Target Semantics Clustering via Text Representations for Robust Universal Domain Adaptation

Weinan He[1], **Zilei Wang**[1*], **Yixin Zhang**[1,2]

[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
hwn2018@mail.ustc.edu.cn, {zlwang, zhyx12}@ustc.edu.cn

Abstract

Universal Domain Adaptation (UniDA) focuses on transferring source domain knowledge to the target domain under both domain shift and unknown category shift. Its main challenge lies in identifying common class samples and aligning them. Current methods typically obtain target domain semantics centers from an unconstrained continuous image representation space. Due to domain shift and the unknown number of clusters, these centers often result in complex and less robust alignment algorithm. In this paper, based on vision-language models, we search for semantic centers in a semantically meaningful and discrete text representation space. The constrained space ensures almost no domain bias and appropriate semantic granularity for these centers, enabling a simple and robust adaptation algorithm. Specifically, we propose TArget Semantics Clustering (TASC) via Text Representations, which leverages information maximization as a unified objective and involves two stages. First, with the frozen encoders, a greedy search-based framework is used to search for an optimal set of text embeddings to represent target semantics. Second, with the search results fixed, encoders are refined based on gradient descent, simultaneously achieving robust domain alignment and private class clustering. Additionally, we propose Universal Maximum Similarity (UniMS), a scoring function tailored for detecting open-set samples in UniDA. Experimentally, we evaluate the universality of UniDA algorithms under four category shift scenarios. Extensive experiments on four benchmarks demonstrate the effectiveness and robustness of our method, which has achieved state-of-the-art performance.

Code — https://github.com/Sapphire-356/TASC

Introduction

Deep neural networks have achieved remarkable success across various computer vision tasks (Carion et al.; Dosovitskiy et al.; He et al.; Deng et al.; He et al.). However, the high cost of annotated data and the limitation of the independent and identically distributed (i.i.d.) assumptions between training and test datasets pose challenges in practical applications. To tackle these issues, Unsupervised Domain Adaptation (DA) (Pan and Yang; Ben-David et al.; Ganin and Lempitsky; Long et al.; Li et al.; Zhang et al.; Li et al.; Zhang, Li, and Wang; Gao et al.) has emerged, aiming to transfer models trained on labeled source domains to unlabeled target domains. Although DA has shown success, traditional DA methods rely on the closed-set assumption, which assumes that the source and target domains share the same label set, and this assumption can be easily violated in real-world scenarios. In light of this, Open-set DA (ODA) (Saito et al. 2018) and Partial DA (PDA) (Cao et al. 2018) consider the existence of private classes in the target and source domains, respectively. However, in ODA and PDA, prior knowledge about the locations of private classes is still required. To adapt to more general scenarios, Universal Domain Adaptation (UniDA) (You et al. 2019) has been proposed, aiming to achieve DA without any prior knowledge of category shifts, i.e., UniDA methods should be able to handle Closed-set DA (CDA), ODA, PDA, and Open-Partial DA (OPDA) simultaneously.

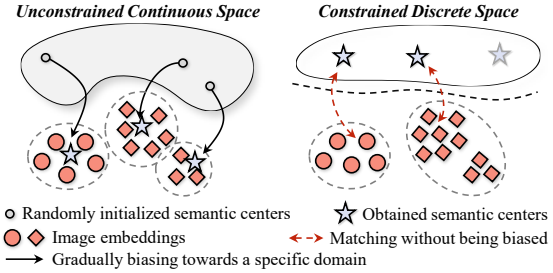


Figure 1: Illustration of our core idea. The left part abstractly represents the updating process of semantic centers in existing clustering-based UniDA methods. These centers gradually become domain-specific and have uncontrollable semantic granularity, such as separating diamonds into two categories. In the right part, we constrain the search space of semantic centers to a semantically meaningful and discrete space, alleviating the issues of domain bias and inappropriate semantic granularity.

Currently, researchers have proposed many solutions, among which the target domain clustering based approach has been widely adopted (Saito et al. 2020; Li et al. 2021a; Chang et al. 2022; Qu et al. 2023). Target domain clustering can effectively mine the intrinsic structure of the tar-

*Corresponding author.

get domain to enhance discriminability, which is almost immune to category shift. Nevertheless, a more critical issue in UniDA is common class detection for cross-domain alignment. In existing methods, the prototype-based classifier in the source domain and the cluster centers in the target domain, both obtained from a continuous image representation space, are expected to accurately represent the corresponding semantic categories. Based on these semantic centers, common class detection is achieved through cluster-level matching (Li et al. 2021a) or sample-semantic center matching (Saito et al. 2020; Chang et al. 2022; Qu et al. 2023), which are all based on representation similarity. However, these semantic centers are actually difficult to use for UniDA: 1) they are domain biased; 2) the number of target domain clusters (or semantic granularity) is hard to estimate. The former makes the representation similarity unreliable and challenging to discriminate common classes across domains. This leads to complex matching and alignment mechanisms (Chang et al. 2022; Li et al. 2021a). The latter causes the algorithm less robust when facing different category shift scenarios as we cannot determine which domain has private classes and how many there are. Consequently, (Chang et al. 2022; Qu et al. 2023) are compelled to assume that private classes must exist in the target domain.

Recently, the emergence of Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021), has offered a promising alternative for visual representation learning. Benefiting from training on web-scale image-text pairs, CLIP exhibits strong cross-modal matching capabilities. The encoded text representations of CLIP possess a certain extent of domain generalization. Moreover, it provides decent representations for open-world images. Therefore, we believe that CLIP provides significant advantages for addressing domain shift and detecting private samples in UniDA. However, how to better leverage these capabilities in CLIP to address the UniDA problem remains largely unexplored.

Considering the above analysis, in this paper, our core idea (Fig. 1) is that uniformly representing the image semantics of both domains in the text representation space to facilitate simple and robust UniDA algorithms on CLIP. For source domain, we obtain a set of embeddings from class names, as its semantic centers. For target domain, we search for a set of optimal text embeddings from the semantically meaningful and discrete text representation space. Within the constrained search space, our algorithms ensure that the two sets of embeddings possess the following properties: 1) almost no domain bias, and the embeddings between common classes across two domains are identical or very close; 2) appropriate semantic granularity. These properties enable: 1) simple matching of common classes through similarity and unifying the losses of common class alignment and private class clustering; 2) simple estimation of the number of private classes, and being aware of category shift.

Specifically, we propose TArget Semantics Clustering (TASC) via Text Representations. It employs information maximization as a unified optimization objective to robustly adapt the model to the target domain in the presence of unknown category shifts. Mathematically, TASC is formulated as a Mixed-Integer Nonlinear Programming problem, which we solve through a two-stage optimization process. In the first stage, the continuous parameters of encoders are fixed, and a set of text embeddings representing the target semantics is searched from a semantically meaningful and discrete text representation space based on a greedy search framework. In the second stage, the discrete variables in the search results are fixed, and the encoders in the continuous parameter space are further optimized based on gradient descent, achieving robust domain adaptation and target private class clustering simultaneously. Additionally, based on the semantic centers of both domains, we explicitly model the category shift. Benefiting from this, we design the Universal Maximum Similarity (UniMS), which is capable of perceiving category shift and is tailored for detecting target private sample in the UniDA task. We optimize the Gaussian Mixture Model to obtain the optimal threshold instead of using hand-tuned ones as in previous approaches.

Experimentally, we simultaneously evaluate UniDA algorithms under three types of category shift scenarios (OPDA, ODA, PDA) and one non-category shift scenario (CDA). We compare our method with existing ones using H-score metric on four common benchmarks. Extensive experimental results demonstrate that our method is sufficiently robust and achieves state-of-the-art performance.

Our contributions are summarized as follows:

- We propose to uniformly represent the image semantics in the semantically rich and discrete text representation space to facilitate simple and robust UniDA algorithms.
- We propose TArget Semantics Clustering (TASC) via Text Representations, which employs information maximization as a unified objective and involves two optimization stages. It achieves robust domain adaptation and target private class clustering.
- We propose Universal Maximum Similarity (UniMS), which is capable of perceiving category shifts and accurately detecting target private samples.
- We evaluate UniDA algorithms under four different category shift scenarios. Our method is robust and achieves state-of-the-art performance on four benchmarks.

Related Work

Universal Domain Adaptation. UniDA (You et al. 2019) aims to address the domain adaptation without prior knowledge of label set relationship. (You et al. 2019; Fu et al. 2020; Zhu et al. 2023) propose multiple criteria for unknown detection. (Saito and Saenko 2021; Yang et al. 2022) design the special classifiers. In (Chen et al.; Chen et al.; Saito et al.; Chen et al.; Lu et al.), neighborhood structures are exploited. (Deng and Jia 2023) explores the foundation models for UniDA. (Liu et al. 2023) address UniDA with few-shot settings. Recently, target domain clustering (Qu et al. 2023; Chang et al. 2022; Li et al. 2021a) has been developed to discover target domain categories and detect common class samples. However, all these methods search for semantic centers in the sub-optimal unconstrained continuous image representation space. In contrast, we explore the semantically meaningful and discrete text representation space.

Vision Task with Vision-language Models. Recently, Vision-language Models (VLMs) have attracted increasing attention in multiple vision tasks (Zhang et al.; Radford et al.; Yao et al.; Li et al.). In this paper, we focus on the image clustering (Cai et al. 2023; Li et al. 2023; Joseph et al. 2022) and open-set adaptation (Min et al. 2023; Zara et al. 2023; Yu, Irie, and Aizawa 2023). (Min et al. 2023) leveraging open-set unlabeled data in the wild for open-set task adaptation. (Zara et al. 2023) consider the adaptation of action recognition model in the open-set scenario. (Yu, Irie, and Aizawa 2023) explore the potential of CLIP for ODA. (Cai et al. 2023; Li et al. 2023) propose leveraging external knowledge from text modality to facilitate clustering. Moreover, (Han et al. 2022) introduces a new task of obtaining class names for unlabeled datasets. In this paper, we emphasize the role of the semantically meaningful text representation space in developing a simple and robust UniDA method.

Preliminary

Before we describe the details of our method, we firstly present the preliminaries used in our framework and formalize Universal Domain Adaptation (UniDA).

Model. In this paper, we focus on adapting the vision-language model CLIP (Radford et al. 2021) to target domain in the UniDA settings. CLIP consists of image encoder f and text encoder g. For a given image $\mathbf{x}$ and a set of class names $\mathcal{T}^s = \{t_1^s, t_2^s, \dots, t_m^s\}$, CLIP can make prediction by comparing image embedding with the text embeddings. Let's denote $\mathbf{z} = f(\mathbf{x})$ and $\mathbf{s}_j = g(t_j^s)$ as the L_2-normalized embeddings of the image and class name j respectively. Then, the probability that $\mathbf{x}$ belongs to class i is calculated as

$$p_i = P(\mathbf{s}_i|\mathbf{z};\tau) = \frac{\exp(\text{sim}(\mathbf{z},\mathbf{s}_i)/\tau)}{\sum_{j=1}^{m}\exp(\text{sim}(\mathbf{z},\mathbf{s}_j)/\tau)}, \quad (1)$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity and τ is the softmax temperature. For simplicity, let define the function h as:

$$h(\mathbf{z};\mathbf{S},\tau) \triangleq [p_1, p_2, \dots, p_m]^T = \mathbf{p}, \quad (2)$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_m]$. Also for simplicity, we omit the prompting strategy as default. Actually, we use the ensemble text templates from (Lin et al.). To efficiently transfer the CLIP model to the target domain, inspired by (Smith et al.; Doveh et al.; Cascante-Bonilla et al.), we fine-tune both the image encoder f and text encoder g via LoRA (Hu et al.).

Universal Domain Adaptation. In the UniDA problem, we are given a labeled source domain $\mathcal{D}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and an unlabeled target domain $\mathcal{D}^t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, where a domain gap exists between them. We denote $\mathcal{C}_s$ and $\mathcal{C}_t$ as the label sets of source and target domain respectively. Additionally, let $\mathcal{T}^s$ and $\mathcal{T}^t$ denote the sets of the class names. Notably, we lack any prior information about $\mathcal{C}_t$ and $\mathcal{T}^t$ during training. The common label set is represented by $\mathcal{C} = \mathcal{C}_s \cap \mathcal{C}_t$. Let $\overline{\mathcal{C}}_s = \mathcal{C}_s \backslash \mathcal{C}$ and $\overline{\mathcal{C}}_t = \mathcal{C}_t \backslash \mathcal{C}$ as label sets of source-private and target-private, respectively. UniDA aims to train a model on $\mathcal{D}^s$ and $\mathcal{D}^t$ that can accurately classifies the target domain common class samples into $|\mathcal{C}|$ classes and, if target-private class samples exist, assigns them to a single *unknown* class.

Notations. First, all vectors in this paper are column vectors. When a matrix is constructed from vectors, these vectors are organized in a column-wise manner. Let $\mathbf{W}^s = [\mathbf{w}_1^s, \mathbf{w}_2^s, \dots, \mathbf{w}_{|\mathcal{C}_s|}^s]$ denote the text embeddings of $\mathcal{T}^s$, where $\mathbf{w}_j^s = g(t_j^s)$. Additionally, Given the prototypes and temperature, the entropy of an embedding can be defined based on predicted probabilities using prototypes. For simplicity, on the basis of Eq. (2), we define the this function:

$$\text{Entropy}(\mathbf{p}) \triangleq -\sum_{i=1}^{m} p_i \log p_i, \quad (3)$$

where $\mathbf{p} = h(\mathbf{z};\mathbf{S},\tau)$ and p_i is the i-th item in $\mathbf{p}$.

Method

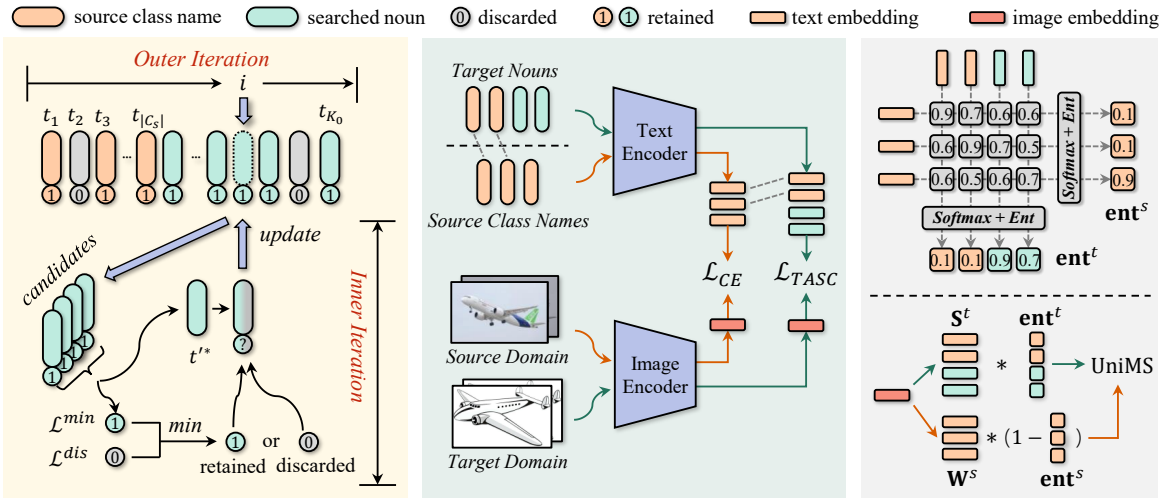
In this work, we aim to solve the challenging UniDA problem with Vision-Language Models (VLMs). As mentioned in the preliminary, we lack any prior knowledge about $\mathcal{C}_t$ and $\mathcal{T}^t$, resulting in uncertainty about the number of private classes in both domains. Empirical evidence (You et al.; Fu et al.; Saito et al.; Cao et al.) suggests that sub-optimal domain alignment will emerge when target-common or target-private classes are mistakenly aligned with source-private or source-common classes, respectively. Therefore, the primary challenge of UniDA lies in common class detection and the subsequent domain alignment.

Existing Clustering-based Methods

Recently, numerous target domain clustering-based UniDA methods (Li et al. 2021a; Chang et al. 2022; Qu et al. 2023) have been proposed and achieved promising results. In this paper, we consider that the key prerequisite for the effectiveness of these methods can be summarized as accurately representing the semantics of both domains:

- DCC (Li et al. 2021a): DCC utilizes source labels to compute source semantic centers and employs K-means centroids as target target semantic centers. Subsequently, common class matching is performed by cycle-consistent matching which is based on representation similarity.
- GLC (Qu et al. 2023): For each class, in the target domain, GLC acquires positive prototypes using prediction results and negative prototypes via K-means. Then, GLC identifies common class samples based on similarity between samples and prototypes.
- UniOT (Chang et al. 2022): UniOT first calculates the similarity between target samples and source prototypes, and then solves the optimal transport (OT) problem to classify samples as either common or unknown. Moreover, UniOT also obtains target prototypes through OT within the target domain.

In summary, although these methods employ different techniques, they all aim to acquire semantic centers (i.e., prototypes or cluster centers) of both domains and detect common classes based on representation similarity. However, the derived semantic centers are less than ideal, leading to complex and less robust algorithms. We argue that this is caused by two factors: 1) the derived semantic centers are



(a) Greedy Search-based Discrete Optimization (b) Model Refinement (c) UniMS

Figure 2: Overview of the proposed Target Semantics Clustering via Text Representations.

domain-biased; 2) the number of classes in target domain is unknown. These factors reduce the reliability of similarity and make it difficult to determine which domain contains private classes and how many such classes exist. Unfortunately, current methods are powerless against them.

In this paper, we argue that this difficulty largely stems from the use of an unconstrained continuous image representation space in the process of representing semantics. First, these semantic centers are iteratively updated based on image representations in an unconstrained space lacking domain-invariant regularization, leading to the incorporation of domain-specific information. Moreover, the points in this space are almost semantically meaningless and can possess any level of semantic granularity. Consequently, regardless of the number of clusters, the clustering loss can be effectively minimized, leading to the inability to estimate the number of clusters.

Based on these analyses, a straightforward idea emerges: constraining the search space of semantic centers to a semantically meaningful and less domain-biased representation space to facilitate simple and robust UniDA algorithm.

Target Semantics Clustering (TASC)

In this section, we will provide a concrete implementation of this idea. Since the embeddings of $\mathcal{T}^s$ are sufficient to serve as the source semantic centers, we focus on the target domain. We first propose the mathematical formulation of Target Semantics Clustering (TASC) and then provide a two-stage optimization method, as shown in Figure 2.

Mathematical Formulation. Benefiting from training on web-scale image-text pairs, text representations encoded by CLIP are inherently semantically meaningful and exhibit less domain bias. Therefore, we leverage the embeddings of source class names $\mathcal{T}^s = (t_1^s, t_2^s, \ldots, t_{|\mathcal{C}_s|}^s)$ and all nouns from WordNet (Miller 1995) $\mathcal{T}^{nouns} = (t_1^n, t_2^n, \ldots, t_N^n)$ to construct the search space, where N is the number of nouns. Base on this, we can formulate the target domain clustering as the process of finding the optimal set of nouns $\mathcal{T}$ and the optimal model parameters θ to minimize $\mathcal{L}_{clu}$, i.e.,

$$\begin{aligned} \min_{\mathcal{T},\theta} \quad & \mathcal{L}_{clu}(\mathcal{T}, \theta; \mathcal{D}^t) \\ \text{s.t.} \quad & \mathcal{T} = (t_1, t_2, \ldots, t_K) \\ & t_i \in \mathcal{T}^s \cup \mathcal{T}^{nouns} \quad \forall\, i = 1, 2, \ldots, K \end{aligned} \tag{4}$$

where K is the number of clusters. During optimization, the embeddings of $\mathcal{T}$ will serve as target semantic centers.

Although the above formulation constrains the search space of semantic centers, it relies on a predefined number of clusters K. To enable the optimization of the number of clusters, we additionally introduce a hidden state vector as: $\mathbf{r} = [r_1, r_2, \ldots, r_{K_0}] \in \{0, 1\}^{K_0}$, where K_0 serves as the fixed upper bound of K. The value of r_i represents the status of t_i, i.e., whether t_i participates in the clustering: if $r_i = 1$, t_i is retained; if $r_i = 0$, t_i is discarded. Therefore, K is dynamically determined by $\mathbf{r}$, i.e., $K = \sum_{i=1}^{K_0} r_i$. Finally, the TArget Semantics Clustering (TASC) via Text Representations can be formulated as:

$$\begin{aligned} \min_{\mathcal{T},\mathbf{r},\theta} \quad & \mathcal{L}_{TASC}(\mathcal{T}, \mathbf{r}, \theta; \mathcal{D}^t) \triangleq \mathcal{L}_{clu}(\mathcal{T}^{\mathbf{r}}, \theta; \mathcal{D}^t) \\ \text{s.t.} \quad & \mathcal{T} = (t_1, t_2, \ldots, t_{K_0}) \\ & t_i \in \mathcal{T}^s \cup \mathcal{T}^{nouns} \quad \forall\, i = 1, 2, \ldots, K_0 \\ & \mathbf{r} = [r_1, r_2, \ldots, r_{K_0}]^T \in \{0, 1\}^{K_0} \\ & \mathcal{T}^{\mathbf{r}} = (t_i | r_i = 1, i = 1, 2, \ldots, K_0) \end{aligned} \tag{5}$$

where $\mathcal{T}^{\mathbf{r}}$ consists of all retained nouns from $\mathcal{T}$ based on $\mathbf{r}$. The text embeddings of $\mathcal{T}^{\mathbf{r}}$ will serve as the target domain semantic centers, denoted as $\mathbf{S}^t = [\mathbf{s}_1^t, \mathbf{s}_2^t, \ldots, \mathbf{s}_K^t]$.

In terms of the loss function, we leave the explicit functional form of $\mathcal{L}_{clu}$ unspecified in the above discussion. This demonstrates that TASC is essentially an optimization framework that constrains the search space of clustering centers and enables estimating of the number of clusters. In this

paper, inspired by numerous works in source-free domain adaptation (Liang, Hu, and Feng; Zhang, Wang, and He), we adopt information maximization as the objective for clustering. Using the notations in Eq. (3), we instantiate $\mathcal{L}_{clu}$ as:

$$\begin{aligned}\mathcal{L}_{clu} &= \mathcal{L}_{ent} + \lambda_{div}\mathcal{L}_{div}, \\ &= \mathbb{E}_{\mathbf{x}^t \in \mathcal{D}^t} \text{Entropy}(\mathbf{p}(\mathbf{x}^t)) - \lambda_{div}\text{Entropy}(\bar{\mathbf{p}}),\end{aligned} \tag{6}$$

where $\mathbf{p}(\mathbf{x}^t) = h(f(\mathbf{x}^t); \mathbf{S}^t, \tau)$ is the prediction score of $\mathbf{x}^t$, $\bar{\mathbf{p}} = \mathbb{E}_{\mathbf{x}^t \in \mathcal{D}^t}\mathbf{p}(\mathbf{x}^t)$ is the mean prediction score of all samples, and λ_{div} is a trade-off hyperparameter.

We have now successfully formulated the Target Semantics Clustering via Text Representations. However, as evident from Eq. (5) and Eq. (6), this is an extremely challenging Mixed-Integer Nonlinear Programming (MINLP) (Belotti et al. 2013) problem. To address this, in this paper, we offer a practicable solution that consists of two stages.

Greedy Search-based Discrete Optimization. In this optimization stage, the model parameters are fixed, and only the discrete variables, i.e., $\mathcal{T}$ and $\mathbf{r}$, are optimized. During initialization, $\mathbf{r}$ is set to an all-ones vector. For $\mathcal{T}$, we place $\mathcal{T}^s$ at the front part of it and randomly initialize the remaining $K_0 - |\mathcal{C}_s|$ nouns, i.e.,

$$\mathcal{T} = (t_1^s, t_2^s, \ldots, t_{|\mathcal{C}_s|}^s, t_{|\mathcal{C}_s|+1}, \ldots, t_{K_0}). \tag{7}$$

Due to the excessively large search space of $\mathcal{T}$ and $\mathbf{r}$, we employ a greedy optimization strategy. Specifically, at the i-th step, we optimize only the noun t_i and its status r_i while keeping the other variables in $\mathcal{T}$ and $\mathbf{r}$ fixed. Regarding a single step, we provide the following summary and analysis:

1. First, construct the feasible space to be searched. For t_i, randomly select n_c candidates $\mathcal{T}^c$ from $\mathcal{T}^{nouns}$; for r_i, it is binary, being either 0 or 1. Moreover, when $r_i = 0$, t_i will not participate in calculating $\mathcal{L}_{TACS}$ since it is discarded. Therefore, we only need to explore $(n_c + 1)$ possible solutions. Let use $\mathcal{T}_{i|t'}$ to denote that the i-th item of $\mathcal{T}$ is replaced by t' and similarly for $\mathbf{r}_{i|0}$ and $\mathbf{r}_{i|1}$.
2. Then, find the optimal t'^* from the n_c candidates when $r_i = 1$, as follows:

$$\mathcal{L}^{min} = \min_{t' \in \mathcal{T}^c} \mathcal{L}_{TASC}(\mathcal{T}_{i|t'}, \mathbf{r}_{i|1}, \theta\,; \mathcal{D}^t). \tag{8}$$

   The loss when $r_i = 0$ is calculated as:

$$\mathcal{L}^{dis} = \mathcal{L}_{TASC}(\mathcal{T}, \mathbf{r}_{i|0}, \theta\,; \mathcal{D}^t). \tag{9}$$

3. Finally, update the discrete variables. Use $\mathcal{T}_{i|t'^*}$ to update $\mathcal{T}$. Set r_i to 1 if $\mathcal{L}^{min} < \mathcal{L}^{dis}$; otherwise, set it to 0.

Based on the single step update, we traverse i from 1 to K_0 to perform a total of K_0 steps, constituting one outer iteration. Throughout the entire discrete optimization, we will conduct N_{outer} outer iterations. To enhance clarity and rigor, we've summarized the above process in Appendix.

It is worth mentioning that, from the perspective of domain adaptation, we should ensure that the semantic centers of common classes across the two domains are identical or highly similar, so that the target domain clustering can substantially contribute to enhancing the classifier's performance. To this end, for the first $|\mathcal{C}_s|$ items in $\mathcal{T}$ and $\mathbf{r}$, we employ two dedicated designs: 1) Starting from the initialization, never update the first $|\mathcal{C}_s|$ items in $\mathcal{T}$. 2) Obtain target domain prototypes $\boldsymbol{\mu}$ using the current predictions based on $\mathcal{T}^{\mathbf{r}}$. When $\text{Entropy}(h(g(t_i^s); \boldsymbol{\mu}, \tau)) < \gamma_{ent}$, set r_i to 1; otherwise, update r_i based on $\mathcal{L}_{min}$ and $\mathcal{L}_{dis}$ as normal.

After this stage, we obtain the optimal $\mathcal{T}^*$ and $\mathbf{r}^*$. Due to the dedicated designs, the source-private classes initially included in $\mathcal{T}$ will be adaptively discarded, while the common classes are retained, and the unknown target-private classes will be represented by the searched nouns.

Model Refinement. In this optimization stage, the discrete variables ($\mathcal{T}^*$ and $\mathbf{r}^*$) are fixed, and only the model parameters θ are optimized. Consequently, based on Eq. (5), TASC reduces to a standard neural network optimization problem that can be addressed via gradient descent:

$$\min_{\theta} \quad \mathcal{L}_{TASC}(\mathcal{T}^*, \mathbf{r}^*, \theta\,; \mathcal{D}^t) = \mathcal{L}_{clu}(\mathcal{T}^{*\mathbf{r}^*}, \theta\,; \mathcal{D}^t). \tag{10}$$

Moreover, consistent with existing DA methods, we utilize the Cross Entropy loss on source domain to guide the adaptation of the generic VLMs to the target task. In summary, we adopt the following losses for model refinement:

$$\min_{\theta} \quad \mathcal{L}_{all} = \mathcal{L}_{CE} + \mathcal{L}_{TASC}. \tag{11}$$

It is worth noting that, we calculate $\mathcal{L}_{CE}$ using the source domain semantic centers $\mathbf{W}^s$ obtained from $\mathcal{T}^s$; whereas in $\mathcal{L}_{TASC}$, we compute it using the target domain semantic centers $\mathbf{S}^t$ derived from $\mathcal{T}^{*\mathbf{r}^*}$. Benefiting from the the dedicated designs in the discrete optimization, the semantic centers of common classes in $\mathbf{W}^s$ and $\mathbf{S}^t$ are nearly identical, enabling $\mathcal{L}_{TASC}$ achieves common class alignment and private class clustering simultaneously. Additionally, from the perspective of loss design, we employ the information maximization criterion derived from Closed-set DA, achieving simple and robust Universal DA, without explicitly identifying target-private samples during optimization.

Universal Maximum Similarity

During inference, we need to assign target-private samples to a single *unknown* class. For unknown detection, existing methods often construct a scoring function based on entropy (You et al.; Fu et al.; Saito et al.), confidence (Fu et al. 2020), similarity (Li et al.; Chang et al.), etc., with a manual threshold for the final decision. However, they lack robustness across different category shift scenarios due to their inability to accurately perceive category shifts. In this paper, based on the high-quality semantic centers, we first explicitly model the category shift and then embed it into the similarity score for more robust unknown detection.

First, functionally speaking, both $\mathbf{S}^t$ and $\mathbf{W}^s$ can serve as image classifiers. More broadly, we can utilize them to classify for each other. Based on the classification scores, the following two entropy vectors can be further defined:

$$\mathbf{ent}^s = [ent_1^s, ent_2^s, \ldots, ent_{|\mathcal{C}_s|}^s]^T \in [0,1]^{|\mathcal{C}_s|}, \tag{12}$$

$$\mathbf{ent}^t = [ent_1^t, ent_2^t, \ldots, ent_K^t]^T \in [0,1]^K, \tag{13}$$

where $ent_i^s = \text{Entropy}(h(\mathbf{w}_i^s; \mathbf{S}^t, \tau))/\log K$ and $ent_j^t = \text{Entropy}(h(\mathbf{s}_j^t; \mathbf{W}^s, \tau))/\log|\mathcal{C}_s|$ are the normalized entropy

Method		Office					Office-Home					DomainNet							VisDA
		OPDA	ODA	PDA	CDA	**Avg**	OPDA	ODA	PDA	CDA	**Avg**	PR	PS	RP	RS	SP	SR	**Avg**	SR
SO	ResNet50	64.9	69.6	87.8	-	-	60.9	55.2	62.9	-	-	57.3	38.2	47.8	38.4	32.2	48.2	43.7	25.7
DCC	ResNet50	80.2	72.7	93.3	-	-	70.2	61.7	70.9	-	-	56.9	43.7	50.3	43.3	44.9	56.2	49.2	43.0
DANCE	ResNet50	80.3	79.8	86.0	-	-	49.2	12.9	71.1	-	-	21.0	37.0	47.3	46.7	27.7	21.0	33.5	42.8
OVANet	ResNet50	86.5	91.7	74.6	-	-	71.8	64.0	49.5	-	-	56.0	47.1	51.7	44.9	47.4	57.2	50.7	53.1
GATE	ResNet50	87.6	89.5	93.7	-	-	75.6	69.0	74.0	-	-	57.4	48.7	52.8	47.6	49.5	56.3	52.1	56.4
GLC	ResNet50	87.8	89.0	94.1	-	-	75.6	69.8	72.5	-	-	63.3	50.5	54.9	50.9	49.6	61.3	55.1	73.1
UniOT	ResNet50	91.1	-	-	-	-	76.6	-	-	-	-	59.3	51.8	47.8	48.3	46.8	58.3	52.0	57.3
SO*	CLIP	75.8	83.5	95.1	87.3	85.4	76.2	64.6	81.8	79.8	75.6	67.1	63.0	64.0	63.2	56.9	66.9	63.5	51.8
DANCE*	CLIP	89.7	90.7	88.1	87.5	89.0	83.9	65.4	84.6	81.7	78.9	66.0	62.1	65.3	63.9	62.6	66.4	64.4	75.3
UniOT*	CLIP	91.1	93.7	71.2	88.9	86.2	86.3	80.3	74.8	82.1	80.9	72.0	63.8	66.5	66.5	62.0	72.8	67.3	75.3
OVANet*	CLIP	**92.1**	94.5	95.2	86.7	92.1	85.0	76.2	81.9	80.0	80.8	72.5	64.5	65.6	65.2	61.8	72.1	67.0	79.1
Ours	CLIP	91.3	**96.1**	**96.3**	**90.1**	**93.5**	**89.4**	**83.8**	**89.2**	**86.7**	**85.4**	**80.5**	**69.7**	**69.1**	**69.3**	**69.6**	**81.5**	**73.3**	**90.4**

Table 1: Performance comparison between state-of-the-arts and our method. The left part shows the results on Office (average on 6 tasks) and Office-Home (average on 12 tasks) under four different category shift scenarios, and the right part shows the results on large-scale DomainNet and VisDA under OPDA scenarios. We report the H-score for OPDA and ODA, as well as the classification accuracy for PDA and CDA. Some results are referred to previous work (Qu et al. 2023).

$\mathcal{L}_{CE}$	$\mathcal{L}_{TASC}$	UniMS	Office				Avg
			OPDA	ODA	PDA	CDA	
✓			82.9	90.9	95.1	87.3	89.1
✓		✓	89.6	95.2	95.1	87.3	91.8
✓	✓	✓	**91.3**	**96.1**	**96.3**	**90.1**	**93.5**

Table 2: Ablation studies on training loss and scoring function. Experiments conduct on Office (average on 6 tasks).

Weights	Office		Office-Home		Avg
	OPDA	ODA	OPDA	ODA	
✗	91.0	**96.3**	84.9	81.9	88.5
✓	**91.3**	96.1	**89.4**	**83.8**	**90.2**

Table 3: H-score (%) on Office-Home (average on 12 tasks) and Office (average on 6 tasks) under OPDA and ODA. "Weights" indicates whether the estimated proportion of target-domain private classes is used in GMM optimization.

that bounded in the range of $[0, 1]$. Due to the constrained space and dedicated designs, the representation similarity is reliable enough for $\mathbf{ent}^s$ and $\mathbf{ent}^t$ to perceive category shifts. In other words, lower values of ent_i^s and ent_j^t indicate a higher probability that $\mathbf{w}_i^s$ and $\mathbf{s}_j^t$ correspond to common classes. To leverage this information, we embed both $\mathbf{ent}^s$ and $\mathbf{ent}^t$ into the widely adopted MLS (Vaze et al. 2021) to derive the Universal Maximum Similarity (UniMS) as:

$$\begin{aligned}\text{UniMS}(\mathbf{x}^t) = &\max\{(1 - ent_i^s) * \text{sim}(f(\mathbf{x}^t), \mathbf{w}_i^s)\}_{i=1}^{|\mathcal{C}_s|} \\ &- \max\{ent_j^t * \text{sim}(f(\mathbf{x}^t), \mathbf{s}_j^t)\}_{j=1}^{K}.\end{aligned} \tag{14}$$

Intuitively, if an image embedding $f(\mathbf{x}^t)$ is close to $\mathbf{w}_i^s$ and ent_i^s is near 0, it will obtain a higher UniMS score; conversely, if it is close to $\mathbf{s}_j^t$ and ent_j^t is near 1, the UniMS score will be suppressed.

Inspired by (Jahan and Savakis; Jang et al.), we optimize a 2-component Gaussian Mixture Model (GMM) to obtain the adaptive threshold of UniMS for unknown detection. During optimization, unlike existing works, we utilize the proportion of target-private classes estimated by TASC and fix it as the mixture weights. Specifically, the number of target-private classes can be estimated via $\sum_{i=|\mathcal{C}_s|+1}^{K_0} r_i$. During prediction, we set the mixture weights to be uniform to achieve theoretically optimal results. More details and theoretical proofs are provided in the Appendix.

Experiments

Setup

Dataset. Our method will be validated on four popular datasets in Domain Adaptation, i.e., Office (Saenko et al. 2010), Office-Home (Venkateswara et al.), VisDA (Peng et al. 2018), and DomainNet (Peng et al. 2019). Due to the large amount of data, we only conduct experiments on three subsets from DomainNet, i.e., Painting(P), Real(R), and Sketch(S), following existing works (Chang et al.; Saito and Saenko; Qu et al.). We evaluate our method on 4 different category shift scenarios to demonstrate its robustness, i.e., CDA, PDA, ODA, and OPDA. Detailed classes split in these scenarios are summarized in Appendix, which is the same as dataset split in (Qu et al. 2023).

Implementation Details. We use pre-trained CLIP model with ViT-B/16 (Dosovitskiy et al.) and Transformer (Vaswani et al.) as image and text encoders, respectively. LoRA (Hu et al. 2021) is used in all transformer blocks in both image and text encoders with $rank = 8$. More details about LoRA can be found in the Appendix. We adopt the same learning rate scheduler $\eta = \eta_0 \cdot (1 + 10 \cdot p)^{-0.75}$ as (Long et al.; Liang, Hu, and

Method	Office		Office-Home		Avg
	OPDA	ODA	OPDA	ODA	
MS-t	58.4	73.3	59.4	51.4	60.6
MS-t w/ $\mathbf{ent}^t$	90.3	91.0	84.2	75.5	85.2
MS-s	93.1	<u>99.0</u>	95.1	<u>92.2</u>	94.8
MS-s w/ $\mathbf{ent}^s$	<u>93.4</u>	<u>99.0</u>	<u>95.9</u>	92.0	<u>95.1</u>
UniMS	**96.6**	**99.1**	**96.4**	**92.6**	**96.2**

Table 4: AUROC (%) on Office-Home (average on 12 tasks) and Office (average on 6 tasks) under OPDA and ODA. Let denote the first item of UniMS as MS-s w/ $\mathbf{ent}^s$ and its unweighted version as MS-s; similarly for the second item.

Feng), where p is the training progress changing from 0 to 1 and $\eta_0 = 0.0001$. For the hyper-parameters, we empirically set the λ_{div} to 0.6, which differs from SHOT-IM (Liang, Hu, and Feng), and we will discuss this in details later. τ is set to 0.02. In the discrete optimization step of TASC, $n_c = 300$, $\gamma_{ent} = 0.3$, and $N_{outer} = 20$. K_0 is set to 100 for Office, Office-Home, and VisDA, but 400 for DomainNet.

Evaluation Protocols. For a fair comparison, we follow the same evaluation metric as previous works (Qu et al.; Liu et al.; Fu et al.; Chang et al.; Li et al.). In OPDA and ODA scenarios, we report the H-score (Fu et al. 2020). The H-score, as defined by (Fu et al. 2020), is the harmonic mean of the accuracy $a_{\mathcal{C}}$ on common classes and the accuracy $a_{\bar{\mathcal{C}}_t}$ on a single unknown class. In PDA and CDA scenarios, we report the classification accuracy over all target samples following (Li et al. 2021a; Qu et al. 2023).

Compared Methods. We select multiple state-of-the-art methods for comparison, including DCC (Li et al. 2021a), DANCE (Saito et al.), OVANet (Saito and Saenko), GLC (Qu et al.), UniOT (Chang et al.). Additionally, we utilize CLIP (Radford et al. 2021) as backbone and fine-tune it via LoRA (Hu et al. 2021), while most existing state-of-the-arts fully fine-tune ResNet-50 or ViT-B/16 pre-trained on ImageNet (Deng et al. 2009). Therefore, for a fair comparison, we conduct experiments on following methods under the same conditions as ours: Source-Only (SO), DANCE, OVANet, UniOT, marked as *.

Results

Comparison with state-of-the-arts. As shown in Table 1, our method achieves the best average performance on all the four benchmarks without tuning any hyper-parameters (except K_0). Notably outstanding is that we exceed existing methods by considerable margins of 6.0% and 11.3% on the large-scale DomainNet and VisDA, respectively. All these results demonstrates that our approach is quite effective and more robust when addressing different category shifts.

Ablation studies of TASC. For all experiments in Table 2, we perform the discrete optimization stage of TASC, followed by the selective use of UniMS and $\mathcal{L}_{TASC}$ in the model refinement stage. The consistent improvements demonstrate their effectiveness.

Estimation of the number of clusters. To evaluate the adaptive estimation mechanism, we vary the dataset split and

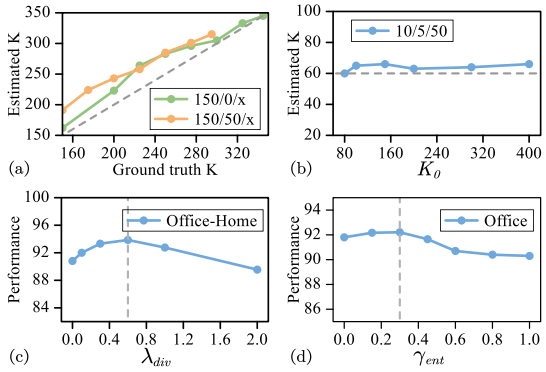


Figure 3: (a) Effectiveness of the adaptive estimation of K. (b) Sensitivity of K_0. (Office-Home, OPDA, AC) (c) Sensitivity of λ_{div}. (Officehome, OPDA, average on AC, CP, PR, RA). (d) Sensitivity of γ_{ent}. (Office, OPDA, average on AD, DW, WA).

plot the estimated K in a line chart. Let use "$|\mathcal{C}|/|\bar{\mathcal{C}}_s|/|\bar{\mathcal{C}}_t|$" to denote different category shifts. In Figure 3(a), we consider two settings: "$150/0/x$" and "$150/50/x$", on SR of DomainNet, where x represents the varying numbers and $K_0 = 400$. As shown by the two color lines, although with the same initial K_0, when varying the number of categories, our algorithm consistently converges towards the ground truth number adaptively.

Effectiveness of perceiving category shifts. First, we conduct ablation studies on the different components of UniMS based on AUROC metric. As shown in Table 4, MS-s serves as a good baseline. Moreover, the category shift information perceived by $\mathbf{ent}^t$ and $\mathbf{ent}^s$ enhances MS-t and MS-s. Finally, MS-t w/ $\mathbf{ent}^t$ provides complementary information to MS-s w/ $\mathbf{ent}^s$, enabling UniMS to achieve the highest performance. Second, we evaluate the effectiveness of setting GMM mixture weights to the estimated proportion of target-private classes by TASC. Table 3 demonstrates that using this proportion is more effective, especially for severely imbalanced category shifts like Office-Home under OPDA (10/5/50) and ODA (25/0/40).

Parameter sensitivity. Figure 3 (b-d) presents our sensitivity analysis on γ_{ent}, K_0, and λ_{div}, showing their stability within specific ranges. Notably, the optimal λ_{div} is lower than 1.0 in SHOT-IM, likely due to partially undiscarded private classes and over-clustering in $\mathcal{T}^{\mathbf{r}}$, necessitating a reduction in the requirement for diversity.

Conclusion

In this paper, we propose that uniformly represent the image semantics of both domains in the semantically rich and discrete text representation space can facilitate simple and robust UniDA algorithm. Based on this idea, we propose target semantics clustering via text representations and universal maximum similarity. Extensive experiments demonstrate the robustness and effectiveness. Finally, considering the universality of TASC, we expect the development of additional UniDA algorithms built upon this framework.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62176246. This work is also supported by Anhui Province Key Research and Development Plan (202304a05020045) and Anhui Province Natural Science Foundation (2208085UD17). This work is also supported by National Natural Science Foundation of China under Grant 62406098 and 62376256, and The Joint Fund for Medical Artificial Intelligence under Grant MAI2022Q011.

References

Belotti, P.; Kirches, C.; Leyffer, S.; Linderoth, J.; Luedtke, J.; and Mahajan, A. 2013. Mixed-integer nonlinear optimization. *Acta Numerica*, 22: 1–131.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79: 151–175.

Cai, S.; Qiu, L.; Chen, X.; Zhang, Q.; and Chen, L. 2023. Semantic-enhanced image clustering. In *AAAI*, volume 37, 6869–6878.

Cao, Z.; Ma, L.; Long, M.; and Wang, J. 2018. Partial adversarial domain adaptation. In *ECCV*, 135–150.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*, 213–229. Springer.

Cascante-Bonilla, P.; Shehada, K.; Smith, J. S.; Doveh, S.; Kim, D.; Panda, R.; Varol, G.; Oliva, A.; Ordonez, V.; Feris, R.; et al. 2023. Going beyond nouns with vision & language models using synthetic data. In *ICCV*.

Chang, W.; Shi, Y.; Tuan, H.; and Wang, J. 2022. Unified optimal transport framework for universal domain adaptation. *NeurIPS*, 35: 29512–29524.

Chen, L.; Du, Q.; Lou, Y.; He, J.; Bai, T.; and Deng, M. 2022a. Mutual nearest neighbor contrast and hybrid prototype self-training for universal domain adaptation. In *AAAI*, volume 36, 6248–6257.

Chen, L.; Lou, Y.; He, J.; Bai, T.; and Deng, M. 2022b. Evidential neighborhood contrastive learning for universal domain adaptation. In *AAAI*, volume 36, 6258–6267.

Chen, L.; Lou, Y.; He, J.; Bai, T.; and Deng, M. 2022c. Geometric anchor correspondence mining with uncertainty modeling for universal domain adaptation. In *CVPR*, 16134–16143.

Deng, B.; and Jia, K. 2023. Universal Domain Adaptation from Foundation Models. *arXiv preprint arXiv:2305.11092*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Doveh, S.; Arbelle, A.; Harary, S.; Schwartz, E.; Herzig, R.; Giryes, R.; Feris, R.; Panda, R.; Ullman, S.; and Karlinsky, L. 2023. Teaching structured vision & language concepts to vision & language models. In *CVPR*.

Fu, B.; Cao, Z.; Long, M.; and Wang, J. 2020. Learning to detect open classes for universal domain adaptation. In *ECCV*, 567–583. Springer.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189. PMLR.

Gao, Y.; Wang, Z.; Zhuang, J.; Zhang, Y.; and Li, J. 2023. Exploit domain-robust optical flow in domain adaptive video semantic segmentation. In *AAAI*, volume 37, 641–649.

Han, K.; LI, Y.; Vaze, S.; and Jia, X. 2022. Semantic Category Discovery with Vision-language Representations.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jahan, C. S.; and Savakis, A. 2024. Unknown sample discovery for source free open set domain adaptation. In *CVPR*, 1067–1076.

Jang, J.; Na, B.; Shin, D. H.; Ji, M.; Song, K.; and Moon, I.-C. 2022. Unknown-aware domain adversarial learning for open-set domain adaptation. *NeurIPS*, 35: 16755–16767.

Joseph, K.; Paul, S.; Aggarwal, G.; Biswas, S.; Rai, P.; Han, K.; and Balasubramanian, V. N. 2022. Novel class discovery without forgetting. In *ECCV*, 570–586. Springer.

Li, G.; Kang, G.; Zhu, Y.; Wei, Y.; and Yang, Y. 2021a. Domain consensus clustering for universal domain adaptation. In *CVPR*, 9757–9766.

Li, J.; Zhang, Y.; Wang, Z.; Hou, S.; Tu, K.; and Zhang, M. 2024. Probabilistic Contrastive Learning for Domain Adaptation. arXiv:2111.06021.

Li, J.; Zhang, Y.; Wang, Z.; and Tu, K. 2021b. Semantic-aware representation learning via probability contrastive loss.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *CVPR*, 10965–10975.

Li, Y.; Hu, P.; Peng, D.; Lv, J.; Fan, J.; and Peng, X. 2023. Image clustering with external guidance. *arXiv preprint arXiv:2310.11989*.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 6028–6039. PMLR.

Lin, Z.; Yu, S.; Kuang, Z.; Pathak, D.; and Ramanan, D. 2023. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *CVPR*, 19325–19337.

Liu, X.; Zhou, Y.; Zhou, T.; Feng, C.-M.; and Shao, L. 2023. COCA: Classifier-Oriented Calibration for Source-Free Universal Domain Adaptation via Textual Prototype. *arXiv preprint arXiv:2308.10450*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*, 97–105. PMLR.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *NeurIPS*, 31.

Lu, Y.; Shen, M.; Ma, A. J.; Xie, X.; and Lai, J.-H. 2024. MLNet: Mutual Learning Network with Neighborhood Invariance for Universal Domain Adaptation. In *AAAI*, volume 38, 3900–3908.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Min, Y.; Ryoo, K.; Kim, B.; and Kim, T. 2023. UOTA: Unsupervised Open-Set Task Adaptation Using a Vision-Language Foundation Model. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *ICCV*, 1406–1415.

Peng, X.; Usman, B.; Kaushik, N.; Wang, D.; Hoffman, J.; and Saenko, K. 2018. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *CVPR Workshops*, 2021–2026.

Qu, S.; Zou, T.; Röhrbein, F.; Lu, C.; Chen, G.; Tao, D.; and Jiang, C. 2023. Upcycling models under domain and category shift. In *CVPR*, 20019–20028.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*, 213–226. Springer.

Saito, K.; Kim, D.; Sclaroff, S.; and Saenko, K. 2020. Universal domain adaptation through self supervision. *NeurIPS*, 33: 16282–16292.

Saito, K.; and Saenko, K. 2021. Ovanet: One-vs-all network for universal domain adaptation. In *ICCV*, 9000–9009.

Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018. Open set domain adaptation by backpropagation. In *ECCV*, 153–168.

Smith, J. S.; Cascante-Bonilla, P.; Arbelle, A.; Kim, D.; Panda, R.; Cox, D.; Yang, D.; Kira, Z.; Feris, R.; and Karlinsky, L. 2023. Construct-vl: Data-free continual structured vl concepts learning. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2021. Open-set recognition: A good closed-set classifier is all you need? *arXiv preprint arXiv:2110.06207*.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 5018–5027.

Yang, S.; Wang, Y.; Wang, K.; Jui, S.; and van de Weijer, J. 2022. OneRing: A Simple Method for Source-free Open-partial Domain Adaptation. *arXiv preprint arXiv:2206.03600*.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

You, K.; Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2019. Universal domain adaptation. In *CVPR*, 2720–2729.

Yu, Q.; Irie, G.; and Aizawa, K. 2023. Open-set domain adaptation with visual-language foundation models. *arXiv preprint arXiv:2307.16204*.

Zara, G.; Roy, S.; Rota, P.; and Ricci, E. 2023. AutoLabel: CLIP-based framework for Open-set Video Domain Adaptation. In *CVPR*, 11504–11513.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y.; Li, J.; and Wang, Z. 2022. Low-confidence samples matter for domain adaptation. *arXiv preprint arXiv:2202.02802*.

Zhang, Y.; Wang, Z.; and He, W. 2023. Class relationship embedded learning for source-free unsupervised domain adaptation. In *CVPR*, 7619–7629.

Zhang, Y.; Wang, Z.; Li, J.; Zhuang, J.; and Lin, Z. 2023. Towards effective instance discrimination contrastive loss for unsupervised domain adaptation. In *ICCV*, 11388–11399.

Zhu, D.; Li, Y.; Yuan, J.; Li, Z.; Kuang, K.; and Wu, C. 2023. Universal domain adaptation via compressive attention matching. In *ICCV*, 6974–6985.