst.*, 35:14274–14289, 2022.

S Tang, Yuji Shi, Zhiyuan Ma, Jian Li, Jianzhi Lyu, Qingdu Li, and Jianwei Zhang. Model adaptation through hypothesis transfer with gradual knowledge distillation. In *Proc. IEEE/RSJ Int. Conf. Intell. Rob. Syst.*, pp. 5679–5685. IEEE, 2021.

Song Tang, Yan Zou, Zihao Song, Jianzhi Lyu, Lijuan Chen, Mao Ye, Shouming Zhong, and Jianwei Zhang. Semantic consistency learning on manifold for source data-free unsupervised domain adaptation. *Neural Networks*, 152:467–478, 2022.

Song Tang, An Chang, Fabian Zhang, Xiatian Zhu, Mao Ye, and Changshui Zhang. Source-free domain adaptation via target prediction distribution searching. *Int. J. Comput. Vis.*, 132(3):654–672, 2024a.

Song Tang, Yuji Shi, Zihao Song, Mao Ye, Changshui Zhang, and Jianwei Zhang. Progressive source-aware transformer for generalized source-free domain adaptation. *IEEE Trans. on Multime.*, 26:4138–4152, 2024b. doi: 10.1109/TMM.2023.3321421.

Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 23711–23720, 2024c.

Song Tang, Wenxin Su, Yan Gan, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Proxy denoising for source-free domain adaptation. In *Int. Conf. Learn. Represent.*, 2025. DOI: https://arxiv.org/abs/2406.01658.

Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. VDM-DA: Virtual domain modeling for source data-free domain adaptation. *IEEE Trans. Circuits Syst. Video Technol.*, 32(6):3749–3760, 2021.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5385–5394, 2017.

Guillaume Vray, Devavrat Tomar, Behzad Bozorgtabar, and Jean-Philippe Thiran. Distill-SODA: Distilling self-supervised vision transformer for source-free open-set domain adaptation in computational pathology. *IEEE Trans. Med. Imag.*, 43(5):2021–2032, 2024.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Proc. Adv. Neural Inform. Process. Syst.*, 32, 2019.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. Causal representation learning for out-of-distribution recommendation. In *Proc. ACM Web Conf.*, pp. 3562–3571, 2022.

Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Proc. Adv. Neural Inform. Process. Syst.*, 34: 29393–29405, 2021a.

Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8978–8987, 2021b.

Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. *Proc. Adv. Neural Inform. Process. Syst.*, 35:5802–5815, 2022.

Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, Shangling Jui, and Jian Yang. Trust your good friends: Source-free domain adaptation by reciprocal neighborhood clustering. *IEEE rans. Pattern Anal. Mach. Intell.*, 2023.

Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Source-free domain adaptation guided by vision and vision-language pre-training. *Int. J. Comput. Vis.*, 133(2):844–866, 2025.

Yixin Zhang, Zilei Wang, and Weinan He. Class relationship embedded learning for source-free unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7619–7629, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022b.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 15659–15669, 2023.

A PROOFS OF THEORETICAL RESULTS

A.1 A PROOF OF LEMMA 1

Restatement of Lemma 1 *Given random variables $\{Z_1, X_1, Y_1\}$ where X_1, Y_1 satisfy $f_1 : X_1 \mapsto Y_1$. When f_1 is compressed, i.e., the output's dimension is smaller than the input's, then $I(Z_1, X_1) \geq I(Z_1, Y_1)$.*

Proof. Since mapping f_1 is compressed, the entropy of X_1, Y_1 have $H(X_1) \geq H(Y_1)$. As for $I(Z_1, X_1)$, we have the relationship below.

$$\begin{aligned} I(Z_1, X_1) &\geq I(Z_1, X_1) - H(Z_1|Y_1) - I(X_1, Y_1) \\ &\geq [H(Z_1) + H(X_1) - H(Z_1, X_1)] - H(Z_1|Y_1) \\ &\quad - [H(X_1) + H(Y_1) - H(X_1, Y_1)] \\ &= H(Z_1) - H(Z_1|Y_1) - H(Y_1) + H(X_1, Y_1) \\ &\geq I(Z_1, Y_1) - H(Y_1) + \max(H(X_1), H(Y_1)) \\ &\geq I(Z_1, Y_1) - H(Y_1) + H(X_1) \\ &\geq I(Z_1, Y_1) - H(Y_1) + H(Y_1) \\ &= I(Z_1, Y_1) \end{aligned} \tag{13}$$

A.2 A PROOF OF THEOREM 1

Lemma 2 *(Data Processing Inequality, DPI) (Cover & Thomas, 2006) Given random variables X, Y and Z. If there is a Markov chain structure $X \to Y \to Z$, then $I(X, Y) \geq I(X, Z)$, which suggests that the mutual information between the prediction X and the post-processed pseudo-label Z is strictly bounded by the mutual information between X and the original feature Y.*

Restatement of Theorem 1 *Suppose that there are five random variables Z, V, Z', Y and Y'. Among them, Z represents the target domain knowledge; V, Z' and Y express the input prompt, intermediate features of the ViL model, and predictions, respectively; Y' depicts a pseudo-label that has a learnable functional relationship f_w (w is parameters) with Z', i.e., $Y' = f_w(Z')$. For the mutual information maximization objective, we have the following lower bound:*

$$\max_{\boldsymbol{v}_d} I(Z, Z') - I(Z', Y) \geq \max_{\boldsymbol{v}_d, w} \underbrace{I(Z, Z')}_{T_1} + \underbrace{I(Y', Y), Y' = f_w(Z')}_{T_2}. \tag{14}$$

Proof. In our network architecture, both the model prediction Y and the pseudo-label Y' are derived exclusively from the ViL's intermediate feature Z'. Specifically, $Y = \phi(Z')$ and $Y' = f_w(Z')$. Because neither Y nor Y' incorporates any additional external information other than what is provided by Z', they are conditionally independent given Z', namely:

$$P(Y, Y'|Z') = P(Y|Z')P(Y'|Z') \tag{15}$$

According to the information theory, this conditional independence expressed in Eq. (15) strictly defines the Markov chain $Y \to Z' \to Y'$ (Cover & Thomas, 2006). By applying DPI Lemma 2 (above) to this Markov chain, we have

$$I(Y', Y) \leq I(Z', Y) \tag{16}$$

By adding the mutual information term $I(Z, Z')$ to both sides of the inequality, we obtain:

$$I(Z, Z') + I(Y', Y) \leq I(Z, Z') + I(Z', Y) \tag{17}$$

By taking the maximum over the prompt context v_e and the network parameters w, we establish the following tractable lower bound for our original causal discovery objective, with $Y' = f_w(Z')$.

$$\max_{v_e} \big[I(Z, Z') + I(Z', Y)\big] \geq \max_{v_e, w} \big[I(Z, Z') + I(Y', Y)\big]. \tag{18}$$

Algorithm 1 Training CausalDA

Input: Unlabelled target data $\mathcal{X}$, pre-trained source model f_{θ_s}, frozen ViL model g_θ, training iteration number M, prompt context $\boldsymbol{v}_e$.
Initialization: Initialize target model f_{θ_t} (by f_{θ_s}) and $\boldsymbol{v}_e$(by fixed template)

1: **for** iter $k = 1$ to M **do**
2: Sample a mini-batch $\mathcal{X}_b$ from $\mathcal{X}$.
3: Discover external invariant factors D_e: update $(\boldsymbol{v}_e, \Sigma_Z)$ by optimizing $\mathcal{L}_{\text{EC}}(\mathcal{X}_b; \boldsymbol{v}_e, \Sigma_Z)$, as fixing the loss $\mathcal{L}_{\text{IC}}$.
4: Convert updated $\boldsymbol{v}_e$ to target pseudo-labels for this batch $\mathcal{X}_b$, i.e., obtain $\mathcal{Q}_b = \phi(g_\theta(\mathcal{X}_b, \boldsymbol{v}_e))$.
5: Discover internal invariant factors D_i: update f_{θ_t} by optimizing $\mathcal{L}_{\text{IC}}$, fixing the loss $\mathcal{L}_{\text{EC}}$.
6: **end for**
7: **return:** The adapted model f_{θ_t}.

B Model Training

We train CausalDA in an alternating manner. In each iteration, we first optimize prompt context $\boldsymbol{v}_e$ with $\mathcal{L}_{\text{EC}}$ whilst the target model is frozen (i.e., `phase-1`), followed by training the target model by $\mathcal{L}_{\text{IC}}$ (i.e., `phase-2`) while the prompt context is frozen. The ViL model is always frozen throughout. More details for training are presented in Alg. 1

C Dataset

C.1 Domain adaptation group

The involved datasets include three challenging benchmarks widely used in the domain adaptation task. In our evaluations, it serves for Closed-set, Partial-set, Open-set, and Generalized SFDA settings. **Office-Home** (Venkateswara et al., 2017) is a middle-scale dataset. It contains 15k images belonging to 65 categories from working or family environments, being divided into four domains, i.e., **Ar**tistic, **Cl**ip Art, **Pr**oduct, and **R**eal-**w**ord. **VisDA** (Peng et al., 2017) is a large-scale dataset. It includes 12 types of **S**ynthetic to **R**eal transfer recognition tasks. The source domain contains 152k synthetic images, whilst the target domain has 55k real object images from COCO. **DomainNet-126** (Peng et al., 2019) is another large-scale dataset. As a subset of DomainNet containing 600k images of 345 classes from 6 domains of different image styles, this dataset has 145k images from 126 classes, sampled from 4 domains, **C**lipart, **P**ainting, **R**eal, **S**ketch, as (Peng et al., 2019) identifies severe noisy labels in the dataset.

C.2 ImageNet variants group

The evaluation on the SF-OODG setting is conducted on this group, which consists of four benchmarks. Specifically, **IN-V2** (i.e., ImageNet-V2) (Recht et al., 2019) is an independent set consisting of natural images, collected from different sources, a total of 10k images of 1000 ImageNet categories. **IN-A** (i.e., ImageNet-A) (Hendrycks et al., 2021b) is a challenging set consisting of these "natural adversarial examples" (misclassified by a vanilla ResNet50 (He et al., 2016)), containing 7.5k images of 200 ImageNet categories. **IN-R** (i.e., ImageNet-R) (Hendrycks et al., 2021a) is an ImageNet sub-set with artistic renditions. It contains 30k images covering 200 ImageNet categories. **IN-K** (i.e., ImageNet-Sketch) (Wang et al., 2019) collects black and white sketches from the ImageNet validation set. This dataset contains 50k images from 1000 ImageNet categories.

D Implementation details

Source models. Following (Zhou et al., 2022b), for source-free out-of-distribution generalization (SF-OODG) we adopt the Pytorch built-in ResNet50 model (pre-trained on ImageNet) as the source model, taking ImageNet variations as evaluation datasets. For other settings, the source model is trained on the source domain in a supervised manner, same as (Liang et al., 2020; Tang et al., 2022).

Networks. Our CausalDA involves two parts: the target model and the ViL model. The target model's structure is the same as the source model. Following (Liang et al., 2020; Yang et al., 2023), a target model consists of a feature extractor and a classifier (an FC layer). We adapt the feature extractor, which is pre-trained on ImageNet. For fair comparison, we use ResNet101 as the backbone for VisDA and ResNet50 for the others.

We choose CLIP (Radford et al., 2021) as the ViL model where the image encoder is different according to the implement variants. Specifically, for CausalDA-C-RN, ResNet101 is employed on the VisDA dataset, while ResNet50 is used on the Office-Home and DomainNet-126 datasets. On the other hand, CausalDA-C-B32 adopts ViT-B/32 (Han et al., 2022) as the backbone across all datasets. This extends to all other methods using CLIP. Similarly, all methods using CLIP are marked with the same meaning for a convenience observation.

Training. There are three parameters with CausalDA: α in L_{EC} (Eq. (9)), τ in L_{UN} (Eq. (11)), and σ in L_{IC} (Eq. (12)). For all settings, we adopt the same configuration $(\alpha, \sigma, \tau) = (0.003, 0.4, 1.0)$. Here, since the value of $\mathcal{L}_{\text{VMI}}$ is large, α is set as small as 0.003. For model training, we use a batch size of 64, an SGD optimizer with momentum 0.9, and 15 epochs on all datasets. The learnable prompt template is initialized with 'a photo of a [CLS].' (Radford et al., 2021), where the [CLS] is a specific target class name.

In the implementation, [CLS] is instantiated strictly with known source class names, meaning our approach does not require access to unknown target class names in non-closed-set settings. To identify unknown target samples, we follow the uncertainty-based rejection strategy from SHOT (Liang et al., 2020). We compute the entropy of the model's prediction distribution over the known source classes and utilize KMeans clustering to separate confident, low-entropy samples from uncertain, high-entropy ones. The high-entropy cluster is then rejected as belonging to unknown categories. Thus, the detection of unknown classes is driven entirely by prediction uncertainty rather than reliance on target-private class names.

E Evaluation Protocol

E.1 Protocol for Generalized SFDA

Unlike Closed-set SFDA, the Generalized SFDA problem highlights the anti-forgetting ability on the seen source domain. In terms of evaluation rule, the same as (Yang et al., 2021b), we adopt the harmonic mean accuracy that is computed by $H = (2 * A_s * A_t)/(A_s + A_t)$ where A_s and A_t are the accuracy on the source domain and the target domain, respectively. Note that the A_s is computed based on the source-testing set. The same as (Yang et al., 2021b; Tang et al., 2024b), on the source domain, the ratio of training and testing sets is 9:1.

E.2 Protocol for unification evaluation

In the SFDA context, unification does not necessarily imply that the proposed method must achieve the absolute best performance in every individual adaptation scenario. Instead, the essence of contributing a unified solution lies in its ability to provide stable and competitive performance across a wide range of settings in a unified manner, without the need of resorting to a diversity of scenario-specific designs.

To evaluate the unification of methods, we adopt three harmonic metrics: (1) Overall mean H_{all} that averages over all settings, (2) worst-case relative gap H_{wrg} that is largest percentage shortfall of a method from the per-setting best across all settings, and (3) leave-one-setting-out averages H_{loso} omits one setting at a time and averages the scores over the rest. For a specific method m evaluated over a number of settings $\mathcal{S}$, the metrics can be computed by:

$$\mathrm{H}_{all}(m) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} x_{m,s};\ \mathrm{H}_{wrg}(m) = \max_{s\in\mathcal{S}} \frac{b_s - x_{m,s}}{b_s};\ \mathrm{H}_{loso}(m, s') = \frac{1}{|\mathcal{S}-1|}\sum_{s\in\mathcal{S}\setminus s'} x_{m,s}, \tag{19}$$

where $x_{m,s}$ specifies the score of method m under a setting s; b_s=$\max_{m'} x_{m',s}$ is the per-setting best; s' is a omitted setting.

Table 9: Closed-set results (%) on **Office-Home**, **VisDA**. SF and ViL means source-free and ViL model-required, respectively.

Method	Venue	SF	ViL	Office-Home 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 S→R
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
CLIP-RN	ICML21	–	✓	51.9	81.5	82.5	72.5	81.5	82.5	72.5	51.9	82.5	72.5	51.9	81.5	72.1	83.7
CLIP-B32	ICML21	–	✓	64.9	86.6	87.2	78.0	86.6	87.2	78.0	64.9	87.2	78.0	64.9	86.6	79.2	85.2
SHOT	ICML20	✓	✗	55.0	78.7	81.3	69.1	78.9	79.1	68.2	53.6	81.6	73.5	59.4	83.5	71.8	82.9
NRC	NeurIPS21	✓	✗	57.2	79.3	81.3	68.9	80.6	80.2	66.6	57.3	82.0	71.0	57.9	84.9	72.3	85.9
AdaCon	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	ICML22	✓	✗	57.3	79.3	81.0	69.3	77.9	79.6	68.1	56.4	82.6	72.9	61.3	83.7	72.4	86.9
PLUE	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
TPDS	IJCV24	✓	✗	59.3	80.3	82.1	70.6	79.4	80.9	69.8	56.8	82.1	74.5	61.2	85.3	73.5	87.6
SHOT-B32	ICML20	✓	✓	64.7	84.0	86.4	79.4	87.3	85.4	78.1	65.8	87.6	81.6	65.0	88.5	79.5	85.2
NRC-B32	NeurIPS21	✓	✓	55.0	83.4	86.7	78.3	86.5	85.1	77.1	69.9	87.2	**90.9**	65.5	88.4	79.5	88.2
AaD-B32	NeurIPS22	✓	✓	69.4	83.0	84.9	76.4	85.2	84.6	73.3	71.1	86.3	79.0	72.1	88.7	79.5	82.6
TPDS-B32	IJCV24	✓	✓	67.7	82.7	86.1	76.5	88.0	86.3	73.6	69.4	87.1	79.3	70.5	89.7	79.7	78.0
MMGA-B32	PR26	✓	✓	68.5	84.6	86.8	78.8	87.5	86.5	75.9	68.3	86.3	80.4	69.9	88.6	80.2	89.2
CausalDA-C-RN	–	✓	✓	60.1	85.6	86.2	77.2	86.0	86.3	76.6	61.0	86.5	77.5	61.4	86.2	77.6	89.3
CausalDA-C-B32	–	✓	✓	**74.2**	**90.8**	**90.3**	**82.3**	**91.0**	**89.9**	**81.0**	**74.8**	**90.0**	82.5	**74.1**	**90.9**	**84.3**	**90.3**

Table 10: Closed-set results (%) on **DomainNet-126**. SF and ViL means: source-free and ViL model-required, respectively.

Method	Venue	SF	ViL	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
CLIP-RN	ICML21	–	✓	70.2	87.1	65.4	67.9	87.1	65.4	67.9	70.2	65.4	67.9	70.2	87.1	72.7
CLIP-B32	ICML21	–	✓	76.2	89.0	73.6	77.9	89.0	73.6	77.9	76.2	73.6	77.9	76.2	89.0	79.2
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	NeurIPS21	✓	✗	62.6	77.1	58.3	62.9	81.3	60.7	64.7	69.4	58.7	69.4	65.8	78.7	67.5
AdaCon	CVPR22	✓	✗	60.8	74.8	55.9	62.2	78.3	58.2	63.1	68.1	55.6	67.1	66.0	75.4	65.4
CoWA	ICML22	✓	✗	64.6	80.6	60.6	66.2	79.8	60.8	69.0	67.2	60.0	69.0	65.8	79.9	68.6
PLUE	CVPR23	✓	✗	59.8	74.0	56.0	61.6	78.5	57.9	61.6	65.9	53.8	67.5	64.3	76.0	64.7
TPDS	IJCV24	✓	✗	62.9	77.1	59.8	65.6	79.0	61.5	66.4	67.0	58.2	68.6	64.3	75.3	67.1
SHOT-B32	ICML20	✓	✓	72.8	84.5	72.0	75.1	85.2	72.2	78.2	75.5	71.6	75.2	73.6	83.2	76.6
NRC-B32	NeurIPS21	✓	✓	72.2	83.4	72.5	75.1	84.8	72.6	78.3	76.7	74.0	75.0	73.7	82.6	76.7
AaD-B32	NeurIPS22	✓	✓	72.2	84.2	72.0	75.3	85.3	72.6	78.2	77.4	73.6	74.8	74.7	82.8	76.9
TPDS-B32	IJCV24	✓	✓	72.5	82.8	71.8	77.7	84.7	72.9	**81.9**	76.5	72.8	73.5	74.2	83.0	77.0
MMGA-B32	PR26	✓	✓	75.4	86.5	69.9	74.6	88.3	72.3	75.9	76.1	70.9	76.2	75.6	87.4	77.4
CausalDA-C-RN	–	✓	✓	75.4	88.2	72.0	75.8	88.3	72.1	76.1	75.6	71.2	77.6	75.9	88.2	78.0
CausalDA-C-B32	–	✓	✓	**79.5**	**89.8**	**76.9**	**81.6**	**90.1**	**77.5**	81.7	**80.4**	**76.6**	**82.5**	**80.1**	**89.9**	**82.2**

F Supplementary Experimental Results

F.1 Full results on Closed-set setting

Complementing the average results presented in Tab. 3, we report the full per-task results in Tab. 9∼10. As shown, our CausalDA-C-RN and CausalDA-C-B32 consistently outperform all competing methods on almost all tasks across the three datasets, except for two cases. Compared with previous best SFDA methods TPDS (on Office-Home), PLUE (on VisDA) and GKD (on DomainNet-126), CausalDA-C-RN improves by **4.1**%, **1.0**% and **9.3**% in average accuracy, respectively. Compared with these UDA methods with both access to labeled source data and CLIP, CausalDA-C-RN still improves by **1.0**%, **0.8**%, and **2.8**% on top of second-best methods PADCLIP-RN (on Office-Home, VisDA) and ADCLIP-RN (on DomainNet-126). The improvement expands further when we switch focus to the strong version CausalDA-C-B32. Such results are unsurprising since the well-pretrained CLIP is employed to identify the external invariant elements.

On the other hand, CausalDA-C-RN defeats CLIP-RN on all tasks, achieving the improvement of **4.5**%, **5.6**%, and **5.3**% on Office-Home, VisDA, and DomainNet-126 on average, respectively. Similarly, compared with CLIP-B32, the corresponding improvement of CausalDA-C-B32 are changed to **5.1**%, **5.1**% and **3.0**%. This indicates the effect of the prompt learning to discover the external elements.

Among the "-B32" models, we have two observations. First, CausalDA-C-B32 outperforms all conventional methods across all tasks on the three datasets, confirming the effectiveness of our design. Second, on the VisDA dataset, improvements with the B32 versions are limited, and they are even outperformed by the ResNet version, such as AaD and TPDS. This is because the rendered synthetic

Table 11: Continual adaptation results (%) on **Office-Home** under Generalized SFDA setting, the first column of each sub-table indicates the adaptation sequence. ↓ means the average accuracy drop of a test domain on the adaptation path compared with the performance when the domain is first seen.

CausalDA-C-RN

	Test			
	Ar	Cl	Pr	Rw
Ar	97.8	46.1	68.4	71.1
Cl	80.0	63.5	64.3	66.0
Pr	78.4	61.7	85.0	76.2
Rw	83.1	60.2	83.2	84.4
↓	17.3	2.5	1.8	–

	Test			
	Cl	Ar	Pr	Rw
Cl	97.5	52.5	65.8	61.5
Ar	80.1	77.8	70.7	75.6
Pr	75.4	74.4	86.9	78.1
Rw	67.6	78.4	85.4	85.6
↓	23.1	1.4	1.6	–

	Test			
	Pr	Ar	Cl	Rw
Pr	99.6	52.2	40.8	71.3
Ar	91.2	78.4	50.2	78.3
Cl	81.5	71.9	61.5	69.9
Rw	87.3	76.9	59.2	85.6
↓	13.0	4.1	2.3	–

	Test			
	Rw	Ar	Cl	Pr
Rw	98.1	63.1	46.3	78.5
Ar	93.4	78.4	50.4	75.6
Cl	83.0	74.7	62.1	69.7
Pr	87.3	72.5	61.5	85.2
↓	10.2	4.8	0.6	–

CausalDA-C-B32

	Test			
	Ar	Cl	Pr	Rw
Ar	97.8	46.1	68.4	71.1
Cl	84.1	77.0	69.3	72.5
Pr	82.5	70.9	91.0	81.1
Rw	86.3	67.6	88.3	87.7
↓	13.5	7.7	2.7	–

	Test			
	Cl	Ar	Pr	Rw
Cl	97.5	52.5	65.8	61.5
Ar	80.9	83.4	72.3	77.0
Pr	74.2	78.1	90.4	82.2
Rw	69.5	81.3	88.9	89.1
↓	22.6	3.7	1.5	–

	Test			
	Pr	Ar	Cl	Rw
Pr	99.6	52.2	40.8	71.3
Ar	90.0	82.5	49.4	80.1
Cl	83.6	75.6	76.2	76.2
Rw	89.8	79.1	69.5	87.7
↓	11.8	5.2	6.7	–

	Test			
	Rw	Ar	Cl	Pr
Rw	98.1	63.1	46.3	78.5
Ar	94.0	82.8	51.2	76.8
Cl	85.2	73.8	76.2	72.5
Pr	89.5	76.9	71.5	90.0
↓	8.5	7.5	4.7	–

data (the source domain in VisDA) lacks visual details, appearing primarily in shades of white and gray light and shadow. When processed in ViT, this results in confused semantics. As for AaD and TPDS, contrastive computation (AaD) and chain-like search (TPDS) amplify the noise in the confused semantics, leading to uncontrolled error propagation.

F.2 Continual adapting results

As a supplement to the one-step adaptation results presented in Tab. 4, we further evaluate continual adaptation performance following the protocol of (Yang et al., 2021b). In the adaptation sequence Ar→Cl→Pr→Rw, we first train an initial model on domain Ar, which then serves as the source for adaptation to domain Cl. This process repeats iteratively until the final adaptation to domain Rw is achieved, resulting in four distinct models. Each model is evaluated across all domains, with **10**% of the data in each domain randomly selected as the test set.

Tab. 11 presents the results of CausalDA-C-RN and CausalDA-C-B32. We highlight two observations with CausalDA-C-B32: **First**, if a domain has been encountered during continual adaptation, the adapted model does not degrade significantly on that domain. For example, in the adaptation sequence Cl→Ar→Pr→Rw, when adapting from Cl to Ar, the accuracy on Ar reaches **83.4**%. Subsequent adaptations to Pr and Rw lead to an average performance drop of **3.7**% (see the second column in the second sub-table). **Second**, the adapted model performs well even on unseen domains. For instance, along the adaptation path Ar→Cl→Pr→Rw, the model's accuracy on Rw gradually improves from **71.1**% to **87.7**% (see the fifth column in the first sub-table). These results suggest that CausalDA effectively captures invariant factors, making it robust to domain shifts.

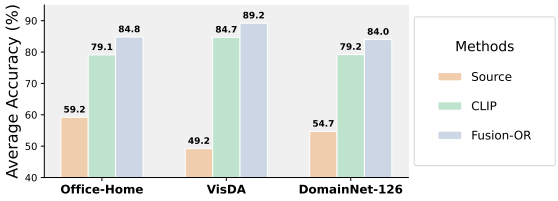


Figure 6: Zero-shot fusion results (Fusion-OR) of CLIP and the source model (Source) under the Closed-set SFDA setting on **Office-Home**, **VisDA**, and **DomainNet-126**. This fusion method follows the logical OR principle: An input instance is considered correctly classified if either CLIP or the source model predicts its ground truth category.

G Model analysis

All experiments use the strong variant CausalDA-C-B32; the suffix "-C-B32" is omitted hereafter.

G.1 Complementarity of internal and external invariant factors

As aforementioned, invariant factors originate from both the domain data itself (encoded internal component) and external knowledge sources (encoded external component). They are proxies of the internal and external causal factors, respectively.

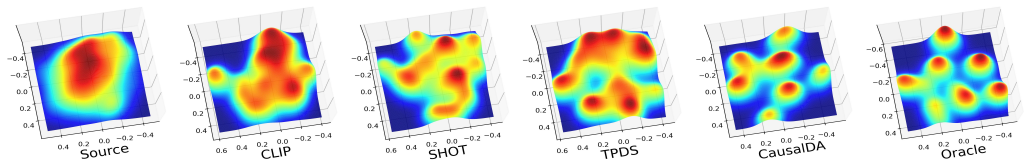


Figure 8: Logit distribution visualization on transfer task Ar→Cl in **Office-Home**. Oracle: Model trained on the real labels of domain Cl. For a clear view, the first ten categories are displayed.

To illustrate this complementarity, we conduct an intuitive empirical experiment on the target domain. Specifically, we fuse the zero-shot predictions of the source model (Source) and a ViL model like CLIP using a logical OR operation. That is, if either CLIP or Source correctly predicts the true category of an input instance $\boldsymbol{x}$, we consider $\boldsymbol{x}$ correctly classified. As shown in Fig. 6, Fusion-OR significantly outperforms both Source and CLIP individually, to some extent, reinforcing the plausibility of our internal-external partition for causality formulation in SFDA.

G.2 Pseudo-label-based analysis for discovery process of invariant factors

In CausalDA, the external invariant elements D_e are identified and converted into pseudo-labels. Thus, the quality of these pseudo-labels serves as an indicator of the discovery process. To evaluate this, we compare pseudo-label-based methods SHOT, COWA, and GKD, along with CLIP's zero-shot predictions (denoted as CLIP).

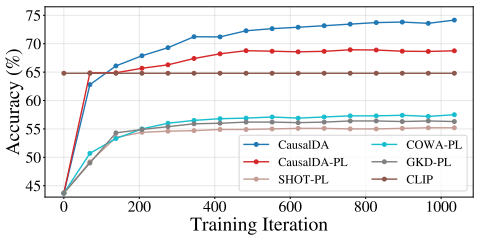


Figure 7: Dynamics of pseudo-label accuracy on task Ar→Cl in **Office-home**. The method "X-PL" means the corresponding accuracy of pseudo-label generated in X.

The results, shown in Fig. 7, demonstrate that CausalDA-PL achieves significantly higher pseudo-label accuracy than prior methods, benefiting from CLIP-assisted S_e discovery. Furthermore, the performance curve of CausalDA remains consistently above that of CausalDA-PL, with a significant gap throughout training. This highlights the effectiveness of capturing internal invariant elements D_i, further enhancing model performance.

G.3 Feature distribution visualization

Feature distribution provides an intuitive way to evaluate classification performance. We compare with Source (source model), CLIP, SHOT, TPDS, and Oracle (trained on domain Cl with real labels). Fig. 8 presents the logit visualization results using a 3D density chart. As we move from Source to CausalDA, the clustering effect progressively improves, with CausalDA exhibiting the most structured distribution, closely resembling Oracle.

G.4 Ablation study

In CausalDA, we separately discover the invariant factors $\{D_e, D_i\}$. To isolate the effect of this strategy, we evaluate four variations: (i) CausalDA-w-D_i: We use only $\mathcal{L}_{\text{UN}}$ from Eq. (12) as the objective for model training, allowing us to assess the impact of capturing D_i. (ii) CausalDA-w-D_e: We apply pseudo-label supervised learning, regulated by $\mathcal{L}_{\text{SCE}}$, where the pseudo-labels are generated from D_e discovery, regulated by L_{EC}. This isolates the effect of capturing D_e. (iii) CausalDA-w-CLIP: We directly use CLIP's zero-shot results as pseudo-labels to regulate the discovery of D_i. (iv) CausalDA-w-P1: We retain only the phase-1 objective, removing the loss $\mathcal{L}_{\text{IC}}$ from Eq. (12).

Tab. 12 presents the results of these methods across four SFDA settings and SF-OODG. Specifically, both CausalDA-w-D_i and CausalDA-w-D_e outperform the Source model, but fall significantly behind the full CausalDA version. This suggests that capturing either D_e or D_i independently is effective, but both are necessary for optimal performance. The comparison between CausalDA-w-

Table 12: Ablation study results (%). CS, OS, PS, and GS mean the Closed-set, Open-set, Partial-set, and Generalized SFDA settings, respectively.

Method	Office-Home					IN-K
	CS	OS	PS	GS	Avg.	SF-OODG
CausalDA-w-D_i	71.3	72.1	71.6	76.6	72.9	36.9
CausalDA-w-D_e	82.5	81.4	85.5	84.9	83.6	44.2
CausalDA-w-CLIP	82.0	78.8	85.1	83.3	82.3	44.2
CausalDA-w-P1	80.3	38.1	80.7	78.1	69.3	40.8
CausalDA	**84.3**	**84.0**	**87.1**	**85.2**	**85.1**	**48.4**

Table 13: Scalability study of IC instantiation.

Loss	Method	Closed-set		SF-OODG
		Office-Home	**VisDA**	**IN-K**
$\mathcal{L}_{\text{UN}}$	**CausalDA**+ENT	83.5	90.4	46.8
	CausalDA+NRC	83.9	89.0	47.5
	CausalDA+TPDS	**84.6**	**90.9**	45.0
$\mathcal{L}_{\text{SCE}}$	**CausalDA**+$\|\|\cdot\|\|_1$	83.4	89.3	44.9
	CausalDA+$\|\|\cdot\|\|_2$	83.8	89.2	40.3
	CausalDA+KL	84.5	90.0	**49.1**
	CausalDA	84.3	90.3	48.4

CLIP and CausalDA highlights the importance of L_{EC}-based prompt learning for D_e discovery. Finally, CausalDA-w-P1 performs considerably worse than CausalDA, emphasizing the value of our two-stage design in the invariant factor capture pipeline.

G.5 Scalability Study of IC Instantiation

The objective $\mathcal{L}_{\text{IC}}$ in Eq. (12), which regulates D_i discovery, is a component with multiple implementation options. To evaluate its scalability, we present three alternative instantiations for the two regulators in $\mathcal{L}_{\text{IC}}$. For $\mathcal{L}_{\text{UN}}$, we consider the following alternatives: (i) CausalDA+ENT: Entropy regularization without category balance constraints, (ii) CausalDA+NRC: Data local structure-based regularization, (iii) CausalDA+TPDS: Domain shift control-based regularization. For $\mathcal{L}_{\text{SCE}}$, we test three popular methods: 1-Norm ($||\cdot||_1$), 2-Norm ($||\cdot||_2$), and KL Divergence.

As shown in Tab. 13, in the Closed-set setting, CausalDA+TPDS achieves the best results, with other methods lagging by up to **1.9**% (on VisDA). This indicates that these typical regulators do not significantly affect model performance in this setting. In the SF-OODG setting, all methods (except CausalDA+KL) show a slight accuracy drop compared to CausalDA. For CausalDA+ENT, removing the category balance regularization leads to a performance decrease. The performance loss for CausalDA+NRC and CausalDA+TPDS is attributed to the fact that these methods do not specifically address the SF-OODG setting. For $\mathcal{L}_{\text{SCE}}$, the results show that probability distribution approximation (KL Divergence) outperforms feature-based methods ($||\cdot||_1$ and $||\cdot||_2$) in preserving model performance in the presence of external invariant factors.

Table 14: Prompt initialization study (%).

#	Initialization template	Closed-set		SF-OODG
		Office-Home	**VisDA**	**IN-K**
1	'X [CLASS].' (#X=4)	83.6	90.1	48.4
2	'X [CLASS].' (#X=16)	82.5	89.6	46.1
3	'There is a [CLASS].'	84.6	90.2	48.6
4	'This is a photo of a [CLASS].'	84.7	89.9	**48.8**
5	'This is maybe a photo of a [CLASS].'	83.9	90.2	**48.8**
6	'This is almost a photo of a [CLASS].'	**84.8**	**90.3**	48.4
7	'This is definitely a photo of a [CLASS].'	84.5	90.0	48.1
8	'a picture of a [CLASS].'	84.6	89.8	48.6
9	'a photo of a [CLASS].'	84.3	**90.3**	48.4

G.6 Effect of Prompt Initialization

In CausalDA, we employ prompt learning to encode the external invariant factors D_e. Here, we examine four different prompt initialization strategies (Tab. 14): (i) the conventional constant strategy (rows 1–2), (ii) the sentence strategy (rows 3–4), (iii) the uncertainty strategy (rows 5–7), and (iv) an innovative approach of sentence initialization with different phrase (rows 8–9), which considers the lack of precise supervision in SFDA.

Tab. 14 shows the comparison results for these nine templates across both the Closed-set setting and SF-OODG. The results indicate that the different templates do not cause significant performance differences, suggesting that CausalDA is relatively insensitive to prompt selection. However, when comparing the constant strategy to the others, it becomes clear that initializing with semantic information is the more effective option, which aligns with our expectations.

Table 15: Training resource demands comparison on task Ar→Cl in the **Office-Home** dataset. The cost variance is measured against SHOT.

Item / Method	SHOT (Liang et al., 2020)	DIFO (Tang et al., 2024c)	CausalDA (Ours)
GPU memory cost (GB)	8.7	10.1 (↑1.4)	10.8 (↑2.1)
Training times (s)	424	879 (↑455)	755 (↑331)

G.7 Training resource demands

In this part, we compare training resources on the transfer task Ar→Cl in Office-Home. For comparison, we select previous SOTA methods SHOT and DIFO. Unlike SHOT, which does not require knowledge of the ViL model, DIFO leverages this external knowledge, aligning with our method. The results presented in Tab. 15 indicate that our CausalDA, alongside DIFO, requires more computational resources and training time compared to SHOT. This is expected, as SHOT does not incur the additional overhead required to transfer task-specific knowledge from a pretrained ViL model. In the ViL model-based group, our CausalDA does not impose significant additional training costs, requiring a computational resource level similar to DIFO. This demonstrates that our performance gains stem fundamentally from the methodological design, rather than simply trading computational resources for accuracy.

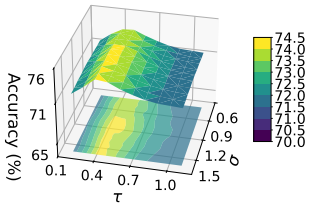


Figure 9: Parameter sensitiveness study (σ, τ) on transfer task Ar→Cl in **Office-Home**.

G.8 Parameters sensitiveness

Taking the transfer task Ar→Cl in Office-Home for example, we evaluate the sensitivity of CausalDA by varying the hyper-parameters $0.2 \leq \sigma \leq 1.1$, $0.7 \leq \tau \leq 1.6$ at the step of 0.1. Fig. 9 indicates that both corresponding terms matter, whilst their sensitivity to performance is not high.

H Limitation

In this paper, the proposed causality modeling for SFDA is characterized by two key properties: (1) it focuses on capturing causal factors, rather than conventional non-causality removal, and (2) these causal factors are further decomposed into internal and external components. Within this framework, we learn invariant factors as proxies for causal factors to approximate their capture.

Although experimental results validate the plausibility of our SFDA causality formulation, in practical optimization, we adopt mutual information maximization for invariant factor discovery, with a ViL model specifically employed to facilitate external factor refinement. These choices are primarily motivated by practical considerations rather than being directly grounded in a formal causal theory, such as intervention-based causal reasoning that removes non-causal factors. Therefore, developing a rigorous theory to address these gaps will be a key focus of our future research.