End-to-End Knowledge Distillation for Unsupervised Domain Adaptation with Large Vision-language Models

Yangtao Wang[1], Xingwei Deng[1], Yanzhao Xie[1], Weilong Peng[1], Siyuan Chen[1], Xiaocui Li[2*], Maobin Tang[1], Meie Fang[1]

[1]School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China
[2]Hunan University of Technology and Business, Changsha, China
{ytaowang@gzhu.edu.cn, XingweiDeng@e.gzhu.edu.cn, yzhx@gzhu.edu.cn, wlpeng@tju.edu.cn, chensiyuan@gzhu.edu.cn, Xiaocuiworld@163.com, tmb178@gzhu.edu.cn, fme@gzhu.edu.cn}

Abstract

Knowledge distillation based on large vision-language models (VLMs) has recently emerged as a significant solution to transfer knowledge from the source domain to the target domain in unsupervised domain adaptation (UDA) tasks. However, existing methods employ a two-stage training pipeline, which not only complicates the training procedure but also lacks interactions between the source and target domains, severely hindering real-time cross-domain knowledge transfer. To address these challenges, we propose **E**nd-to-End **K**nowledge **D**istillation for UD**A** with large VLMs (termed as EKDA). (1) EKDA employs a lightweight prompt learning mechanism to first embed the knowledge from the source domain into VLMs, and then simultaneously utilize the image encoder and text encoder of VLMs to perform knowledge distillation on the target domain, significantly reducing the domain gap. (2) EKDA designs a teacher-student alternating training strategy to implement real-time collaborative interactions across domains, enabling an end-to-end paradigm to provide accurate source domain-aware supervision for the target domain. We conduct extensive experiments on 4 widely recognized benchmark datasets including Office-31, Office-Home, VisDA-2017, and Mini-DomainNet. Experimental results demonstrate that EKDA achieves significant performance improvement over the state-of-the-art UDA approaches, while maintaining a much lower model complexity. Take Office-Home for example, EKDA has gained at least 2.7% performance improvement while reducing the learnable parameters by over 80% compared with the state-of-the-art UDA baselines.

Introduction

There has always been a significant bottleneck in the data-centric era: the prohibitive cost and expertise required to acquire annotated data, particularly in specialized fields such as medical imaging and remote sensing (Zhang et al. 2024; Shin et al. 2023; Cai, Shang, and Yin 2024). This results in a striking paradox: despite the abundance of unlabeled data, the scarcity of labeled samples severely limits their deployment in the real world. To bridge this gap, unsupervised domain adaptation (UDA) (Ben-David et al. 2010; Ganin and Lempitsky 2015; Wilson and Cook 2020; Zhu et al. 2023) emerges as a pivotal solution, transferring semantically invariant knowledge from a labeled source domain to another unlabeled target domain. By addressing domain shift (Wilson and Cook 2020) caused by variations in imaging protocols, sensor specifications, or demographic disparities, UDA enables models to generalize effectively across domains with minimal supervision.

Early UDA methods can be primarily categorized into two types: distribution alignment-based and adversarial learning-based approaches. Generally, the former learns domain-invariant features by minimizing the distribution discrepancy between the source and target domains, such as MMD (Long et al. 2015), CORAL (Sun, Feng, and Saenko 2016) and MDD (Zhang et al. 2019), while the latter introduces a domain discriminator to distinguish samples and trains another generator to deceive the discriminator. However, directly completing cross-domain alignment may lead to a significant loss of semantic knowledge and a struggle to handle complex domain differences. To leverage multi-modal semantic alignment for domain adaptation, researchers have begun to explore pretrained large vision-language models (VLMs) like CLIP (Radford et al. 2021) to address domain shifts. To measure the semantic similarity between two domains, PDA (Bai et al. 2024) uses a two-branch paradigm: the base branch integrates class-related representations for discrimination, while the alignment branch builds feature banks and employs cross-attention modules to combine features. Other representative VLMs-based solutions such as DAMP (Du et al. 2024) and UniMoS (Li et al. 2024a) design modality-aware discriminators to respectively align visual and textual embeddings, allowing prompts to capture cross-domain relationships potentially. Although these methods bridge the semantic associations between the source domain and the target domain through VLMs, they require the design of additional cross-attention or discriminative components to achieve domain alignment, which introduces excessive and redundant computational costs.

Recently, VLMs-based knowledge distillation means have been proposed to address domain shifts in a more lightweight manner, requiring only a small number of learnable parameters to achieve domain alignment. For example, KDPL (Mistretta et al. 2024) (see Figure 1(a)) first trains a teacher model with data from the source domain, then di-

*Corresponding author

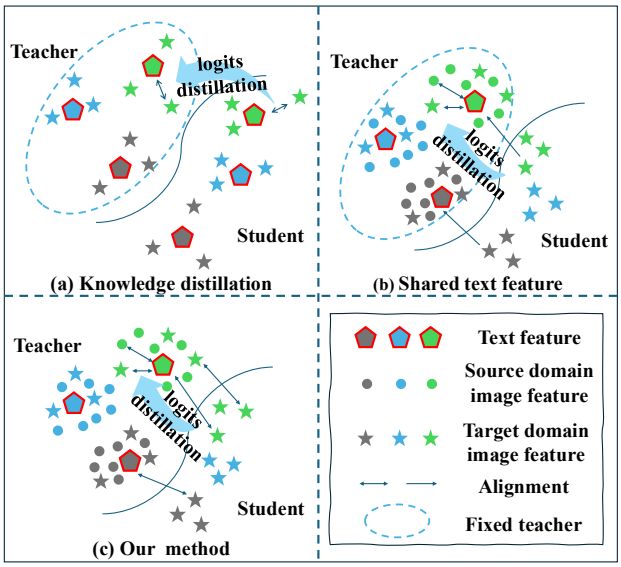

Figure 1: Existing VLMs-based knowledge distillation approaches utilize a fixed teacher model (separated from the target domain) pre-trained on the source domain: (a) conventional methods leverage respective text features (centers) of two domains to implement logits distillation; (b) recent methods leverage shared text features (centers) of two domains to implement logits distillation. (c) Our method distills the knowledge of the teacher into the student in real time, ensuring efficient cross-domain alignment.

rectly implements efficient logits distillation in the student model with data from the target domain, but suffers from a large domain gap due to the separation of text features (centers) between two domains. PromptKD (Li et al. 2024b) (see Figure 1(b)) addresses this issue by using the shared teacher model's text features to guide the student model's visual encoder, significantly reducing the domain gap. However, on the one hand, this two-stage training process is overly cumbersome. On the other hand, once the teacher model is trained, it becomes fixed, resulting in a lack of real-time interaction between the source domain and the target domain, which can lead to large semantic deviations.

To address these challenges, we propose **E**nd-to-End **K**nowledge **D**istillation for UD**A** with large VLMs (termed as EKDA). The core idea is to achieve collaborative optimization (see Figure 1(c)) of cross-domain prompt learning and knowledge distillation through alternating training of teacher-student models. Specifically, on the one hand, the teacher model optimizes only lightweight visual/text prompts based on the source domain data, which can output source domain-aware image features and pseudo-labels for each target domain sample. The student model, on the other hand, implicitly aligns target domain visual features with the teacher's text features through visual prompts and a feature projection layer. Finally, we achieve teacher-student alternating training through end-to-end knowledge distillation, which transfers the knowledge of the source domain to the target domain with real-time cross-domain alignment.

In summary, our work makes three key contributions as follows.

- **New Paradigm**. To the best of our knowledge, this is the first attempt to propose an end-to-end knowledge distillation paradigm that facilitates real-time collaborative interactions between the source domain and target domain, breaking new ground for UDA with VLMs-based cross-domain knowledge transfer.
- **Novel Method.** We introduce a lightweight cross-domain distillation method EKDA that (1) leverages text features from the source domain as shared semantic centers, and (2) extracts source domain-aware image features and pseudo-labels for the target domain to implement real-time knowledge distillation, significantly reducing the domain gap.
- **High Performance.** We conduct extensive experiments on 4 UDA benchmark datasets including Office-31, Office-Home, VisDA-2017, and Mini-DomainNet. Experimental results demonstrate that EKDA achieves significant performance improvement over the state-of-the-art approaches, while maintaining a much lower model complexity.

Related Works

Unsupervised Domain Adaptation

Recent unsupervised domain adaptation (UDA) advancements address distribution shifts through several approaches. Early methods like MMD (Long et al. 2015) and CORAL (Sun, Feng, and Saenko 2016) align feature distributions via statistical matching, while adversarial methods (DANN (Ganin and Lempitsky 2015), ADDA (Tzeng et al. 2017)) learn domain-invariant representations using domain classifiers. Recently, large VLMs (Radford et al. 2021; Jia et al. 2021) enhance UDA via cross-modal capabilities: CMKD (Zhou and Zhou 2024) fine-tunes CLIP's visual encoder at high cost, whereas PDA (Bai et al. 2024), UniMoS (Li et al. 2024a), and DAMP (Du et al. 2024) use frozen CLIP with cross-attention modules but incur computational redundancy. KDPL (Mistretta et al. 2024) employs lightweight distillation yet retains a complex two-stage pipeline with static teachers, limiting cross-domain interaction. To overcome these issues, we propose an end-to-end knowledge distillation paradigm with lightweight prompts for efficient real-time UDA.

Prompt Tuning for VLMs

Large VLMs like CLIP (Radford et al. 2021) are typically trained on massive-scale multi-modal data (e.g., hundreds of millions of image-text pairs) to align visual and textual representations. These pre-trained VLMs exhibit remarkable zero-shot generalization for downstream tasks by leveraging hand-crafted prompts (e.g., "a photo of a [CLASS]") during inference. However, their massive parameter sizes pose great challenges for direct deployment in resource-constrained scenarios. Prompt learning has emerged as an efficient alternative to fine-tuning, where only lightweight learnable

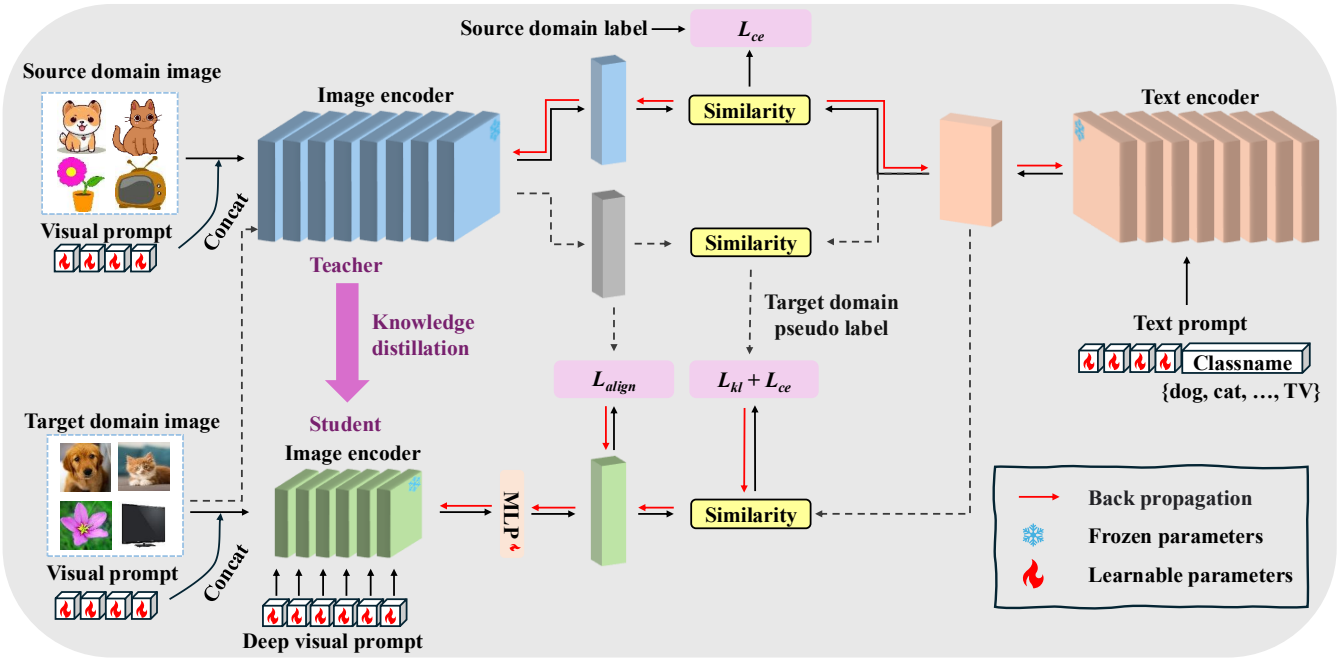

Figure 2: The overall architecture of EKDA. The teacher model generates image features of the source domain and shared text features for each category. Based on this, during the knowledge distillation process, the teacher model transfers the knowledge of the source domain to the student model (on the target domain) in real time, thereby achieving efficient domain alignment.

prompts are optimized while keeping VLMs frozen. For example, CoOp (Zhou et al. 2022b) introduces learnable context vectors to replace static text prompts, enabling task-specific adaptation. Based on this, CoCoOp (Zhou et al. 2022a) extends CoOp with conditional prompts that dynamically adjust to input images. Furthermore, MaPLe (Khattak et al. 2023a) jointly optimizes prompts in both vision and language branches of VLMs, boosting cross-modal alignment. Recently, PDA (Bai et al. 2024) employs a dual-branch prompt-tuning paradigm, combining class-related representations with cross-domain feature alignment to significantly improve UDA's performance. These methods demonstrate that lightweight prompt tuning can significantly boost model adaptability without costly full-model fine-tuning. Inspired by these advances, our approach integrates multi-modal prompt learning with knowledge distillation to fully leverage the cross-modal alignment capabilities of VLMs while minimizing computational overhead.

Knowledge Distillation with VLMs

Knowledge distillation (Cho and Hariharan 2019; Jin, Wang, and Lin 2023; Mirzadeh et al. 2020; Zagoruyko and Komodakis 2017) transfers knowledge from complex teachers to lightweight students via logit matching, feature imitation, or relational alignment. While traditionally single-modality focused, recent work extends distillation to multimodal VLMs: KDPL (Mistretta et al. 2024) enables unsupervised prompt learning without labels; CLIP-KD (Yang et al. 2024) identifies feature mimicry with MSE loss as optimal for CLIP distillation; PromptKD (Li et al. 2024b) aligns logits using KL divergence with teacher-generated prompts. However, these methods employ a two-stage pipeline (pre-training teachers then distilling to fixed students), increasing computational complexity and risking semantic loss due to absent real-time cross-domain interaction. To address this, we propose an end-to-end knowledge distillation approach for UDA that simultaneously trains both teacher and student models, enabling real-time source-to-target knowledge distillation to simplify training and enhance cross-domain alignment.

Proposed Methodology

Overall Architecture

Formally, given a labeled source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and another unlabeled target domain $\mathcal{D}_t = \{x_i^t\}_{i=1}^{N_t}$, y_i^s denotes the ground truth label of x_i^s and both domains share the same label space $\mathcal{Y} \in \{1, 2, \cdots, C\}$. We aim to predict the accurate label for each $x_i^t \in \mathcal{D}_t$ based on given $\mathcal{D}_s$ and $\mathcal{D}_t$. As shown in Figure 2, our proposed EKDA establishes an end-to-end knowledge distillation paradigm with VLMs through multi-modal prompt learning and alternating teacher-student training, which comprises two key components as follows. (1) The teacher model learns transferable representations from source domain data via visual-textual prompt tuning while maintaining CLIP's frozen backbone. (2) The student model aligns the target domain features with the knowledge of the source domain through three distillation objectives using alternating optimization. The teacher model first updates its prompt parameters using source domain supervision,

then guides the student's adaptation via both pseudo-label supervision and feature alignment, enabling real-time knowledge transfer across domains. We further describe the workflow of EKDA below.

Training Teacher Model

Based on the source domain data, we embed the knowledge of each image and its corresponding class label into the teacher model. Following mainstream prompt learning-based methods (Khattak et al. 2023a; Du et al. 2024), we utilize CLIP (Radford et al. 2021) as the foundational model. As illustrated in Figure 2, we freeze the image encoder f_I^{tea} and text encoder f_T^{tea} of the teacher model. For the visual branch, we input the i-th image $x_i^s \in D_s$ along with a learnable visual prompt $P_I^{tea} = \{v_1^{tea}, v_2^{tea}, \cdots, v_M^{tea}\}$ (i.e., M denotes the prompt token length) into the frozen image encoder f_I^{tea}, yielding the image feature $I_i^s = f_I^{tea}(x_i^s, P_I^{tea})$. As for the text branch, the input to the text encoder f_T^{tea} is designed as $t = (P_t^{tea}, [\textit{Classname}])$, where $P_t^{tea} = \{t_1^{tea}, t_2^{tea}, \ldots, t_M^{tea}\}$ represents a learnable text prompt and $[\textit{Classname}]$ denotes one name of C categories (e.g., dog, cat). This input t is then fed into the frozen text encoder f_T^{tea}, yielding C text feature $T = \{T_1, T_2, \cdots, T_C\}$, each of which represents a shared class center. After obtaining each image feature I_i^s and all C text features, we employ a commonly-used contrastive learning loss (Khattak et al. 2023b) to train the teacher model as follows:

$$\mathcal{L}_{ce}^{tea} = -\sum_{i=1}^{N^s} y_i^s \log \frac{\exp\left(\text{sim}\left(I_i^s, T_{y_i}\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(\text{sim}\left(I_i^s, T_c\right)/\tau\right)}, \quad (1)$$

where $\text{sim}(I_i^s, T_{y_i})$ denotes the cosine similarity between the i-th image feature I_i^s and its corresponding text feature T_{y_i}, and τ is the temperature parameter that adjusts the concentration level of the softmax distribution.

Note that after training each batch of source domain images, the teacher model will be temporarily frozen and used to distill the student model. This process involves alternately updating the teacher model and the student model, where the teacher model is updated based on the next batch of source domain images and subsequently employed to refine the student model.

Training Student Model

(1) Similarity comparison with the teacher model. During the student model training, each unlabeled target domain image $x_i^t \in D_t$ is first fed into the teacher model to obtain its corresponding probability distribution and pseudo label based on the shared class centers. Specifically, the teacher model first generates source domain-aware image feature $I_i^t = f_I^{tea}(x_i^t, P_I^{tea})$ for x_i^t, then calculates each similarity score between I_i^t and all C shared text features T as follows:

$$p_{ic}^{tea} = \frac{\exp\left(\text{sim}\left(I_i^t, T_c\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(\text{sim}\left(I_i^t, T_c\right)/\tau\right)}. \quad (2)$$

Based on this, $p_i^{tea} = (p_{i1}^{tea}, p_{i2}^{tea}, \cdots, p_{iC}^{tea})$ denotes the probability distribution of I_i^t from the teacher model. In addition, we can obtain its pseudo-label $\hat{y}_i^t = \arg\max(p_i^{tea})$ based on the category to which the index of the maximum element in the probability distribution belongs.

(2) Similarity comparison with the student model. The same i-th target domain image x_i^t is input into the student model. The image encoder of the student model not only incorporates the visual prompt $P_I^{stu} = \{v_1^{stu}, v_2^{stu}, \cdots, v_M^{stu}\}$ at the initial stage but also introduces deep visual prompt $D_{prompt} = \{d_1, d_2, ..., d_L\}$ at each layer, where L denotes the number of layers in the image encoder. In this way, the student model can more finely adjust and output the feature representation $F_i^t = f_I^{stu}(x_i^t, P_I^{stu}, \{d_1, d_2, ..., d_L\})$, where f_I^{stu} indicates the frozen image encoder of the student model. To align with the feature dimension of the teacher model, F_i^t is transformed through a Multi-Layer Perceptron (MLP) layer $g(\cdot)$: $F_i^{t'} = g(F_i^t)$. Similarly, the student model calculates each similarity score between $F_i^{t'}$ and all C shared text features T as follows:

$$p_{ic}^{stu} = \frac{\exp\left(\text{sim}\left(F_i^{t'}, T_c\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(\text{sim}\left(F_i^{t'}, T_c\right)/\tau\right)}. \quad (3)$$

Based on this, $p_i^{stu} = (p_{i1}^{stu}, p_{i2}^{stu}, \cdots, p_{iC}^{stu})$ denotes the probability distribution of $F_i^{t'}$ from the student model.

(3) Cross-domain knowledge distillation. We implement three key operations to distill knowledge from the teacher model to the student model. **(i) Feature alignment loss:** we design a feature alignment loss to measure the discrepancy between the student model's feature $F_i^{t'}$ and the teacher model's feature I_i^t:

$$\mathcal{L}_{align} = \frac{1}{N^t} \sum_{i=1}^{N^t} \|F_i^{t'} - I_i^t\|_2^2. \quad (4)$$

(ii) Cross-entropy loss: we construct the cross-entropy loss between the student model's prediction p_i^{stu} and its corresponding teacher model's pseudo label $\hat{y}_i^t$:

$$\mathcal{L}_{ce}^{stu} = -\sum_{i=1}^{N^t} \hat{y}_i^t \log(p_i^{stu}). \quad (5)$$

(iii) KL divergence loss: we implement the KL divergence loss to align the student model's probability distribution p_i^{stu} with the teacher model's probability distribution p_i^{tea}:

$$\mathcal{L}_{kl} = \frac{1}{N^t} \sum_{i=1}^{N^t} KL(\phi(p_i^{tea}), \phi(p_i^{stu})), \quad (6)$$

where $KL(\cdot||\cdot)$ denotes the KL divergence between two probability distributions, and $\phi(\cdot)$ denotes the softmax operation.

Finally, we train the student model with the overall student loss:

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{align} + \mathcal{L}_{ce}^{stu} + \mathcal{L}_{kl}. \quad (7)$$

By this means, the student model can learn the feature distribution of the teacher model in real time, thus efficiently accomplishing cross-domain alignment.

Experiments

Experimental Settings

(1) Datasets. We adopt 4 widely-used UDA benchmarks: **Office-31** (Saenko et al. 2010) (domains: Amazon, Webcam, DSLR), **Office-Home** (Venkateswara et al. 2017) (domains: Art, Clipart, Product, Real World), **VisDA2017** (Peng et al. 2017) (synthetic-to-real transfer), **Mini-DomainNet** (Litrico, Bue, and Morerio 2023) (domains: Clipart, Painting, Real, Sketch).

(2) Implementation details. We use PyTorch on 4 NVIDIA A40 GPUs. For fair comparison with mainstream methods (Zhou et al. 2022b; Khattak et al. 2023a; Bai et al. 2024; Du et al. 2024; Li et al. 2024a; Zhou and Zhou 2024), we adopt CLIP (Radford et al. 2021) with ViT/L-14 as the teacher's frozen image encoder. Students use ViT/B-16 or ResNet-50 (image encoder frozen), updating only lightweight visual/text prompts and an MLP layer. Training employs SGD with identical epochs/batch-size for teacher and student models. Prompt length M=4 is set empirically. Please check our open-sourced project https://anonymous.4open.science/r/EKDA for more implementation details.

(3) Baselines. To evaluate the effectiveness of EKDA, we compare it with representative state-of-the-art (SOTA) baselines: (i) ResNet-based methods including DANN (Ganin and Lempitsky 2015), JAN (Long et al. 2017), MDD (Zhang et al. 2019), SHOT (Liang, Hu, and Feng 2020), RCE (Ding et al. 2024), SDAT (Rangwani et al. 2022) and CMKD (Zhou and Zhou 2024); (ii) ViT-based methods including Deit (Touvron et al. 2021), CDTrans (Xu et al. 2022), SSRT (Sun et al. 2022) and TVT (Yang et al. 2023); (iii) recent prompt learning-based methods including CLIP (Radford et al. 2021), CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), VPT (Jia et al. 2022), IVLP (Khattak et al. 2023a), MaPLe (Khattak et al. 2023a), ADCLIP (Singha et al. 2023), DAPrompt (Ge et al. 2025), PDA (Bai et al. 2024), DAMP (Du et al. 2024), RADA-prompt (Jin et al. 2024), UniMoS (Li et al. 2024a) and ADAPT (Cui et al. 2025).

Performance Comparisons

(1) High performance on Office-31 and Office-Home. We compare EKDA with SOTA baselines across various architectures, including ResNet50/ViT/prompt learning-based approaches. On Office-31 (see Table 1), EKDA obtains obvious advantages with an average accuracy of 91.7% (ResNet50) and 93.2% (ViT/prompt learning), surpassing the best candidate (PDA) by 6.6% and 2.0%, respectively. In addition, on Office-Home (see Table 2), EKDA achieves an average accuracy of 83.9% with ResNet50, surpassing the best method CMKD (79.3%) by a significant margin of 4.6%. When evaluated with ViT, EKDA further improves the average accuracy to 89.8%, greatly outperforming the strongest baselines such as DAMP (87.1%) and PDA (85.7%). In particular, regardless of the image encoder (ResNet50/ViT-B/16) we adopted for the target domain, our solution leads the pack in the vast majority of cross-domain scenarios (highlighted in red), including all tasks on Office-31 (ResNet50), 4 out of 6 tasks on Office-31 (ViT/prompt learning), 10 out of 12 tasks on Office-Home (ResNet50), and 11 out of 12 tasks on Office-Home (ViT/prompt learning). This is mainly because the real-time cross-domain distillation strategy can significantly reduce the domain gap, thereby efficiently achieving semantic alignment between the source domain and the target domain. These results demonstrate the effectiveness of EKDA, which applies to diverse real-world scenarios.

Method	A-D	A-W	D-A	D-W	W-A	W-D	**Avg**
ResNet50							
CLIP⁎	74.1	67.0	72.8	67.0	72.8	74.1	**71.3**
CoOp	82.3	78.2	77.9	90.7	76.3	96.4	**83.6**
CoCoOp	82.9	76.7	75.6	88.8	76.7	93.6	**82.4**
VPT⁎	74.9	68.4	73.9	68.4	74.1	76.1	**72.6**
DAPrompt	77.3	71.9	76.7	74.7	77.4	79.7	**76.3**
PDA	85.1	81.1	76.6	92.8	77.3	97.8	**85.1**
EKDA	94.4	94.0	75.4	94.0	76.9	99.2	**91.7**
ViT/Prompt learning							
CLIP⁎	77.7	75.8	79.0	75.8	79.0	77.7	**77.5**
CoOp	88.5	88.5	82.0	96.1	82.4	99.0	**89.4**
CoCoOp	86.9	88.2	83.2	94.1	82.8	98.0	**88.9**
VPT⁎	89.6	86.5	81.9	96.5	82.8	99.2	**89.4**
DAPrompt	81.7	80.3	81.2	81.8	81.0	81.3	**81.2**
IVLP	85.7	89.2	81.9	98.4	80.3	99.2	**89.1**
MaPLe	86.9	88.6	83.0	97.7	82.0	99.4	**89.6**
PDA	91.2	92.1	83.5	98.1	82.5	99.8	**91.2**
EKDA	95.0	96.1	85.3	98.2	86.0	98.6	**93.2**

Table 1: Performance (%) comparisons on Office-31 with ResNet50/ViT/Prompt learning based methods. We mark (i) the best result in red and (ii) the average (Avg) result in bold. Note that ⁎ denotes the highest performance of a baseline that has multiple architecture variants.

(2) Efficiency-effectiveness trade-off on VisDA-2017 and Mini-DomainNet. To demonstrate the robustness and generalization of EKDA in more complex scenarios, we further conduct cross-domain validation on complicated VisDA-2017 and Mini-DomainNet compared with representative ViT/prompt learning-based methods. On VisDA-2017 (see Table 3), EKDA attains an average accuracy of 90.1%, surpassing the strongest baselines such as DAPrompt (89.5%) and PDA (89.7%). Notably, EKDA achieves the best performance in 5 out of 12 categories, including challenging classes like Horse and Truck, demonstrating its robustness to significant domain shifts. Similarly, on Mini-DomainNet (see Table 4), EKDA outperforms existing methods with an average accuracy of 88.4%, exceeding the performance of ADCLIP (86.9%) and DAMP (87.6%). It achieves the highest accuracy in 6 out of 12 domain adaptation tasks, particularly excelling in challenging cross-domain scenarios such as C-S (Clipart→Sketch) and R-S (Real→Sketch). In addition, we also record the efficiency-effectiveness trade-off between the strongest baselines in Table 5. As we see, while the SOTA methods like PDA rely on more complex training strategies or additional learnable cross-attention modules, EKDA achieves superior performance with minimal trainable parameters (2.02MB), accounting for only 19.03% of the optimal candidate PDA (10.61MB). Furthermore, EKDA

Method	A-C	A-P	A-R	C-A	C-P	C-R	P-A	P-C	P-R	R-A	R-C	R-P	**Avg**
ResNet50													
DANN	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**
JAN	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	**57.6**
MDD	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	**68.1**
SHOT	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	**71.8**
SDAT	56.0	72.2	78.6	62.5	73.2	71.8	62.1	55.9	80.3	75.0	61.4	84.5	**69.5**
DAPrompt	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**
RADA-prompt	59.0	78.4	82.3	67.7	75.9	75.0	69.2	57.2	84.2	77.8	62.7	86.5	**72.9**
RCE	57.1	81.0	83.3	70.7	77.8	80.3	71.9	59.2	82.4	78.9	62.4	86.3	**74.3**
PDA	55.4	85.1	85.8	75.2	85.2	85.2	74.2	55.2	85.8	74.7	55.8	86.3	**75.3**
UniMoS	59.5	89.4	86.9	75.2	89.6	86.8	75.4	58.4	87.2	76.9	59.5	89.7	**77.9**
DAMP	59.7	88.5	86.8	76.6	88.9	87.0	76.3	59.6	87.1	77.0	61.0	89.9	**78.2**
CMKD	65.9	86.6	87.3	74.4	87.7	85.8	75.9	64.4	87.9	79.1	67.2	90.0	**79.3**
EKDA	61.1	90.2	89.5	88.2	91.8	90.8	88.3	66.3	91.4	89.4	67.1	92.3	**83.9**
ViT/Prompt learning													
CLIP*	67.6	89.0	89.4	82.4	89.0	89.4	82.4	67.6	89.4	82.4	67.6	89.0	**82.1**
Deit*	61.8	79.5	84.3	75.4	78.8	81.2	72.8	55.7	84.4	78.3	59.3	86.0	**74.8**
CDTrans*	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**
SDAT	69.1	86.6	88.9	81.9	86.2	88.0	81.0	66.7	89.7	86.2	72.1	91.9	**82.4**
SSRT	75.2	89.0	91.1	85.1	88.3	89.9	85.0	74.2	91.2	85.7	78.6	91.8	**85.4**
CoOp	70.0	90.8	90.9	83.2	90.9	89.2	82.0	71.8	90.5	83.8	71.5	92.0	**83.9**
CoCoOp	70.4	91.4	90.4	83.5	91.8	90.3	83.4	70.9	91.0	83.4	71.2	91.7	**84.1**
VPT*	71.6	89.9	90.3	82.8	91.0	89.7	82.0	71.5	90.3	84.6	71.7	91.6	**83.9**
DAPrompt	70.7	91.0	90.9	85.2	91.0	91.0	85.1	70.7	90.9	85.3	70.4	91.4	**84.4**
TVT	74.9	86.8	89.5	82.8	87.9	88.3	79.8	71.9	90.1	85.5	74.6	90.6	**83.6**
IVLP	71.4	91.7	90.8	83.6	90.2	89.3	82.2	72.4	90.4	84.1	72.1	92.0	**84.2**
MaPLe	72.2	91.6	90.3	82.6	90.9	89.8	82.4	71.6	90.1	85.1	72.0	92.1	**84.2**
PDA	72.9	90.9	91.4	86.3	91.3	91.6	86.2	73.8	91.6	86.4	73.3	92.4	**85.7**
DAMP	75.7	94.2	92.0	86.3	94.2	91.9	86.2	76.3	92.4	86.1	75.6	94.0	**87.1**
EKDA	78.6	94.0	94.1	89.9	95.1	94.4	90.2	80.4	94.4	90.9	80.1	95.0	**89.8**

Table 2: Performance (%) comparisons on Office-Home with ResNet50/ViT/Prompt learning based methods.

Method	Plane	Bicycle	Bus	Car	Horse	Knife	Mcycl	Person	Plant	Sktbrd	Train	Truck	**Avg**
Deit*	98.2	73.0	82.5	62.0	97.3	63.5	96.5	29.8	68.7	86.7	96.7	23.6	**73.2**
CDTrans*	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**
SDAT	96.3	80.7	74.5	65.4	95.8	99.5	92.0	83.7	93.6	88.9	85.8	57.2	**84.5**
SSRT	98.9	87.6	89.1	84.8	98.3	98.7	96.3	81.1	94.8	97.9	94.5	43.1	**88.8**
TVT	92.9	85.6	77.5	60.5	93.6	98.2	89.3	76.4	93.6	92.0	91.7	55.7	**83.9**
CoOp	98.7	89.8	94.2	69.7	99.0	71.5	96.3	53.9	91.5	96.3	95.8	35.7	**82.7**
CoCoOp	99.1	92.4	92.0	71.7	99.1	95.0	95.8	22.7	90.3	95.6	96.0	60.6	**84.2**
VPT*	98.7	78.2	96.0	68.7	98.8	83.6	97.0	82.5	87.4	94.5	94.3	54.6	**86.2**
DAPrompt	99.1	92.6	93.1	77.4	98.4	92.2	94.6	84.7	88.3	96.1	93.7	63.4	**89.5**
MaPLe	98.6	85.8	93.0	68.8	99.2	72.4	96.8	77.1	84.7	96.0	95.9	33.1	**83.5**
PDA	99.2	91.1	91.9	77.1	98.4	93.6	95.1	84.9	87.2	97.3	95.3	65.3	**89.7**
EKDA	99.2	90.8	90.1	70.1	99.2	91.9	97.2	85.7	93.8	96.1	97.0	70.0	**90.1**

Table 3: Performance (%) comparisons on VisDA-2017 with ViT/Prompt learning (ViT-B/16) based methods.

exhibits faster convergence speed, requiring only 10 training epochs in all cases, whereas other candidates require many more training epochs (such as 50 epochs of UniMos) to achieve domain alignment. Despite the increased complexity of these larger benchmarks, EKDA achieves highly competitive results while maintaining remarkable efficiency. This excellent cross-domain alignment capability of EKDA does not stem from the stacking of parameters or modules, but rather from our ingenious design of distilling knowledge from the source domain to the target domain in real time. These results demonstrate that EKDA strikes an optimal efficiency-effectiveness trade-off, possessing strong generalization capabilities that can be extended to more complex domain adaptation scenarios.

(3) Model complexity analysis. We further conduct model complexity analysis of EKDA compared with competitive methods. Take Office-Home for example (see Figure 3), we compare the learnable parameters and average accuracy under different image encoders. (i) With the default ViT-B/16 image encoder, EKDA (marked as red square) achieves an accuracy of 89.8% with around 2 MB learnable parameters, showing the highest performance and almost the lowest model complexity among others. Note that although CLIP, CoOp, and CoCoOp have lower parameter counts, they can hardly complete effective UDA and yield relatively poor results under the same conditions. (ii) With ResNet50 (RN50) image encoder, EKDA (marked as orange square) produces an accuracy of 83.9% with around 2 MB learnable parameters. Other candidates such as PDA (RN50), CMKD (RN50), and UniMos (RN50) introduce more training parameters, but result in lower performance. In summary, our EKDA demonstrates good performance and parameter efficiency with strong generalization under different image encoders.

Method	C-P	C-R	C-S	P-C	P-R	P-S	R-C	R-P	R-S	S-C	S-P	S-R	**Avg**
ViT*	63.3	79.0	56.4	62.6	83.3	55.4	62.0	70.3	53.5	63.0	63.6	75.8	**65.7**
CLIP*	80.3	90.5	77.8	82.7	90.5	77.8	82.7	80.3	77.8	82.7	80.3	90.5	**82.8**
DAPrompt	83.3	92.4	81.1	86.4	92.1	81.0	86.7	83.3	80.8	86.8	83.5	91.9	**85.8**
ADAPT	83.5	92.4	81.9	88.1	93.0	82.8	88.1	83.7	82.2	88.7	83.7	92.9	**86.8**
ADCLIP	84.3	93.7	82.4	87.5	93.5	82.4	87.3	84.5	81.6	87.9	84.8	93.0	**86.9**
UniMoS	86.2	93.2	83.2	86.9	93.2	83.2	86.8	86.0	82.8	87.0	86.2	93.3	**87.3**
DAMP	86.4	93.3	83.5	87.2	93.4	84.1	87.2	86.5	82.5	87.3	86.6	93.4	**87.6**
EKDA	85.1	93.2	85.2	90.3	92.9	84.8	90.3	84.9	87.1	90.5	84.1	92.7	**88.4**

Table 4: Performance (%) comparisons on Mini-DomainNet with ViT/Prompt learning (ViT-B/16) based methods.

Method	VisDA-2017			Mini-DomainNet		
	Epoch	Param	Avg (%)	Epoch	Param	Avg (%)
PDA	10	10.61	89.7	-	-	-
UniMoS	10	11.08	88.1	50	11.08	87.3
EKDA	10	2.02	90.1	10	2.02	88.4

Table 5: Efficiency-effectiveness trade-off. Note that Param denotes the scale of learnable parameters (MB).

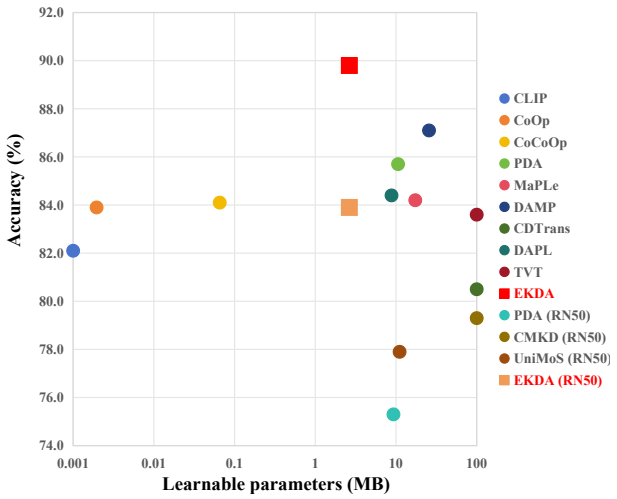

Figure 3: Comparisons of learnable parameters and accuracy.

Ablation Studies

(1) Impact of key knowledge distillation operations. We implement 3 key knowledge distillation operations (see Equation (7)) to transfer the knowledge from the source domain to the target domain. In Table 6, we record the average accuracy on the target domain with/without key knowledge distillation operations. Firstly, if we refrain from implementing any distillation operations, we will inevitably encounter suboptimal outcomes across all scenarios, owing to the dearth of interaction between the source domain and the target domain, which gives rise to a significant domain gap. Secondly, the introduction of the KL divergence loss (L_{kl}) plays a pivotal role in knowledge transfer to bring remarkable performance improvement (e.g., +7.3% on Office-Home and +11.5% on Office-31), indicating its criticality in aligning domain distributions and preserving semantic consistency. Thirdly, the cross-entropy loss (L_{ce}^{stu}) further refines the target-domain predictions by encouraging confident outputs, contributing to marginal yet consistent improvements (e.g., +0.4% on Office-Home). Finally, the alignment loss (L_{align}) enhances cross-domain feature adaptation, particularly benefiting VisDA-2017 (+0.4% over the two-component setup). While L_{ce}^{stu} and L_{align} provide auxiliary benefits, their effects are secondary compared to L_{kl}, emphasizing that KL divergence serves as the cornerstone for effective knowledge transfer. These results collectively validate the necessity of a hierarchical distillation strategy: L_{kl} ensures robust foundational alignment, while L_{ce}^{stu} and L_{align} synergistically refine domain invariance and prediction reliability.

L_{kl}	L_{ce}^{stu}	L_{align}	Office-Home	Office-31	VisDA-2017
			82.1	77.5	88.9
✓			89.4	93.0	89.6
✓	✓		89.8	93.1	89.7
✓	✓	✓	89.8	93.2	90.1

Table 6: Performance (%) change with/without key knowledge distillation operations on the target domain.

Visual prompt	Deep visual prompt	Office-Home	Office-31	VisDA-2017
✓		89.6	93.0	89.5
✓	✓	89.8	93.2	90.1

Table 7: Performance (%) change with/without deep visual prompt for the student model.

(2) Impact of deep visual prompt. As shown in Figure 2, during the training of the student model, in addition to employing traditional multi-modal prompting mechanisms that incorporate shallow visual prompt at the input level, we have also incorporated deep visual prompt within each layer of the student's image encoder. In Table 7, we record the average accuracy on the target domain with/without deep visual prompt across diverse cross-domain scenarios. As we see, compared with shallow visual prompt, deep visual prompt yields higher performance on all datasets, with +0.2% on Office-Home, +0.2% on Office-31, and +0.6% on VisDA-2017. Shallow visual prompt primarily enables the student model's output to align with the teacher model's output (e.g., text features) by freezing the student model's image encoder. In contrast, deep visual prompt further ensures that the output distribution (such as image features and probability distributions) of the student model dynamically matches that of

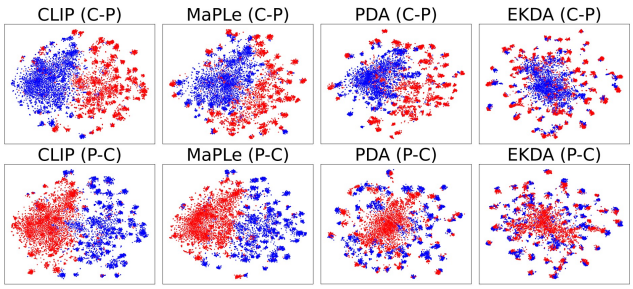

Figure 4: T-SNE visualization results of image features. The blue denotes the source domain while the red denotes the target domain.

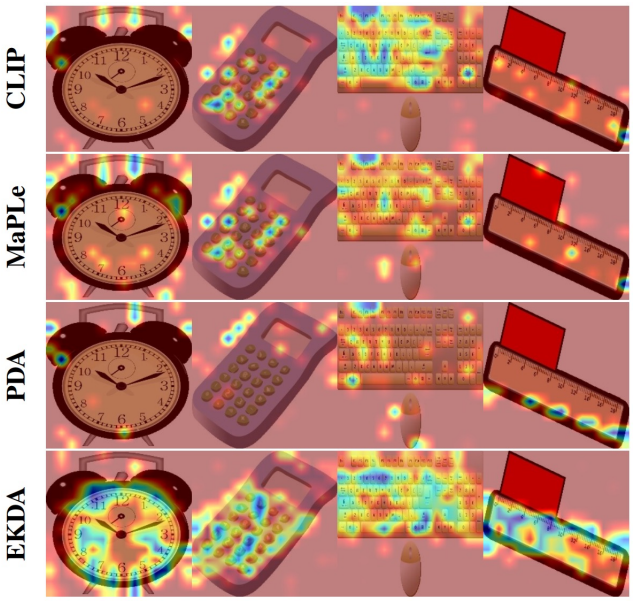

Figure 5: Grad-CAM visualization results.

the teacher model, thereby facilitating more effective knowledge distillation, particularly on complex datasets.

Visualization

T-SNE and Grad-CAM visualization. We employ two complementary techniques to demonstrate our method's efficacy. Firstly, T-SNE (Van der Maaten and Hinton 2008) visualization of image features (see Figure 4) reveals that CLIP and MaPLe exhibit significant source-target domain gaps due to absent cross-domain interactions, while PDA shows improved alignment but poor category separation with clustered data points. In contrast, our EKDA achieves both clear category distinction and enhanced domain alignment. Secondly, Grad-CAM (Selvaraju et al. 2020) analysis (see Figure 5) indicates current baseline methods partially identify core image regions but activate biased semantics, while our EKDA consistently highlights semantically critical areas, enabling precise feature extraction for superior domain adaptation.

Conclusion and Future Works

In this paper, we propose EKDA to implement end-to-end knowledge distillation for UDA that enables real-time cross-domain interactions between the source domain and target domain, greatly reducing the domain gap. Extensive experiments on Office-31, Office-Home, VisDA-2017, and Mini-DomainNet demonstrate that EKDA achieves significant performance improvement over the SOTA UDA approaches, while maintaining a much lower model complexity. In addition, the ablation studies and visualization results further demonstrate the effectiveness of our proposed training strategy and key operations. In the future, we will continue to explore (i) more effective knowledge distillation techniques to measure the distribution discrepancies across domains, and (ii) more robust prompt learning methods to integrate the prompt information from the source domain and the target domain.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62406082, No. 62506085, No. 62502156, No. 62394334), Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515110650, No. 2023A1515110659), Guangzhou Science and Technology Planning Project (No. 2024A03J0013, No. 2025A04J4590), Guangdong Provincial Department of Education Innovation Strong School Youth Innovation Talent Project (No. 2023KQNCX055), National Key Research and Development Program of China (No. 2023YFC3306204), the Project of Xiangjiang Laboratory under Grant 24XJJCYJ01002, and Science and Technology Program of Guangzhou (No. SL2023A03J00284).

References

Bai, S.; Zhang, M.; Zhou, W.; Huang, S.; Luan, Z.; Wang, D.; and Chen, B. 2024. Prompt-Based Distribution Alignment for Unsupervised Domain Adaptation. In *AAAI*, 729–737.

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

Cai, Y.; Shang, Y.; and Yin, J. 2024. MultiDAN: Unsupervised, Multistage, Multisource and Multitarget Domain Adaptation for Semantic Segmentation of Remote Sensing Images. In *MM*, 1168–1177.

Cho, J. H.; and Hariharan, B. 2019. On the Efficacy of Knowledge Distillation. In *ICCV*, 4793–4801.

Cui, C.; Liu, Z.; Gong, S.; Zhu, L.; Zhang, C.; and Liu, H. 2025. When Adversarial Training Meets Prompt Tuning: Adversarial Dual Prompt Tuning for Unsupervised Domain Adaptation. *IEEE Trans. Image Process.*, 34: 1427–1440.

Ding, F.; Li, J.; Tian, W.; Zhang, S.; and Yuan, W. 2024. Unsupervised Domain Adaptation via Risk-Consistent Estimators. *IEEE Trans. Multim.*, 26: 1179–1187.

Du, Z.; Li, X.; Li, F.; Lu, K.; Zhu, L.; and Li, J. 2024. Domain-Agnostic Mutual Prompting for Unsupervised Domain Adaptation. In *CVPR*, 23375–23384.

Ganin, Y.; and Lempitsky, V. S. 2015. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, volume 37, 1180–1189.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2025. Domain Adaptation via Prompt Learning. *IEEE Trans. Neural Networks Learn. Syst.*, 36(1): 1160–1170.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, volume 139, 4904–4916.

Jia, M.; Tang, L.; Chen, B.; Cardie, C.; Belongie, S. J.; Hariharan, B.; and Lim, S. 2022. Visual Prompt Tuning. In *ECCV*, volume 13693, 709–727.

Jin, X.; Lan, C.; Zeng, W.; and Chen, Z. 2024. Domain Prompt Tuning via Meta Relabeling for Unsupervised Adversarial Adaptation. *IEEE Trans. Multim.*, 26: 8333–8347.

Jin, Y.; Wang, J.; and Lin, D. 2023. Multi-Level Logit Distillation. In *CVPR*, 24276–24285.

Khattak, M. U.; Rasheed, H. A.; Maaz, M.; Khan, S. H.; and Khan, F. S. 2023a. MaPLe: Multi-modal Prompt Learning. In *CVPR*, 19113–19122.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.; and Khan, F. S. 2023b. Self-regulating Prompts: Foundational Model Adaptation without Forgetting. In *ICCV*, 15144–15154.

Li, X.; Li, Y.; Du, Z.; Li, F.; Lu, K.; and Li, J. 2024a. Split to Merge: Unifying Separated Modalities for Unsupervised Domain Adaptation. In *CVPR*, 23364–23374.

Li, Z.; Li, X.; Fu, X.; Zhang, X.; Wang, W.; Chen, S.; and Yang, J. 2024b. PromptKD: Unsupervised Prompt Distillation for Vision-Language Models. In *CVPR*, 26607–26616.

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*, volume 119, 6028–6039.

Litrico, M.; Bue, A. D.; and Morerio, P. 2023. Guiding Pseudo-labels with Uncertainty Estimation for Source-free Unsupervised Domain Adaptation. In *CVPR*, 7640–7650.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning Transferable Features with Deep Adaptation Networks. In *ICML*, volume 37, 97–105.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *ICML*, volume 70, 2208–2217.

Mirzadeh, S.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved Knowledge Distillation via Teacher Assistant. In *AAAI*, 5191–5198.

Mistretta, M.; Baldrati, A.; Bertini, M.; and Bagdanov, A. D. 2024. Improving Zero-Shot Generalization of Learned Prompts via Unsupervised Knowledge Distillation. In *ECCV*, volume 15142, 459–477.

Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. VisDA: The Visual Domain Adaptation Challenge. *CoRR*, abs/1710.06924.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Rangwani, H.; Aithal, S. K.; Mishra, M.; Jain, A.; and Radhakrishnan, V. B. 2022. A Closer Look at Smoothness in Domain Adversarial Training. In *ICML*, volume 162, 18378–18399.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting Visual Category Models to New Domains. In *ECCV*, volume 6314, 213–226.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.*, 128(2): 336–359.

Shin, H.; Kim, H.; Kim, S.; Jun, Y.; Eo, T.; and Hwang, D. 2023. SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation. In *CVPR*, 7412–7421.

Singha, M.; Pal, H.; Jha, A.; and Banerjee, B. 2023. AD-CLIP: Adapting Domains in Prompt Space Using CLIP. In *ICCV*, 4357–4366.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of Frustratingly Easy Domain Adaptation. In *AAAI*, 2058–2065.

Sun, T.; Lu, C.; Zhang, T.; and Ling, H. 2022. Safe Self-Refinement for Transformer-based Domain Adaptation. In *CVPR*, 7181–7190.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, 10347–10357.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial Discriminative Domain Adaptation. In *CVPR*, 2962–2971.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep Hashing Network for Unsupervised Domain Adaptation. In *CVPR*, 5385–5394.

Wilson, G.; and Cook, D. J. 2020. A Survey of Unsupervised Deep Domain Adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5): 51:1–51:46.

Xu, T.; Chen, W.; Wang, P.; Wang, F.; Li, H.; and Jin, R. 2022. CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation. In *ICLR*.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *CVPR*, 15952–15962.

Yang, J.; Liu, J.; Xu, N.; and Huang, J. 2023. TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation. In *WACV*, 520–530.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *ICLR*.

Zhang, X.; Wu, Y.; Angelini, E. D.; Li, A.; Guo, J.; Rasmussen, J. M.; O'Connor, T. G.; Wadhwa, P. D.; Jackowski, A. P.; Li, H.; Posner, J.; Laine, A. F.; and Wang, Y. 2024. MAPSeg: Unified Unsupervised Domain Adaptation for Heterogeneous Medical Image Segmentation Based on 3D Masked Autoencoding and Pseudo-Labeling. In *CVPR*, 5851–5862.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. I. 2019. Bridging Theory and Algorithm for Domain Adaptation. In *ICML*, volume 97, 7404–7413.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *CVPR*, 16795–16804.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *Int. J. Comput. Vis.*, 130(9): 2337–2348.

Zhou, W.; and Zhou, Z. 2024. Unsupervised Domain Adaption Harnessing Vision-Language Pre-Training. *IEEE Trans. Circuits Syst. Video Technol.*, 34(9): 8201–8214.

Zhu, D.; Li, Y.; Shao, Y.; Hao, J.; Wu, F.; Kuang, K.; Xiao, J.; and Wu, C. 2023. Generalized Universal Domain Adaptation with Generative Flow Networks. In *MM*, 8304–8315.