Counterfactual Knowledge Maintenance for Unsupervised Domain Adaptation

Yao Li[1,2], **Yong Zhou**[1,2*], **Jiaqi Zhao**[1,2], **Wen-liang Du**[1,2], **Rui Yao**[1,2], **Bing Liu**[1,2]
[1]School of Computer Sciences and Technology, China University of Mining and Technology
[2]Mine Digitization Engineering Research Center of the Ministry of Education
{liyao0719, yzhou, jiaqizhao, wldu, ruiyao, liubing}@cumt.edu.cn,

Abstract

Traditional unsupervised domain adaptation (UDA) struggles to extract rich semantics due to backbone limitations. Recent large-scale pre-trained visual-language models (VLMs) have shown strong zero-shot learning capabilities in UDA tasks. However, directly using VLMs results in a mixture of semantic and domain-specific information, complicating knowledge transfer. Complex scenes with subtle semantic differences are prone to misclassification, which in turn can result in the loss of features that are crucial for distinguishing between classes. To address these challenges, we propose a novel counterfactual knowledge maintenance UDA framework. Specifically, we employ counterfactual disentanglement to separate the representation of semantic information from domain features, thereby reducing domain bias. Furthermore, to clarify ambiguous visual information specific to classes, we maintain the discriminative knowledge of both visual and textual information. This approach synergistically leverages multimodal information to preserve modality-specific distinguishable features. We conducted extensive experimental evaluations on several public datasets to demonstrate the effectiveness of our method. The source code is available at https://github.com/LiYaolab/CMKUDA.

1 Introduction

Deep learning has achieved significant success with large datasets [Huang *et al.*, 2022], but models often require vast amounts of labeled data and are highly dependent on data-driven features. When applied to different domains with similar tasks, performance can degrade due to domain discrepancies. Data-driven models, assuming independent and identically distributed (i.i.d.) data, are vulnerable to distribution shifts from training data [Zhao *et al.*, 2025]. Unsupervised domain adaptation (UDA) offers a solution by allowing models to be trained on labeled data from a source domain and tested on unlabeled data from a target domain, reducing the impact of domain differences. Traditional UDA methods aim to minimize the distribution gap using techniques like entropy minimization [Wang and Deng, 2018], moment matching [Li *et al.*, 2021], and adversarial learning [Du *et al.*, 2021]. However, aligning visual features without addressing distribution differences can cause semantic distortions and reduce class distinguishability [Tang *et al.*, 2020a]. Additionally, converting text labels into numerical labels limits the model's ability to capture rich, nuanced class information, leading to suboptimal performance [Ge *et al.*, 2023].

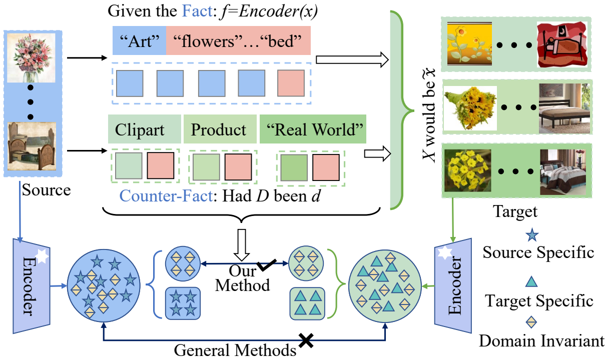


Figure 1: Current prompt-based UDA methods lead to a coarse-grained alignment that encompasses both domain-related and semantically-related features. Our method decouples semantic information and domain features by counterfactual disentanglement, only aligning semantic related features. This approach minimizes the bias introduced by domain-specific knowledge.

To overcome the limitations of traditional UDA methods, large-scale pre-trained visual language models (VLMs) have been increasingly leveraged. VLMs, such as CLIP [Radford *et al.*, 2021] and ALIGN [Jia *et al.*, 2021], are pre-trained on extensive image and text pairs, aligning visual and textual features within a joint embedding space. This pre-training process encodes rich visual and textual information, enhancing the adaptability and generalization of the model across different domain contexts through alignment. Therefore, VLMs possess a robust conceptual knowledge foundation and strong zero-shot transfer capabilities, making them highly effective for addressing domain adaptation challenges [Bai *et al.*, 2024; Ge *et al.*, 2023; Singha *et al.*, 2023].

However, these methods primarily leverage the pre-

[*]Corresponding author

training knowledge of VLMs to enhance generalization and build statistical dependencies between data and labels, without elucidating the underlying relationship. UDA requires learning not only statistical correlations between visual features and labels across domains but also potential causal mechanisms [Wang *et al.*, 2023]. By establishing causal relationships instead of correlations between visual features and labels, models can better classify the same visual feature in various scenarios. Existing UDA methods often lead to spurious correlations between inference results and visual features. Additionally, VLMs involve two distinct modalities: visual and textual. For classification, visual features may be more influential for some samples, while textual features may be more useful for others. For example, categories with distinct visual characteristics, like chairs and sofas, are easier to classify using visual features. In contrast, categories with similar visual features, such as filing cabinets and shelves, are better classified using textual features. Current UDA methods mainly rely on the extensive visual knowledge in CLIP for classification, leading to suboptimal results when visual semantic information is ambiguous.

To address these challenges, we propose a novel UDA method using counterfactual disentanglement and discrimination knowledge maintenance. First, we introduce a structural causal model that identifies semantic information relevant to data types and categories as causal factors, while domain-specific information unrelated to categories is treated as non-causal factors. Only semantic features causally affect category labels. As shown in Figure 1, given a source domain sample x, the coupled feature f can be obtained by the encoder. By intervening in the domain information $D = d$ and setting it as a counterfactual with target domain knowledge, the sample $\tilde{x}$ in the target domain can be obtained. During this process, semantic information remains unchanged across domains. By employing counterfactual disentanglement, we decouple semantic and domain features, thereby mitigating biases inherent in CLIP's knowledge base. Secondly, we adaptively integrate the visual and textual embeddings extracted by VLMs, assessing which modality is more conducive to inference for each sample. By maintaining discrimination knowledge, we preserve the distinctive features of each modality and clarify class-specific ambiguous visual information. Our method not only eliminates spurious correlations caused by confounding factor domain confidence, but also synergistically combines the advantages of visual and textual modalities to enhance task performance.

Overall, our significant contributions can be summarized as follows:

- We propose a novel counterfactual knowledge maintenance framework and construct a causal model from a causal perspective for the UDA problem.
- To solve the confusion between semantic and domain features, counterfactual disentanglement is proposed to decouple and represent mixed knowledge, obtaining domain-invariant knowledge.
- To distinguish complex scenes with slight semantic details, discrimination knowledge maintenance is used to clarify class-specific ambiguous visual information based on modal preferences of discriminative features.
- Extensive experiments are conducted on public datasets to demonstrate the effectiveness of the proposed method.

2 Related Work

2.1 Causal Mechanism

Deep learning models traditionally rely on vast datasets to generate predictions, which can lead to capturing spurious correlations [Yue *et al.*, 2023]. This results in models that lack interpretability, robustness, and generalization capabilities. Causal learning is used to establish robust causal relationships within complex and unstructured data. To mitigate the impact of confounding factors in data, Zhu et al. [Zhu *et al.*, 2023] introduced a method that leverages causal relationships to reweight and resample online data, thereby enhancing model performance. Ding et al. [Ding *et al.*, 2023] employed causal models to solve the problem of inconsistent data distribution in the training dataset. Furthermore, Yue et al. [Yue *et al.*, 2021] proposed a transport causal mechanism to identify the representation of confusion layers and domain invariant solution causal mechanisms. Yang et al. [Yang *et al.*, 2023b] introduced a learned causal representation.

2.2 Prompt Learning

Visual language models integrate rich visual and textual information from large datasets and excel in various computer vision tasks. Recently, prompts have been integrated into VLMs to learn universal visual representations. CLIP [Radford *et al.*, 2021] is the most groundbreaking. Zhou et al. [Zhou *et al.*, 2022b; Zhou *et al.*, 2022a] proposed the CoOp and CoCoOp method, which model prompts using continuous representations, automatically learning task-related prompts. However, these methods overlooked domain transfer issues. DAPL [Ge *et al.*, 2023] processes distribution transitions in UDA by learning domain-independent and domain-specific cues. PADCLIP [Lai *et al.*, 2023] introduces domain names and dynamically adjusts depolarization strength and momentum in prompts to bridge domain gaps. PDA [Bai *et al.*, 2024] proposed a prompt-based distribution alignment method that incorporates domain knowledge into prompt learning. AD-CLIP [Singha *et al.*, 2023] addresses the domain adaptation problem in the prompt space with a domain-independent CLIP prompt learning strategy. DAMP [Du *et al.*, 2024] proposes domain-agnostic mutual prompts, which align visual and textual embeddings with domain-invariant semantics.

However, previous methods directly use coupled domain and semantic knowledge from CLIP, without considering the varying effectiveness of different modalities for specific tasks. To address this, we employ causal decoupling to separate semantic knowledge from domain-specific knowledge while maintaining discrimination knowledge for ambiguous visual semantic information.

3 Proposed Method

3.1 Overview of Our Framework

Formally, the problem definition and symbolic representation of UDA can be described as follows: given labeled

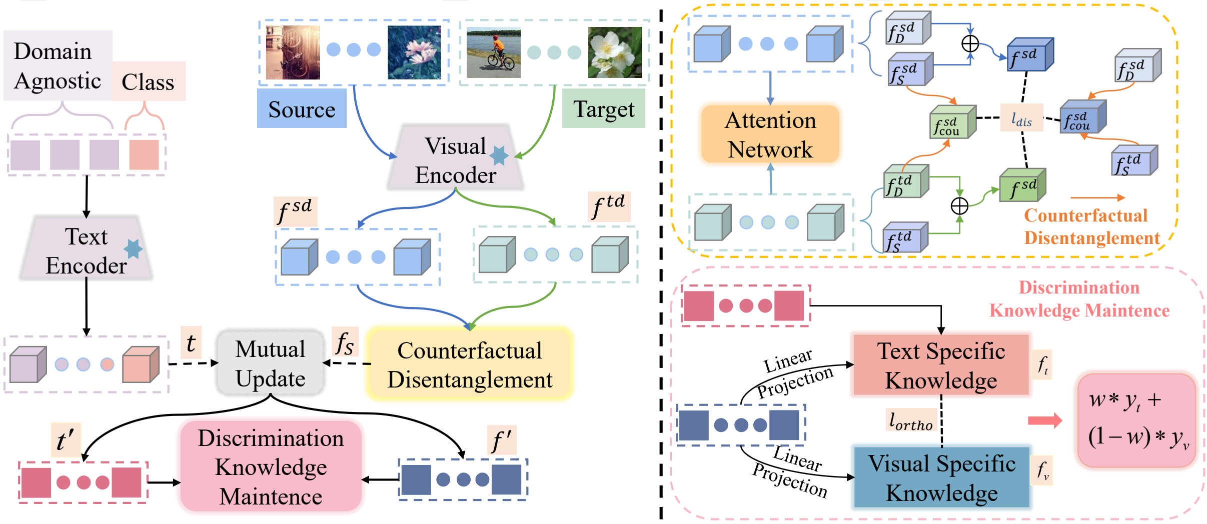


Figure 2: The overall framework of our method. Initially, a frozen text and visual encoder are employed to extract text features t and visual features f, respectively. Given confounding visual information, we apply counterfactual disentanglement to isolate semantic information f_S that exclusively contains domain-invariant features. Subsequently, the extracted text features and semantic features are fed into existing mutual updating mechanisms to generate fused multimodal outputs t' and f'. This process ultimately enhances the classification accuracy of ambiguous visual information samples by maintaining discriminative knowledge.

source domain image data $\mathcal{D}_{sd} = \{x_i^{sd}, y_i^{sd}\}_{i=1}^{N_{sd}}$ and unlabeled target domain image data $\mathcal{D}_{td} = \{x_i^{td}\}_{i=1}^{N_{td}}$. N_{sd} and N_{td} represent the number of samples from the source and target domains, respectively. The objective of UDA is to train a model on the source domain data with distribution $x_{sd}^i \sim P_{sd}(X)$ such that it performs well on the target domain data with distribution $x_{td}^i \sim P_{td}(X)$. The primary challenge in this process is that the samples originate from two distinct distributions$P_{sd}(X) \neq P_{td}(X)$, which is often referred to as the domain shift problem.

Therefore, based on the CLIP framework, the counterfactual disentanglement is first introduced to decouple domain-related and semantic-related features, thereby mitigating the confounding effects of domain-specific information. Secondly, a discrimination knowledge maintenance strategy that tailors the combination of text and visual information to the needs of individual samples, enhancing task performance. The framework is illustrated in Figure 2.

For clarity, superscripts in the text denote domains: sd for the source domain and td for the target domain. Symbols without superscripts apply to both domains. Specifically, input the sample x from the source or target domain into the frozen visual encoder $\mathcal{V}$ to extract the coupled visual feature f. Construct a naive prompt $\{t_i\}_{i=1}^{K}$ for “a [DOM] photo of a [CLA]”, where [DOM] represents domain names, [CLA] represents category names, and K represents the number of categories. Input the prompt information into frozen text encoder $\mathcal{T}$ to obtain the text feature t. Then, the coupled visual features f are decoupled by counterfactual disentanglement. The domain-related features f_D and semantic features f_S are separated to eliminate the confounding caused by domain-related features. Inspired by [Du *et al.*, 2024], the semantic feature f_S and text t_i are mutually prompted to fully integrate visual and textual information, yielding updated visual feature f' and text feature t'. By orthogonally decomposing the updated visual feature f', the text-specific feature f_t' and the visual-specific feature f_v' are extracted. The final logits y_{all} are obtained by modal adaptive fusion f_t' and f_v' using weight w, where w is a learnable parameter. Finally, the obtained y_{all} is used to calculate the cross entropy loss with the source domain label y_{sd}^i and the pseudo label of the target $\hat{y}_{td}^i$ domain for training.

3.2 Counterfactual Disentanglement

To explore the causal relationship between data and labels, a Structural Causal Model (SCM) is integrated into the CLIP-based UDA framework. The visual-language model (VLM) P extracts both visual information V and textual information T. In V, causal variables are those directly related to categories and labels, like the ”shape” feature in digit recognition, and these relationships are consistent across domains. Non-causal variables, such as ”handwriting style” in digit recognition, are domain-specific. Each sample X is a combination of causal variables C and non-causal variables U, with only C having a direct causal impact on the category label Y, as shown in Figure 3.

The objective of the model is to identify and extract causal factor C from the raw data X, and establish a stable causal relationship framework. Before that, two basic theorems [Reichenbach and Morrison, 1956] are given.

Theorem 1 (Common Cause Principle)**.** *If variables X and Y exhibit statistical correlation, it implies the presence of a common causal factor C that influences both X and Y, accounting for their observed associations. To put it succinctly, the correlation between X and Y vanishes when the influence of C is considered, rendering X and Y conditionally independent given C.*

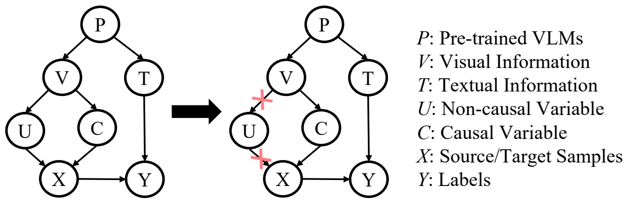


Figure 3: SCM of our method. The solid arrow indicates that the parent node causes the child node, and it is necessary to block statistical dependencies in the causal path. We take domain-invariant semantic information as causal variables C and domain-related information as non-causal variables U.

According to Theorem 1, formalize the structural causal model of Figure 3.

$$X := \mathcal{F}(C, U, V_1), C \perp U \perp V_1, \tag{1}$$

$$Y := \mathcal{H}(C, V_2) = \mathcal{H}(\mathcal{G}(X), V_2), V_1 \perp V_2, \tag{2}$$

where, V_1 and V_2 are independent noise variables that cannot be explained. The functions $\mathcal{F}$, $\mathcal{H}$, and $\mathcal{G}$ are unknown structural functions.

In addition, causal factors need to meet the requirement of mutual independence [Peters *et al.*, 2017], as expressed in theorem 2.

Theorem 2 (Independent Causal Mechanism Principle)**.** *Each variable's distribution, conditioned on its causal factors, is independent and does not influence or inform other causal mechanisms.*

The joint distribution of causal factors can be decomposed into the following conditions, namely causal decomposition:

$$P(c_1, c_2, \ldots, c_N) = \prod_{i=1}^{N} P(c_i \mid PV_i), \tag{3}$$

where, PV_i is the parents of c_i in causal graph.

In essence, the model must identify key causal factors which are independent of domain-specific factors and can significantly affect classification outcomes.

Given a sample x, we generate counterfactual samples based on the following three steps.

- **Abduction** "given the fact that $f = Encoder(x)$". Given evidence $X = x$, we obtain the source domain sample attribute $Encoder(x)$.
- **Action** "had D been d". $d \in \{sd, td\}$ is the domain label. In this step, we intervene in D by discarding the inferred value and setting D as d.
- **Prediction** "X would be $\tilde{x}$". Under the condition of inferring $f = Encoder(x)$ (fact) and intervention target $D = d$ (counterfactual), we can generate counterfactual samples $\tilde{x}$ from $P_\theta(X|f = Encoder(x), D = d)$.

Counterfactual is used to decouple the causal factor C and non-causal factor U by generating new data with perturbed domain-related information. The generated data differs in non-causal factors U but retains the same causal factor C, ensuring the representation remains unchanged. Specifically, for an image sample x, it is fed into a fixed visual encoder $\mathcal{V}$ to extract the coupled visual features f. To decouple domain-related and semantic-related features, an attention network $\mathcal{N}$ is utilized, which splits the features into two distinct components. Multiply the coupled visual features by the attention network weights to obtain semantic-related features f_S, while the rest are domain-related features f_D.

After extracting domain-related and semantic features, the feature map's dimensionality is reduced using global average pooling and projection. Separate projection layers are used for each type of feature due to their significant differences. A counterfactual approach is applied to decouple these features, separating domain-related and semantic features while preserving semantic information. This process integrates semantic features from one domain with domain features from another.

$$f_{cou}^{sd} = f_S^{sd} + f_D^{td}, \tag{4}$$

$$f_{cou}^{td} = f_S^{td} + f_D^{sd}. \tag{5}$$

To further ensure the independence of f_D and f_S, a domain classifier is employed to differentiate the domain-specific information between the original and counterfactual features. The source domain is labeled as 0, and the target domain as 1. For counterfactual features, f_{cou}^{sd} encapsulates the semantic information from both the source and target domains, as does f_{cou}^{td}. Consequently, the domain discrimination loss is:

$$\begin{aligned} \mathcal{L}_{dis} =& \ell_{bce}(p_{dom}(f^{sd}), 0) + \ell_{bce}(p_{dom}(f^{td}), 1) \\ &+ \ell_{bce}(p_{dom}(f_{cou}^{td}), 0) + \ell_{bce}(p_{dom}(f_{cou}^{sd}), 1). \end{aligned} \tag{6}$$

3.3 Discrimination Knowledge Maintenance

After decoupling to extract the transferable causal feature f_S, the mutual prompting method [Du *et al.*, 2024] is employed to effectively integrate visual and textual knowledge with the causal feature f_S and the textual prompt t, ultimately obtaining updated visual f' and textual prompt t'. However, both visual and textual cues are unique and modality-specific, and certain samples are best classified using specific modalities. Pre-trained CLIP may sometimes misclassify these items into visually similar categories. For complex items with subtle semantic differences, visual classifiers might make errors, while CLIP can leverage its extensive knowledge base for accurate zero-shot predictions. In summary, the visual branch excels at recognizing visual features of specific categories, while the text branch clarifies uncertainties in classification through semantic information. Therefore, discrimination knowledge is maintained to adaptively fuse visual and textual information for each sample, enhancing classification performance.

Specifically, the updated visual feature f' is decomposed into text-specific feature f'_t and visual-specific feature f'_v through a modal partitioning network by linear projection.

During this process, orthogonal loss is used to ensure the independence of separation.

$$\mathcal{L}_{ort} = |f_t'^{sd} \cdot f_v'^{sd}|_F^2 + |f_t'^{td} \cdot f_v'^{td\top}|_F^2. \tag{7}$$

After obtaining modality-specific features, each modality is processed to obtain the final output. For the text specific

Method	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**
RN-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [Ganin and Lempitsky, 2015]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
GSDA [Hu *et al.*, 2020]	**61.3**	76.1	79.4	65.4	73.3	74.3	65.0	53.2	80.0	72.2	60.6	83.1	70.3
GVB-GD [Cui *et al.*, 2020]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
SPL [Wang and Breckon, 2019]	54.5	77.8	81.9	65.1	78.0	81.1	66.0	53.1	82.8	69.9	55.3	86.0	71.0
ToAlign [Wei *et al.*, 2024]	57.9	76.9	80.8	66.7	75.6	77.0	67.8	57.0	82.5	75.1	60.0	84.9	72.0
SRDC [Tang *et al.*, 2020b]	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
CLIP [Radford *et al.*, 2021]	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
PADCLIP [Lai *et al.*, 2023]	57.5	84.0	83.8	77.8	85.5	84.7	76.3	59.2	85.4	78.1	$\underline{60.2}$	86.7	76.6
DAPL [Ge *et al.*, 2023]	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5
AD-CLIP [Singha *et al.*, 2023]	55.4	85.2	85.6	$\underline{76.1}$	85.8	86.2	**76.7**	56.1	85.4	76.8	56.1	85.5	75.9
DAMP [Du *et al.*, 2024]	59.7	$\underline{88.5}$	$\underline{86.8}$	**76.6**	$\underline{88.9}$	$\underline{87.0}$	$\underline{76.3}$	$\underline{59.6}$	$\underline{87.1}$	$\underline{77.0}$	**61.0**	**89.9**	$\underline{78.2}$
Ours	$\underline{60.2}$	**89.1**	**87.5**	75.8	**89.0**	**87.6**	76.2	**61.7**	**87.5**	**77.3**	**61.0**	$\underline{89.0}$	**78.5**

Table 1: Using ResNet50 as the backbone, comparison of our method with state-of-the-art methods for UDA task on Office-Home dataset. The best and second-best accuracy are indicated in bold and underlined respectively.

Method	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**
CDTrans* [Xu *et al.*, 2022]	68.8	85.0	86.9	81.5	87.1	87.3	79.6	63.3	88.2	82.0	66.0	90.6	80.5
TVT [Yang *et al.*, 2023a]	74.9	86.8	89.5	82.8	88.0	88.3	79.8	71.9	90.1	85.5	74.6	90.6	83.6
SSRT [Sun *et al.*, 2022]	75.2	89.0	91.1	85.1	88.3	90.0	85.0	74.2	91.3	85.7	$\underline{78.6}$	91.8	85.4
CLIP [Radford *et al.*, 2021]	67.8	89.0	89.8	82.9	89.0	89.8	82.9	67.8	89.8	82.9	67.8	89.0	82.4
PADCLIP [Lai *et al.*, 2023]	$\underline{76.4}$	90.6	90.8	**86.7**	92.3	92.0	86.0	74.5	91.5	86.9	**79.1**	93.1	86.7
DAPL [Ge *et al.*, 2023]	70.6	90.2	91.0	84.9	89.2	90.9	84.8	70.5	90.6	84.8	70.1	90.8	84.0
AD-CLIP[Singha *et al.*, 2023]	70.9	92.5	**92.1**	85.4	92.4	**92.5**	**86.7**	74.3	**93.0**	**86.9**	72.6	93.8	86.1
DAMP [Du *et al.*, 2024]	75.7	$\underline{94.2}$	$\underline{92.0}$	$\underline{86.3}$	$\underline{94.2}$	91.9	$\underline{86.2}$	**76.3**	92.4	86.1	75.6	$\underline{94.0}$	$\underline{87.1}$
Ours	**77.2**	**94.4**	91.8	**86.7**	**94.8**	$\underline{92.1}$	85.8	$\underline{76.1}$	$\underline{92.9}$	$\underline{86.5}$	76.0	**94.7**	**87.4**

Table 2: Using ViT-B/16 as the backbone, comparison of our method with state-of-the-art methods for UDA task on Office-Home dataset. Whereas, CDTrans* has used DeiT-base backbone only. The best and second-best accuracy are indicated in bold and underlined respectively.

modality f_t',

$$\hat{y}_t = (\hat{l}_1, \hat{l}_2, \cdots \hat{l}_k), \quad \hat{l}_i = \cos(t_i, f_t')/temp, \tag{8}$$

where, $temp$ is temperature in pretrained CLIP. Since the target domain data has no labels, we first obtain the pseudo label $\hat{y}^{td}$ of the target domain data, and then calculate the cross entropy loss. For source data, cross-entropy loss is directly applied using labeled source data.

$$\mathcal{L}_t = \mathrm{CE}(\hat{y}_t^{td}, \hat{y}^{td}) + \alpha \mathrm{CE}(\hat{y}_t^{sd}, y^{sd}), \tag{9}$$

where the value of α can be adjusted based on the influence of supervised source domain data.

For the visual specific modality f_v', pass it through two learnable linear layers ψ_1 and ψ_2 to obtain the final output.

$$\hat{y}_v = \psi_2(\psi_1(f_v')). \tag{10}$$

Utilizing Eq. (8) and Eq. (10), the ensemble output $\hat{y}_{ens}$ is formulated as:

$$\hat{y}_{all} = w * \hat{y}_t + (1 - w) * \hat{y}_v, \tag{11}$$

where, w is a learnable parameter obtained by the weight generation network $\mathcal{W}$.

Like Eq. (9), we will calculate the cross entropy loss for the adaptive fusion features yens and labels obtained. In addition, to further embed the updated target domain under the learned semantic structure, and enhance individual discriminability and global diversity, an additional information maximization loss $\mathcal{L}_{im}$ is added to regularize unlabeled target data.

$$\mathcal{L}_{im} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{vi}^c \log y_{vi}^c - \sum_{k=1}^{K} \hat{y}_v^k \log \hat{y}_v^k, \tag{12}$$

where, N is the total number of image samples, K is the total number of image categories.

Therefore, the loss of specific visual modalities is:

$$\mathcal{L}_v = CE(\hat{y}_{all}^{td}, \hat{y}^{td}) + \beta CE(\hat{y}_{all}^{sd}, y^{sd}) + \mathcal{L}_{im}, \tag{13}$$

where the value of β can be adjusted based on the influence of supervised source domain data.

3.4 Overall Training Objective

We train our method with the supervised loss and above regularizations in an end-to-end manner. As depicted in Figure 2, the pre-trained text and vision encoder are frozen. We optimize parameters of a counterfactual disentangled network and discrimination knowledge maintenance network denoted as θ_{cau} and θ_{mod} respectively. Combining Eq. (6), Eq. (9), Eq. (13), we define the following optimization problem:

$$\theta_{cau} = \arg\min_{\theta_{cau}} \mathcal{L}_{dis}, \tag{14}$$

$$\theta_{mod} = \arg\min_{\theta_{mod}} \mathcal{L}_t + \mathcal{L}_v. \tag{15}$$

The overall loss of the entire network is:

$$\mathcal{L}_{dis} = \gamma_1 * \mathcal{L}_{dis} + \gamma_2 * (\mathcal{L}_t + \mathcal{L}_v). \tag{16}$$

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We evaluated our method on two prominent public datasets: Office-Home (Venkateswara et al., 2017), and VisDA-2017 (Peng et al., 2018). Office-Home comprises images across four distinct domains, encompassing 65 categories. VisDA-2017 features 152,000 synthetic images in the source domain and 55,000 real images in the target domain.

Method	plane	bicycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg**
RN-101 [He *et al.*, 2016]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [Ganin and Lempitsky, 2015]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN [Long *et al.*, 2017]	75.7	18.7	82.3	86.3	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
MODEL [Li *et al.*, 2020]	94.8	73.4	68.8	74.8	93.1	95.4	88.6	**84.7**	89.1	84.7	83.5	48.1	81.6
STAR [Lu *et al.*, 2020]	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
CLIP [Radford *et al.*, 2021]	98.2	83.9	**90.5**	73.5	97.2	84.0	<u>95.3</u>	65.7	79.4	89.9	91.8	63.3	84.4
DAPL [Ge *et al.*, 2023]	<u>97.8</u>	83.1	88.8	77.9	97.4	91.5	94.2	79.7	88.6	89.3	<u>92.5</u>	62.0	86.9
AD-CLIP [Singha *et al.*, 2023]	98.1	83.6	91.2	76.6	98.1	93.4	**96.0**	81.4	86.4	91.5	92.1	64.2	87.7
PADCLIP [Lai *et al.*, 2023]	96.7	88.8	87.0	<u>82.8</u>	97.1	93.0	91.3	83.0	<u>95.5</u>	<u>91.8</u>	91.5	63.0	<u>88.5</u>
DAMP [Du *et al.*, 2024]	97.3	<u>91.6</u>	89.1	76.4	<u>97.5</u>	<u>94.0</u>	92.3	<u>84.5</u>	91.2	88.1	91.2	<u>67.0</u>	88.4
Ours	**97.9**	**94.3**	<u>89.5</u>	81.7	96.9	**99.7**	90.6	83.0	**97.1**	**92.9**	**94.8**	**68.2**	**90.6**

Table 3: Using ResNet101 as the backbone, comparison of our method with state-of-the-art methods for UDA task on VisDA-2017 dataset. The best and second-best accuracy are indicated in bold and underlined respectively.

Method	plane	bicycle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	**Avg**
CDTrans* [Xu *et al.*, 2022]	97.1	90.5	82.4	77.5	96.6	96.1	93.6	**88.6**	**97.9**	86.9	90.3	62.8	88.4
TVT [Yang *et al.*, 2023a]	97.1	92.9	85.3	66.4	97.1	97.1	89.3	75.5	<u>95.0</u>	94.7	94.5	55.1	86.7
SSRT [Sun *et al.*, 2022]	98.9	87.6	89.1	<u>84.8</u>	98.3	<u>98.7</u>	96.3	81.1	94.9	97.9	94.5	43.1	88.8
CLIP [Radford *et al.*, 2021]	99.1	91.7	<u>93.8</u>	76.7	98.4	91.7	95.3	82.7	86.5	<u>96.0</u>	94.6	60.5	88.9
DAPL [Ge *et al.*, 2023]	99.2	92.5	93.3	75.4	98.6	92.8	95.2	82.5	89.3	96.5	95.1	63.5	89.5
AD-CLIP [Singha *et al.*, 2023]	99.6	92.8	**94.0**	78.6	<u>98.8</u>	95.4	**96.8**	83.9	91.5	95.8	<u>95.5</u>	65.7	90.7
PADCLIP [Lai *et al.*, 2023]	98.1	<u>93.8</u>	87.1	**85.5**	98.0	96.0	94.4	<u>86.0</u>	94.9	93.3	93.5	<u>70.2</u>	<u>90.9</u>
DAMP [Du *et al.*, 2024]	<u>98.7</u>	92.8	91.7	80.1	**98.9**	96.9	94.9	83.2	93.9	94.9	94.8	<u>70.2</u>	<u>90.9</u>
Ours	**98.9**	**96.6**	<u>93.8</u>	79.2	98.7	**99.8**	<u>96.6</u>	83.2	94.4	93.8	**96.6**	**70.3**	**91.8**

Table 4: Using ViT-B/16 as the backbone, comparison of our method with state-of-the-art methods for UDA task on VisDA-2017 dataset. Whereas, CDTrans* has used DeiT-base backbone only. The best and second-best accuracy are indicated in bold and underlined respectively.

Training Configuration. We employed ResNet50 [He *et al.*, 2016], and ViT-B/16 [Dosovitskiy *et al.*, 2021] as visual encoders $\mathcal{V}$, and utilized a pre-trained CLIP text encoder as the text encoder $\mathcal{T}$. The parameters of both encoders were frozen during training. We set the length of the learnable text prompt L to 32. For optimization, we used the Adam optimizer [Kingma and Ba, 2017] with an initial learning rate of $3e-3$ and trained the model for 30 epochs with 32 batch sizes. All experiments were conducted on an NVIDIA RTX A6000 GPU.

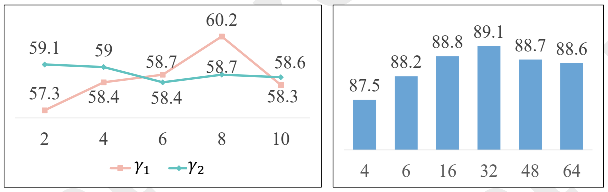
4.2 Comparasion with State-of-the-Arts

Results on Office-Home dataset. For a fair comparison, we evaluate our approach against a comprehensive set of traditional UDA methods based on CNN and transformer and CLIP-based approaches. Initially, we employ ResNet-50 as the backbone for our visual encoder, with the results detailed in Table 1. Next, we use ViT-B/16 as the visual encoder, and the corresponding results are shown in Table 2. Overall, our method consistently demonstrates outperformed performance. Especially, our method achieved the best results in 8 out of the 12 tasks evaluated.

Results on VisDA-2017 dataset. Similarly, we conduct comparisons using ResNet-101 and ViT-B/16 as visual encoders.The results are shown in Table 3 and Table 4, respectively. Our method further outperforms the state-of-the-art method in terms of average accuracy, by 2.2% and 0.9% respectively. The recognition accuracy has been improved across specific categories, such as ”bicycle”, ”knife”, ”train”, etc. Collectively, these improvements underscore the significance of extracting causal factors and maintaining discriminative knowledge.

f_v	CD	DKM	avg	f_v	avg
RN-50	×	×	74.5	ViT-B/16	84.0
	✓	×	78.4		87.1
	×	✓	78.3		87.2
	✓	✓	**78.5**		**87.4**

Table 5: Ablation study of our method with Counterfactual Disentanglement (CD) and Discrimination Knowledge Maintenance (DKM). Where × and ✓respectively represent removing or adding the module while maintaining all other configurations constant.



(a) Weight values on loss (b) Different prompt lengths

Figure 4: The results of different weight values on losses.

4.3 Ablation Study

Effectiveness of each module. We conduct ablation experiments to substantiate the efficacy of counterfactual disentanglement and discriminative knowledge maintenance. The experimental outcomes are detailed in Table 5. Utilizing DAPL

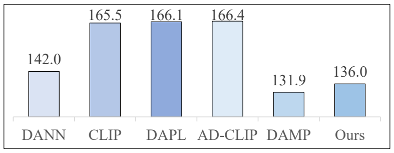
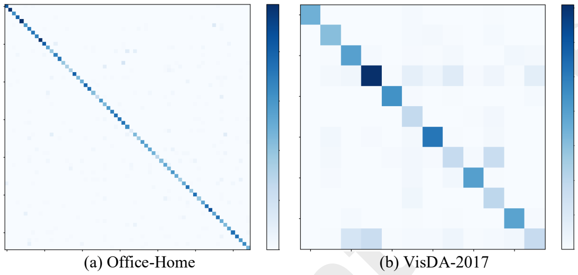


Figure 5: Comparison of the computational complexity in terms of GFLOPs.


(a) Office-Home

(b) VisDA-2017

Figure 6: The confusion matrix visualization on the Office-Home and VisDA-2017 datasets.

[Ge *et al.*, 2023] as the baseline. A major observation is that the omission of any module invariably resulted in performance decrements of varying magnitudes, emphasizing the positive contribution of each module to the overall performance. Similarly, as our method is based on CLIP, we demonstrated the results of our ablation experiment using different backbone networks.

Parameter sensitivity analysis. We also conduct sensitivity analysis on key parameters, focusing on the weights of causal loss and knowledge maintenance loss, as well as the length of learnable tokens. According to Figure 4 (a), we can see that the weights of both losses affect the performance of the model to varying degrees. When γ_1 is set to 8 and γ_2 is set to 4, the experimental results are the best. According to Figure 4 (b), we can see that as the length of the learnable token increases, the accuracy of the model also increases. However, when the length of the learnable token exceeds 32, the accuracy of the model gradually begins to decline. Therefore, we set the length of the learnable token to 32 in the experiment.

Model Complexity. As shown in Figure 5, the result of our method requires fewer 4.23%, 17.82%, 18.12%, 18.27% than DANN [Ganin and Lempitsky, 2015], CLIP [Radford *et al.*, 2021], DAPL [Ge *et al.*, 2023], and AD-CLIP [Singha *et al.*, 2023] computational resources than most others.

4.4 Analysis

Confusion matrix visualization. We generated confusion matrices using our method on the Office-Home dataset and the VisDA-2017 dataset. As depicted in (a) and (b) of Figure 6, both matrices exhibit a pronounced diagonal structure.

Feature map visualization. Figure 7 shows the visualization of the model's feature map on the Office-Home dataset. The left portion illustrates the adaptation from the "Clipart" domain to the "Product" domain, and the right portion depicts the adaptation from the "Art" domain to the "Real World" domain. The feature map reveals that our method facilitates the model in primarily learning category-related causal features. By mitigating the interference from background information, our approach enhances the model's generalization capability.

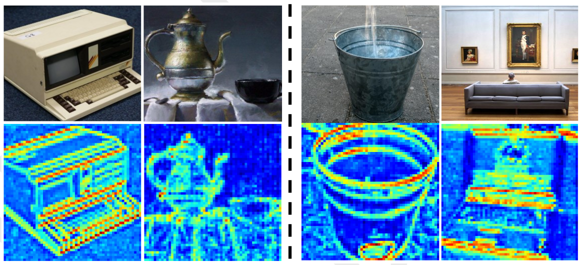
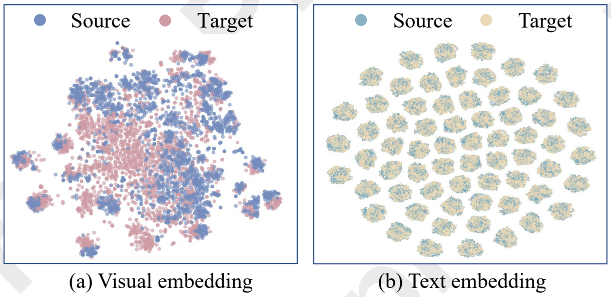


Figure 7: The feature visualization of "Real World" domain on Office-Home dataset.



(a) Visual embedding

(b) Text embedding

Figure 8: t-SNE visualizations of visual embedding and text embeddings from "Art" to "Real World" domains on Office-Home dataset.

t-SNE visualization. Figure 8 shows the t-SNE visualization of the "Art" to "Real World" task on the Office-Home dataset. The visualization reveals that our method successfully aligns visual and text embeddings, achieving both domain invariance and discriminability. The overlapping areas indicate the capability to learn domain invariant knowledge. Moreover, the text embeddings exhibit marked separation between distinct categories, which underscores the model's strong discriminative power in differentiating various categories.

5 Conclusion

In this paper, we introduce a novel UDA method, which employs counterfactual disentanglement and discriminative knowledge maintenance. By leveraging counterfactuals to disentangle domain-specific and semantics-related features, our approach mitigates the confounding effects of domain-related features. To further address the challenge of classifying samples with ambiguous semantic information, we propose modal adaptive fusion to enhance the extraction of class-discriminative features. The proposed method lead to improved feature disentanglement and class recognizability. Extensive experiments demonstrate that our method consistently outperforms two strong baselines, offering a robust method for UDA to harness source code and pre-trained VLM knowledge. Future work will explore the application of UDA in open-world scenarios, where models need to adapt to dynamically changing environments.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272461, Grant 62172417, Grant 62276266 and Grant 62277046; in part by the Double First-Class Project of China University of Mining and Technology for Independent Innovation and Social Service under Grant 2022ZZCX06; and in part by the Six Talent Peaks Project in Jiangsu Province under Grant 2015-DZXX-010 and Grant 2018-XYDXX-044; and in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Prowince under Grant 123625314; and Funded by the Graduate Innovation Program of China University of Mining and Technology under Grant 123625314.

References

[Bai *et al.*, 2024] Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. In *In Proceedings of the AAAI Conference on Artificial Intelligence*, pages 729–737, 2024.

[Cui *et al.*, 2020] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12452–12461, 2020.

[Ding *et al.*, 2023] Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Causalaf: Causal autoregressive flow for safety-critical driving scenario generation. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *In Proceedings of Machine Learning Research*, pages 812–823. PMLR, 14–18 Dec 2023.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Du *et al.*, 2021] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3936–3945, 2021.

[Du *et al.*, 2024] Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, and Jinzjing Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23375–23384, 2024.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org, 2015.

[Ge *et al.*, 2023] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Hu *et al.*, 2020] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4051, 2020.

[Huang *et al.*, 2022] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens van der Maaten, and Kilian Q. Weinberger. Convolutional networks with dense connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8704–8716, 2022.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *In Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *arXiv*, 2017.

[Lai *et al.*, 2023] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16109–16119, 2023.

[Li *et al.*, 2020] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9638–9647, 2020.

[Li *et al.*, 2021] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3918–3930, 2021.

[Long *et al.*, 2017] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In Doina Precup and Yee Whye Teh, editors, *In Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2208–2217. PMLR, 06–11 Aug 2017.

[Lu *et al.*, 2020] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9108–9117, 2020.

[Peters *et al.*, 2017] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *In Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[Reichenbach and Morrison, 1956] Maria Reichenbach and P. Morrison. *The Direction of Time*. University of California Press, Berkeley, CA, 1956.

[Singha *et al.*, 2023] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 4357–4366, 2023.

[Sun *et al.*, 2022] Tao Sun, Cheng Lu, Tianshuo Zhang, and Haibin Ling. Safe self-refinement for transformer-based domain adaptation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7191–7200, June 2022.

[Tang *et al.*, 2020a] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8722–8732, 2020.

[Tang *et al.*, 2020b] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8722–8732, 2020.

[Wang and Breckon, 2019] Qian Wang and T. Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *In Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[Wang and Deng, 2018] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[Wang *et al.*, 2023] Shanshan Wang, Yiyang Chen, Zhenwei He, Xun Yang, Mengzhu Wang, Quanzeng You, and Xingyi Zhang. Disentangled representation learning with causality for unsupervised domain adaptation. In *In Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 2918–2926, New York, NY, USA, 2023. Association for Computing Machinery.

[Wei *et al.*, 2024] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: task-oriented alignment for unsupervised domain adaptation. In *In Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc.

[Xu *et al.*, 2022] Tongkun Xu, Weihua Chen, Pichao WANG, Fan Wang, Hao Li, and Rong Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. In *International Conference on Learning Representations*, 2022.

[Yang *et al.*, 2023a] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023.

[Yang *et al.*, 2023b] Shuai Yang, Kui Yu, Fuyuan Cao, Lin Liu, Hao Wang, and Jiuyong Li. Learning causal representations for robust domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2750–2764, 2023.

[Yue *et al.*, 2021] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8599–8608, October 2021.

[Yue *et al.*, 2023] Zhongqi Yue, QIANRU SUN, and Hanwang Zhang. Make the u in uda matter: Invariant consistency learning for unsupervised domain adaptation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 26991–27004. Curran Associates, Inc., 2023.

[Zhao *et al.*, 2025] Jiaqi Zhao, Yao Li, Yong Zhou, Wen-Liang Du, Xixi Li, Rui Yao, and Abdulmotaleb El Saddik. Ddci: Unsupervised domain adaptation for remote sensing images based on diffusion causal distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–12, 2025.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16795–16804, 2022.

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

[Zhu *et al.*, 2023] Wenxuan Zhu, Chao Yu, and Qiang Zhang. Causal deep reinforcement learning using observational data. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4711–4719. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.