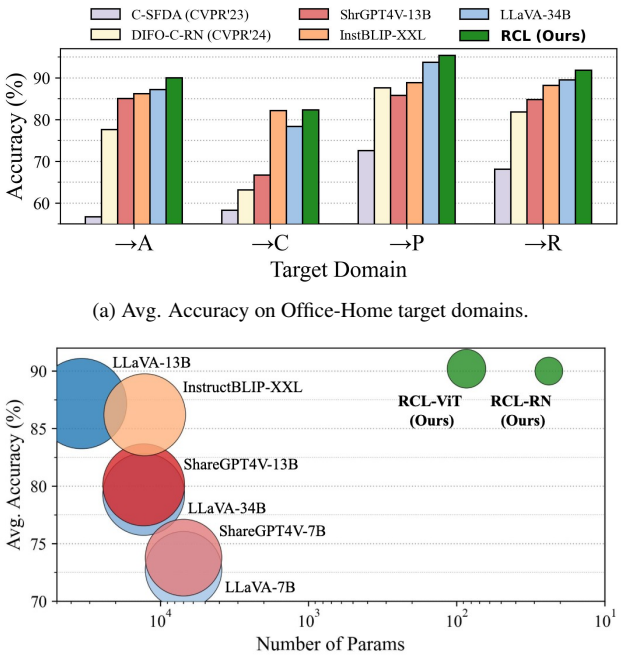
Empowering Source-Free Domain Adaptation via MLLM-Guided Reliability-Based Curriculum Learning

Dongjie Chen[1*], Kartik Patwari[1*], Zhengfeng Lai[1], Xiaoguang Zhu[1], Sen-ching Cheung[1,2], Chen-Nee Chuah[1]

[1] University of California, Davis [2] University of Kentucky

Abstract

Source-Free Domain Adaptation (SFDA) aims to adapt a pre-trained source model to a target domain using only unlabeled target data. Current SFDA methods face challenges in effectively leveraging pre-trained knowledge and exploiting target domain data. Multimodal Large Language Models (MLLMs) offer remarkable capabilities in understanding visual and textual information, but their applicability to SFDA poses challenges such as instruction-following failures, intensive computational demands, and difficulties in performance measurement prior to adaptation. To alleviate these issues, we propose ***Reliability-based Curriculum Learning (RCL)****, a novel framework that integrates multiple MLLMs for knowledge exploitation via pseudo-labeling in SFDA. Our framework incorporates Reliable Knowledge Transfer, Self-correcting and MLLM-guided Knowledge Expansion, and Multi-hot Masking Refinement to progressively exploit unlabeled data in the target domain. RCL achieves state-of-the-art (SOTA) performance on multiple SFDA benchmarks, e.g.,* ***+9.4%*** *on DomainNet, demonstrating its effectiveness in enhancing adaptability and robustness without requiring access to source data. Our code is available at* https://github.com/Dong-Jie-Chen/RCL



(a) Avg. Accuracy on Office-Home target domains.

(b) Accuracy vs. Model Size (Number of Parameters, log scale).

Figure 1. Comparisons with existing methods, MLLMs (zero-shot with proposed STS), and RCL on OfficeHome dataset. RCL achieves SOTA results across domains while being lightweight.

1. Introduction

Source-Free Domain Adaptation (SFDA) aims to adapt a pre-trained model to a new target domain without requiring access to labeled source data. This setting is particularly useful when privacy, storage, or proprietary constraints prevent access to source datasets. One of the key aspects that make SFDA effective is the use of externally pre-trained models. Earlier works utilized models pre-trained on ImageNet [20], while more recent works shift towards using large-scale pre-trained models like CLIP [7, 15, 33, 41]. These models capture rich, transferable representations that bridge the domain gap. However, the importance of pre-trained knowledge has been less explored in previous works [20].

Recently, multimodal large language models (MLLMs) such as LLaVA [24] and InstructBLIP [4] have set new performance records in tasks like visual question answering (VQA), detailed description, and complex reasoning retrieval [21, 25]. With their strong generalization ability, MLLMs have been widely adopted for efficient fine-tuning in downstream tasks, particularly in scenarios with limited training data. However, a critical observation is that the datasets used to train MLLMs are typically inaccessible

* Equal contribution.
Correspondence to: Dongjie Chen <cdjchen@ucdavis.edu>.

due to proprietary restrictions or their scale and complexity. Consequently, utilizing MLLMs for downstream tasks inherently aligns with the principles of SFDA, as they transfer knowledge without access to the source training data. This necessitates the exploration of SFDA with MLLMs, especially in specialized domains such as medical and industrial applications, where source data is often sensitive or restricted, and MLLMs could have transformative impacts.

Nevertheless, the application of MLLMs in SFDA is still limited. One of the primary challenges in employing MLLMs for SFDA is that MLLMs are not suitable for zero-shot classification tasks. As illustrated in Figure 3, MLLMs often fail to follow classification instructions as they are primarily designed for text generation. Remarkably, with improvements on instruction following (with STS as described in Sec. 3.1), these MLLMs achieve superior results without having been exposed to any images from either the source or target domain. As shown in Figure 1(a), InstructBLIP and LLaVA demonstrate better performance compared to models pre-trained on ImageNet (C-SFDA [13]) or distilling knowledge from CLIP (DIFO-C-ViT [41]). Based on this observation, we hypothesize that pre-trained knowledge can play an equally or more important role than pre-trained models in improving the performances of SFDA.

While MLLMs can inject valuable pre-trained knowledge, their applications to SFDA still face two major challenges. First, their inference process is time-consuming and computationally intensive, which prevents their wide adoption. Second, MLLMs exhibit variability in zero-shot SFDA performance, (see Figure 1(b)). Their effectiveness depends not only on model size but also on differences in their pretraining datasets and architectures, leading to diverse knowledge representations. Consequently, relying on a single MLLM for SFDA can result in suboptimal adaptation, particularly when the target domain shifted with MLLM's training data in a large margin. Additionally, fine-tuning MLLMs for specific domains requires substantial computational resources and a GPT-4-based instruction-following dataset [4, 25], creating scalability challenges across diverse downstream tasks. This underscores the need for efficient methods that utilize multiple MLLMs for knowledge transfer while preserving scalability for domain-specific adaptations.

To overcome these issues, we first propose Semantic Textual Similarity (STS), a method that reformulates image classification as a VQA task by prompting MLLMs with predefined class names. We observe that MLLMs sometimes fail to strictly follow instructions (see Fig 3), producing open-ended responses influenced by their pretraining. STS aligns these outputs with target class labels, ensuring MLLMs function reliably for classification. As shown in Figure 1(a), zero-shot MLLMs with our STS can rival SOTA SFDA methods without any fine-tuning, making them an attractive choice for adaptation. Beyond using a single MLLM, a fundamental limitation remains: knowledge transfer from one MLLM is inherently constrained—pseudo-labels derived from a single model do not fully capture the diverse reasoning capabilities of multiple MLLMs. A naive approach would be to directly distill knowledge from an ensemble of MLLMs, but this leads to inconsistent pseudo-labels and over-reliance on a single teacher's biases. Instead, we introduce **Reliability-based Curriculum Learning (RCL)**, a novel multi-MLLM distillation framework that incorporates reliability-driven learning. RCL integrates multiple MLLMs for SFDA through agreement-driven knowledge distillation. Unlike traditional multi-teacher knowledge distillation (MTKD), which assumes that teacher models provide uniformly reliable guidance, RCL introduces the concept of pseudo-label reliability based on MLLM agreement. This is particularly critical because MLLMs perform zero-shot classification, making their outputs inherently uncertain.

To this end, we: (1) Quantify pseudo-label reliability by measuring agreement among multiple MLLMs, ensuring that knowledge transfer is guided by robust, consensus-driven supervision. (2) Design a structured curriculum learning strategy that progressively incorporates pseudo-labels based on their reliability: the model first learns from high-confidence samples, then integrates less reliable pseudo-labels with adaptive correction, and finally incorporates uncertain or previously unlabeled samples through Multi-hot Masking Refinement (MMR). This staged approach prevents early overfitting to unreliable labels and enables a self-correcting adaptation process. (3) Introduce MMR within the curriculum learning framework to address uncertainty by refining noisy pseudo-labels rather than discarding them, ensuring that all available target data contributes to model adaptation. This refinement process leverages multi-hot masking and consistency regularization to improve pseudo-label accuracy, allowing the model to effectively utilize ambiguous samples while mitigating errors.

Leveraging MLLM knowledge in a structured adaptation framework, RCL achieves SOTA SFDA performance on Office-Home, DomainNet, and VisDA, surpassing single-MLLM zero-shot and recent SFDA methods. Our results highlight that a multi-MLLM distillation approach, when guided by reliability and curriculum learning, enhances adaptation robustness without requiring fine-tuning or computationally expensive retraining. While our primary focus is SFDA, the principles of reliability-driven multi-teacher distillation and curriculum learning could extend beyond SFDA to broader applications in knowledge transfer, model adaptation, and robust AI training.

2. Related Work

Source-Free Domain Adaptation. SFDA adapts a pre-trained source model to a target domain using unlabeled target data, making pseudo-labeling a key technique [3, 22, 29].

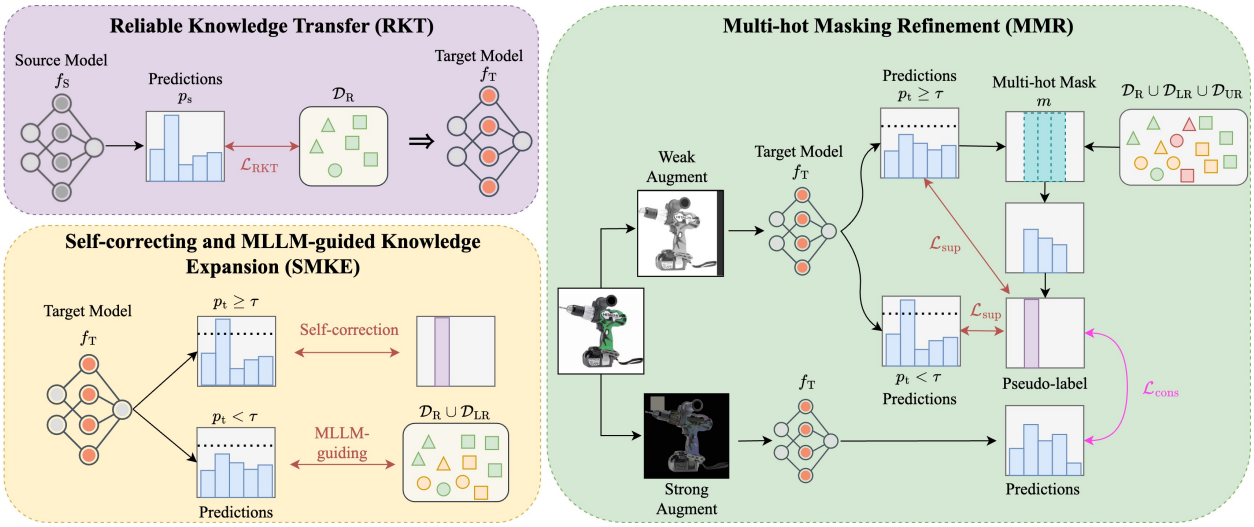


Figure 2. An overview of our proposed Reliability-based Curriculum Learning (RCL) framework.

Some works improve pseudo-labeling by leveraging target data structure [10, 35, 37] or aligning source and target distributions [5, 46]. Others explore synthetic target-style data generation [19, 42, 49]. Our approach leverages MLLMs' zero-shot capabilities for pseudo-labeling and performs multi-teacher distillation through a curriculum learning framework to enhance adaptation.

VLMs/MLLMs. Pretrained vision-language models (VLMs) like CLIP [30] and ALIGN [9] capture vision-language-aligned features, while recent MLLMs (e.g., LLaVA [24], InstructBLIP [4]) enhance multimodal understanding via LLM backbones, enabling strong zero-shot capabilities. While some works use CLIP for domain adaptation, they require labeled source data [16, 43, 50] or fine-tuning [15]. DIFO [41], the latest SOTA work, adapts CLIP for target-domain learning by iteratively customizing it with prompt learning and distilling its knowledge into a task-specific model. While effective, this approach requires adapting a large VLM, whereas our method leverages multiple MLLMs in a zero-shot manner without finetuning.

Multi-Teacher Knowledge Distillation (MTKD). MTKD extends traditional KD by distilling knowledge from multiple teachers [11, 28], typically assuming labeled source data and fixed teacher ensembles [18, 44]. Wu et al. [44] proposed aligning teacher outputs to ensure consistent supervision, while Li et al. [18] introduced knowledge integration to enhance student generalization beyond simple imitation. Kang et al. [12] explored heterogeneous teacher architectures in recommendation tasks, demonstrating the benefits of aggregating diverse model types for improved performance. Ding et al. [6] studied balancing teacher ensemble size by learning a categorical distribution for stochastic teacher selection, optimizing the trade-off between capacity and efficiency. Unlike prior approaches that assume labeled source data and static teacher ensembles, our method leverages MLLMs for SFDA without source access. RCL uniquely refines pseudo-labels via curriculum learning, incorporating reliability scoring to adapt MLLM knowledge from multiple teachers—to the best of our knowledge, this has not been explored before. The proposed MMR (Sec. 4.3) enhances robustness by mitigating pseudo-label noise across domains. Without stochastic teacher selection, RCL ensures structured, self-correcting adaptation, achieving strong SFDA performance even under varying pseudo-label confidence.

3. Pseudo-labeling and Reliability Measurement with MLLMs

First, we formally define SFDA for image classification. We denote $\mathcal{D}_s = (x_s^i, y_s^i)_{i=1}^{N_s}$ as the labeled source-domain dataset with N_s images, where $x_s^i \in \mathcal{X}_s$ refers to an image and $y_s^i \in \mathcal{Y}_s$ is its corresponding one-hot label. A pre-trained source model $f_{\theta_s} : \mathcal{X}_s \to \mathcal{Y}_s$ is trained on $\mathcal{D}_s$, where θ_s represents its learned parameters. The target domain contains an unlabeled dataset, $\mathcal{D}_t = \{x_t^i\}_{i=1}^{N_t}$, where $x_i^t \in \mathcal{X}_t$ (target domain images), and N_t is the number of unlabeled images. The goal of SFDA is to adapt a pre-trained source model f_{θ_s} to $\mathcal{D}_t$ without access to $\mathcal{D}_s$ during the adaptation process. The goal is to train a target model $f_{\theta_t} : \mathcal{X}_t \to \mathcal{Y}_t$, where $\mathcal{Y}_t$ is the target domain label space.

3.1. Pseudo-labeling with MLLMs

We leverage multiple MLLMs for initial target-image labeling. As MLLMs are primarily designed for text generation, their responses may deviate from classification requirements. Therefore, we design prompts to repurpose MLLMs for class label prediction. Figure 3 shows how we reframe VQA as

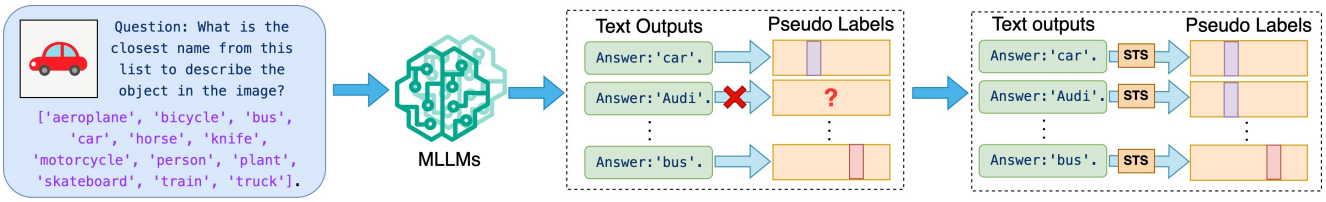


Figure 3. Pseudo-labeling with MLLMs. Directly prompting MLLMs for classification can lead to failures: we propose Semantic Textual Similarity (STS) to correct pseudo-labelling.

a zero-shot classification task using MLLMs. We design the prompt to incorporate all class names and a question instructing MLLMs to select the most appropriate match. This prompt and the image x_t^i are then fed to multiple pre-trained MLLMs, such as LLaVA [24, 25], InstructBLIP [4], and ShareGPT4V [2]. Each MLLM generates a text output $T_1^i, T_2^i, \ldots, T_M^i$, where M is the number of MLLMs.

However, as shown in Figure 3, MLLMs sometimes fail to follow prompts, generating responses beyond classification constraints. This happens when MLLMs rely on prior knowledge rather than selecting from provided options (e.g., predicting 'Audi' instead of 'car'). Unlike MTKD, which distills structured soft labels, our approach must handle inherently diverse and inconsistent MLLM outputs. To mitigate this, we propose Semantic Textual Similarity (STS) to align outputs with class names, ensuring pseudo-label consistency.

We derive pseudo-labels by computing STS between class names and MLLM-generated text. Formally, for the m-th MLLM and the i-th image x_t^i, the pseudo-label $\hat{y}^{mi}$ is determined by:

$$\hat{y}^{mi} = \underset{c}{\operatorname{argmax}}\, \mathrm{STS}(T_m^i, T_t^c), \tag{1}$$

where T_t^c represents the name of the c-th class. The STS between two text sequences T_1 and T_2 is computed as:

$$\mathrm{STS}(T_1, T_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2} - 1, \tag{2}$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are vector representations (using [31]) of T_1 and T_2, respectively. STS refines pseudo-labels by aligning MLLM outputs with class semantics, even when instructions are ignored. To facilitate pseudo-labeling, we design specific prompt templates for different MLLMs: (1) LLaVA models: *"Question: What is the closest name from this list to describe the object in the image? Return the name only. <class names>"* (2) ShareGPT4V models: *"Question: What is the closest name from this list to describe the object in the image? List: <class names> Return the closest name from the list only. Use *exact* names from the list only. Answer:"* (3) InstructBLIP models: *"Question: What is the closest name from this list to describe the object in the image? <class names>. Use the closest name from the list only. Pick the answer from the list only. Answer:"*

3.2. Consensus-based Reliability Measurement

Pseudo-labels from different MLLMs may vary for the same target sample. This disagreement raises a key question: **how can we measure the reliability of the pseudo-labels from multiple MLLMs?** While STS helps correct deviations from instructions, it does not assess pseudo-label reliability, as it cannot detect when MLLMs generate incorrect labels. To address this, we propose a consensus-based reliability metric for pseudo-labels.

We define a reliability score $\mathcal{R}(x_t^i)$ for each target domain sample x_t^i based on agreement among MLLM-assigned pseudo-labels:

$$\mathcal{R}(x_t^i) = \frac{1}{M(M-1)} \sum_{m=1}^{M} \sum_{n=1, n\neq m}^{M} \mathbb{1}(\hat{y}^{mi} = \hat{y}^{ni}), \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function. $\mathcal{R}(x_t^i)$ quantifies the proportion of MLLM pairs that assign the same pseudo-label to x_t^i. Using $\mathcal{R}(x_t^i)$, we categorize $\mathcal{D}_t$ into three subsets: (1) **Reliable** ($\mathcal{D}_R$): All MLLMs agree ($\mathcal{R}(x_t^i) = 1$). (2) **Less Reliable** ($\mathcal{D}_{LR}$): Partial agreement ($0 < \mathcal{R}(x_t^i) < 1$). (3) **Unreliable** ($\mathcal{D}_{UR}$): No agreement ($\mathcal{R}(x_t^i) = 0$). Thus, the target dataset is partitioned as: $\mathcal{D}_t = \{\mathcal{D}_R, \mathcal{D}_{LR}, \mathcal{D}_{UR}\}$. As shown in Figure 4, higher reliability scores correlate with improved pseudo-label accuracy, validating our consensus-based reliability metric. Furthermore, Figure 4 shows the accuracy and distribution of the pseudo-label at different levels of reliability $\mathcal{R}(x_t^i)$, highlighting the disagreement between LLaVA, InstructBLIP, and ShareGPT4V. For samples with $\mathcal{R}(x_t^i) > 0$, accuracy is computed using a majority vote across MLLMs, while for $\mathcal{R}(x_t^i) = 0$, we report the accuracy of each individual MLLM.

4. Reliability-based Curriculum Learning

While pseudo-labels from MLLMs provide a strong starting point, their reliability varies across samples. Directly training on all pseudo-labels can introduce noise and hinder adaptation. To address this, we propose Reliability-based Curriculum Learning (RCL), shown in Figure 2, which builds on consensus-based reliability to strategically utilize target domain data. As shown in Figure 2, RCL consists of three

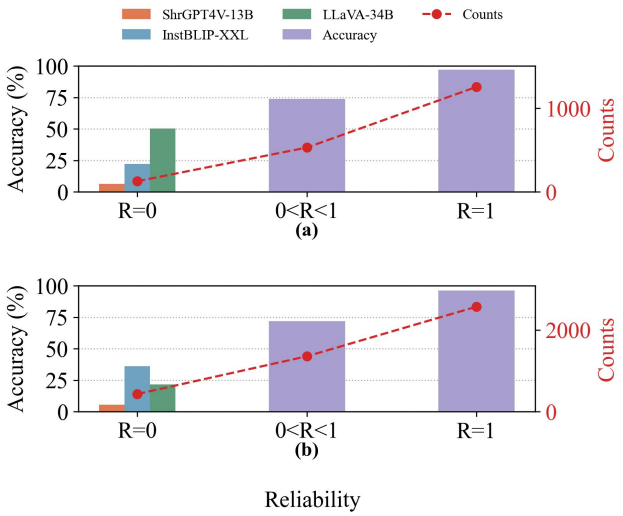


Figure 4. Pseudo-label accuracy and distribution across MLLMs in Office-Home (a) Clipart and (b) Art domains (65 classes each).

stages: (1) **Reliable Knowledge Transfer (RKT)**, (2) **Self-correcting and MLLM-guided Knowledge Expansion (SMKE)**, and (3) **Multi-hot Masking Refinement (MMR)**. In RKT (Sec.4.1), we begin by training the target model on the reliable subset $\mathcal{D}_R$. Next, in SMKE (Sec.4.2), we fine-tune the model using both the reliable and less reliable subsets $\mathcal{D}_R$. Finally, in MMR (Sec. 4.3), we refine the model on the full target dataset $\mathcal{D}_t$, incorporating the proposed Multi-hot Masking and consistency regularization for the unreliable subset $\mathcal{D}_{UR}$. This staged learning strategy enables the model to first learn from reliable samples before gradually incorporating noisier pseudo-labels, reducing errors from premature exposure to unreliable data. By progressively refining its training, the model's learning trajectory is effectively regularized, preventing instability that could arise from training on all data at once.

4.1. Reliable Knowledge Transfer (RKT)

We begin by transferring the most reliable MLLM knowledge to the target model. RKT trains the target model using pseudo-labels from the reliable subset $\mathcal{D}_R$ in a supervised manner. At this stage, the model exclusively relies on high-confidence MLLM pseudo-labels, as it has not yet learned from the target domain. The reliable subset $\mathcal{D}_R$ consists of samples for which all MLLMs agree on the pseudo-label:

$$\mathcal{D}_R = \{(x_r^i, y_r^i) \mid \mathcal{R}(x_r^i) = 1\}, \tag{4}$$

where x_r^i is the i-th sample in the reliable subset, $\tilde{y}_r^i$ is the corresponding pseudo-label agreed upon by all MLLMs, and $\mathcal{R}(x_r^i)$ is the reliability measure defined in the previous section. The target model f_{θ_t} is trained using a supervised cross-entropy loss on the reliable subset $\mathcal{D}_R$:

$$\mathcal{L}_{RKT} = -\frac{1}{|\mathcal{D}_R|} \sum_{(x_r^i, y_r^i) \in \mathcal{D}_R} y_r^i \cdot \log f_{\theta_t}(x_r^i), \tag{5}$$

where $|\mathcal{D}_R|$ denotes the number of samples in the reliable subset. Training exclusively on $\mathcal{D}_R$ ensures that only the most confident and consistent MLLM pseudo-labels shape the target model's initial learning phase. RKT provides a strong foundation before introducing less reliable pseudo-labels, ensuring stable knowledge transfer.

4.2. Self-correcting and MLLM-guided Knowledge Expansion (SMKE)

Following RKT, RCL integrates less reliable pseudo-labels to expand the target model's knowledge. Since the target model has been pre-trained on the source model and fine-tuned with RKT, we transition from direct distillation to a more adaptive learning process. When the target model is confident, it refines its predictions through self-correction. In cases of lower confidence, MLLMs serve as guidance rather than fixed teachers, allowing the model to expand its knowledge dynamically.

To facilitate this, we propose SMKE, which fine-tunes the target model on both the reliable and less reliable subsets, $\mathcal{D}_R \cup \mathcal{D}_{LR}$. This enables the target model to learn from a larger portion of the target domain data and benefit from the additional information provided by $\mathcal{D}_{LR}$. The pseudo-label $\tilde{y}^i$ used for training the target model is determined based on the confidence of the target model's predictions. Let $\hat{y}_t^i$ be the pseudo-label predicted by the target model for the target domain sample x_t^i, and let p_t^i be the corresponding confidence score, which is calculated as the maximum value of the target model's predictive probabilities. We define a confidence threshold τ to determine whether to use the target model's pseudo-label or the MLLMs' pseudo-label. The pseudo-label $\tilde{y}^i$ is as follows:

$$\tilde{y}^i = \begin{cases} \hat{y}_t^i, & \text{if } p_t^i \geq \tau, \\ \text{mode}(\hat{y}^{1i}, \hat{y}^{2i}, \ldots, \hat{y}^{Mi}), & \text{if } p_t^i < \tau, \end{cases} \tag{6}$$

where $\text{mode}(\cdot)$ returns the most frequent pseudo-label among the MLLMs.

In SMKE, if the target model's confidence score $p_t^i > \tau$, (τ being given threshold), we employ the target model's pseudo-label $\hat{y}_t^i$ for self-correction. Otherwise, the model adopts the most frequent MLLM pseudo-label to mitigate uncertainty and expand knowledge. The adaptive training approach is optimized through the loss function:

$$\mathcal{L}_{\text{SMKE}} = -\frac{1}{|\mathcal{D}_R \cup \mathcal{D}_{LR}|} \sum_{x_t^i \in \{\mathcal{D}_R \cup \mathcal{D}_{LR}\}} \tilde{y}^i \cdot \log f_{\theta_t}(x_t^i), \tag{7}$$

By incorporating the less reliable pseudo-labels and leveraging the target model's confidence scores, SMKE stage of the curriculum learning framework can expand the knowledge transferred to the target model by utilizing both the target model's predictions and the MLLMs' pseudo-labels.

4.3. Multi-hot Masking Refinement (MMR)

Finally, we expand to the full training set to incorporate the unreliable subset $\mathcal{D}_{UR}$, which consists of target domain samples where the MLLMs disagree on the pseudo-labels. This disagreement makes it challenging to assign reliable pseudo-labels to the samples in $\mathcal{D}_{UR}$. Thus, these samples are not included during RKT and SMKE. To fully exploit $\mathcal{D}_t$, we propose Multi-hot Masking Refinement (MMR), integrating joint selection, multi-hot masking, and consistency regularization to utilize these challenging samples. The full algorithm for MMR is shown in Algo. 1.

Multi-hot masking. Let $\mathbf{z}_t^i \in \mathbb{R}^C$ be the predictive probabilities of the target model for the target domain sample x_t^i, where C is the number of classes. The confidence score of the target model's prediction is given by $p_t^i = \max_c \mathbf{z}_t^i$. We define a Multi-hot mask $\mathbf{m}^i \in \{0,1\}^C$ based on the pseudo-labels assigned by the MLLMs: $\mathbf{m}^i = 1 - \prod_{m=1}^{M}(1 - \mathbb{1}(\hat{y}^{mi}))$ where the mask $\mathbf{m}^i$ is formed by adding up one-hot vectors indicating the presence of each class as predicted by the MLLMs for the sample x_t^i. We then apply the Multi-hot mask to mask out the target model's logits, forming a refined pseudo-label $\tilde{y}^i$ based on the confidence threshold τ:

$$\tilde{y}^i = \begin{cases} \text{argmax}_C(\mathbf{z}_t^i), & \text{if } p_t^i \geq \tau, \\ \text{argmax}_C(\mathbf{z}_t^i \odot \mathbf{m}^i), & \text{if } p_t^i < \tau, \end{cases} \tag{8}$$

where $\odot$ denotes element-wise multiplication and τ is the confidence threshold. If $p_t^i > \tau$, the original prediction is retained; otherwise, it is adjusted based on the multi-hot mask, filtering out less likely classes.

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{D}_R \cup \mathcal{D}_{LR} \cup \mathcal{D}_{UR}|} \sum_{x_t^i \in \{\mathcal{D}_R \cup \mathcal{D}_{LR} \cup \mathcal{D}_{UR}\}} \tilde{y}^i \cdot \log f_{\theta_t}(x_t^i), \tag{9}$$

The consistency loss $\mathcal{L}_{\text{cons}}$ is computed using the refined pseudo-labels from both weakly and strongly augmented samples, reinforcing target model predictions to align with MLLMs, especially when the model is not confident:

$$\mathcal{L}_{\text{cons}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N_t} \text{H}(\tilde{y}^i, \mathbf{z}_{st}^i), \tag{10}$$

where $\mathbf{z}_{st}^i$ denotes the target model's logit for strong augmentation samples and $\text{H}(\cdot,\cdot)$ denotes the cross-entropy loss. The target model is then optimized through the combined loss $\mathcal{L}_{\text{MMR}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}$ where λ_{cons} is a fixed hyper-parameter to balance the supervised and consistency losses. Through the MMR phase, the target model not only uses the MLLMs' pseudo-labels to refine its training strategy but also ensures robust learning even from samples whose initial predictions lack confidence.

Algorithm 1 Multi-hot Masking Refinement (MMR)

Input: Unlabeled dataset $\mathcal{D}_t$, confidence threshold τ, model f_{θ_t}, MLLM outputs $\hat{y}^{mi}$
for each sample $x_t^i \in \mathcal{D}_t$ **do**
  Generate augmented views: $x_{t,weak}^i$ (weak) and $x_{t,strong}^i$ (strong)
  Obtain model predictions $\mathbf{z}_{t,weak}^i$ and $\mathbf{z}_{t,strong}^i$
  Compute confidence score $p_t^i = \max_c \mathbf{z}_{t,weak}^i$
  if $p_t^i \geq \tau$ **then**
    Assign pseudo-label $\tilde{y}^i = \text{argmax}_C\, \mathbf{z}_{t,weak}^i$
  else
    Compute multi-hot mask $\mathbf{m}^i$ from MLLM outputs:

$$\mathbf{m}^i = 1 - \prod_{m=1}^{M} (1 - \mathbb{1}(\hat{y}^{mi}))$$

    Refine pseudo-label:

$$\tilde{y}^i = \begin{cases} \text{argmax}_C(\mathbf{z}_t^i), & \text{if } p_t^i \geq \tau, \\ \text{argmax}_C(\mathbf{z}_t^i \odot \mathbf{m}^i), & \text{if } p_t^i < \tau, \end{cases}$$

  end if
  Compute supervised loss:

$$\mathcal{L}_{\text{sup}} = -\tilde{y}^i \cdot \log f_{\theta_t}(x_{t,weak}^i)$$

  Compute consistency loss:

$$\mathcal{L}_{\text{cons}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N_t} \text{H}(\tilde{y}^i, \mathbf{z}_{st}^i)$$

  Update model by minimizing:

$$\mathcal{L}_{\text{MMR}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}$$

end for

5. Experiments

Datasets. We evaluate our method on three standard benchmark datasets: Office-Home, DomainNet-126, and VisDA-C 2017. Office-Home [32] has 4 domains – Real (R), Clipart (C), Art (A), and Product (P), encompassing 65 classes with a total of 15.5k images. VisDA [26] is a large-scale synthetic-to-real object recognition dataset, where the source domain includes 152k synthetic images and the target domain contains about 55k real object images across 12 classes. DomainNet [27] is a challenging large-scale, featuring 6 domains with a total of around 600k images across 345 classes. We follow the standard DomainNet-126 setup with 145k im-

ages from 126 classes, sampled from four domains, Clipart (C), Painting (P), Real (R), Sketch (S).

Model details. Following [20, 39, 41], we use ResNet-101 [8] for VisDA and ResNet-50 for Office-Home and DomainNet. For VisDA and Office-Home, we adopt pre-trained source models from SHOT [20], while for DomainNet, we train source models from scratch following [23]. Similar to DIFO [41], we also report results with ViT-B/32 backbone (RCL-ViT). For main results, we use the strongest open-source MLLMs for RCL: LLaVA-v1.6-34B [25], ShareGPT4V-13B [2], and InstructBLIP-T5-XXL [4].

Training Details. The parameters used in the training process of RCL are shown in Table 1. We use the Adam optimizer [14] for RCL training. All experiments were conducted using PyTorch on NVIDIA A100 GPUs.

Table 1. RCL Training Parameters per stage (RKT, SMKE, MMR).

	Office-Home			DomainNet			VisDA		
	RKT	SMKE	MMR	RKT	SMKE	MMR	RKT	SMKE	MMR
learning rate	1e-04	1e-05	1e-05	1e-05	1e-05	1e-05	1e-05	1e-06	1e-06
τ	–	0.7	0.95	–	0.7	0.9	–	0.7	0.6
batch size	64	256	128	64	256	64	64	256	256
max iter	3000	5000	5000	8000	10000	5000	6000	6000	5000

5.1. Main Results

Tables 2, 3, and 4 present our results for Office-Home, DomainNet, and VisDA, respectively. From top to bottom, we report domain adaptation methods that (1) use source data with CLIP-based techniques, (2) are source-free without multimodal or CLIP, and (3) employ CLIP-based multimodal approaches. Zero-shot MLLM (w/ STS)/CLIP performance is also included as a reference. *RCL consistently achieves state-of-the-art performance across all datasets, with notable improvements: +6.4% on Office-Home, +9.4% on DomainNet, and +2.9% on VisDA-C*. The prior best-performing methods, DIFO-C-B32 [41] and DIFO-C-RN [41], employ a ViT-B/32 CLIP encoder and ResNet CLIP backbone, respectively, using prompt learning. In contrast, *RCL surpasses these methods while using only a ResNet backbone with guided curriculum training through MLLM pseudo-labels—without prompt learning or additional tuning*. Similarly, PSAT-GDA [39], another competitive method, trains transformers for source guidance and domain alignment, whereas RCL relies solely on zero-shot MLLM inference and requires no additional tuning or training of large VLMs. Our self-refinement and curriculum learning processes outperform standard MLLM zero-shot performance by capturing valuable latent information beyond MLLMs. Additionally, the reliability of our pseudo-labels enables full-data training, further distinguishing RCL from prior approaches that rely on prompt tuning or handcrafted domain alignment. RCL leverages MLLM inference with STS and *does not require customization, prompt learning, or heavy training of multimodal models.*

5.2. Ablation Studies

This section covers ablation studies on RCL across its components, MLLMs used, backbones, and hyperparameters.

5.2.1. Impact of RCL components.

Table 5 shows the results of evaluating individual RCL components. Using only RKT yields the lowest performance, as it relies on the most reliable MLLM pseudo-labels, which may lack class coverage and diversity (see `Supplementary`). Nonetheless, *RKT provides sufficient initial supervision for identifying essential features.* Applying SMKE after RKT outperforms MMR after RKT, as *SMKE leverages less reliable pseudo-labels, improving robustness and expanding knowledge.* In contrast, MMR directly after RKT performs worse, indicating the need for pseudo-label diversity before semi-supervised integration. Finally, *MMR following SMKE consistently improves performance, allowing the model to learn from even the most unreliable labels in a semi-supervised manner, maximizing dataset utilization.* Figure 8 visually compares the feature distributions of SOTA methods (DIFO) and RCL, showing RKT, SMKE, and MMR progressively refine target features.

5.2.2. Synergy between RCL and MLLMs.

Table 6 shows that without MLLMs, RCL provides only a marginal 0.1% gain over the best existing SFDA method (LCFD-C-B32), indicating limited improvement when applied to standard adaptation approaches (TPDS, LCFD-C-B32, DIFO-C-B32). In contrast, integrating MLLMs (LLaVA-34B, ShareGPT4V-13B, InstBLIP-XXL) into RCL yields an 2.8% improvement over the best MLLM model (LLaVa-34B) indicated in Table 2 and a 6.4% boost over RCL without MLLMs. Unlike ImageNet-pretrained models, MLLMs are trained on broad multimodal corpora, providing valuable auxiliary knowledge that enhances pseudo-label quality and adaptation effectiveness. These results confirm that traditional SFDA methods lack the generalization capacity of MLLMs, while RCL effectively leverages their complementary knowledge for improved adaptation.

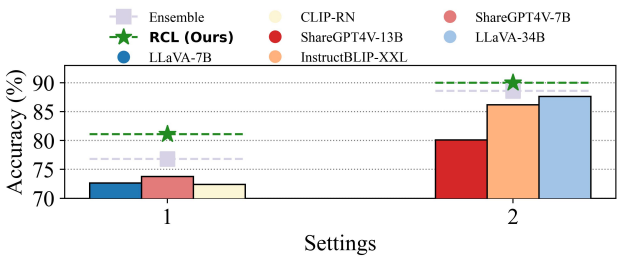


Figure 5. RCL's sensitivity and robustness against MLLMs with weaker capability and MLLM ensemble.

Table 2. Accuracy (%) on **Office-Home** dataset. SF: source-free, CP, ViT: method uses CLIP, ViT. We highlight the best result and underline the second-best one. (*) represents pre-trained CLIP/MLLM zero-shot performance with proposed STS.

Method	SF	CP	ViT	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.
Source	-	✗	✗	44.7	64.2	69.4	48.3	57.9	60.3	49.5	40.3	67.2	59.7	45.6	73.0	56.7
DAPL-RN [7]	✗	✓	✗	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
PADCLIP-RN [15]	✗	✓	✗	57.5	84.0	83.8	77.8	85.5	84.7	76.3	59.2	85.4	78.1	60.2	86.7	76.6
ADCLIP-RN [33]	✗	✓	✗	55.4	85.2	85.6	76.1	85.8	86.2	76.7	56.1	85.4	76.8	56.1	85.5	75.9
SHOT [20]	✓	✗	✗	56.7	77.9	80.6	68.0	78.0	79.4	67.9	54.5	82.3	74.2	58.6	84.5	71.9
NRC [45]	✓	✗	✗	57.7	80.3	82.0	68.1	79.8	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2
GKD [34]	✓	✗	✗	56.5	78.2	81.8	68.7	78.9	79.1	67.6	54.8	82.6	74.4	58.5	84.8	72.2
AaD [47]	✓	✗	✗	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7
AdaCon [1]	✓	✗	✗	47.2	75.1	75.5	60.7	73.3	73.2	60.2	45.2	76.6	65.6	48.3	79.1	65.0
CoWA [17]	✓	✗	✗	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5
SCLM [36]	✓	✗	✗	58.2	80.3	81.5	69.3	79.0	80.7	69.0	56.8	82.7	74.7	60.6	85.0	73.0
ELR [48]	✓	✗	✗	58.4	78.7	81.5	69.2	79.5	79.3	66.3	58.0	82.6	73.4	59.8	85.1	72.6
PLUE [23]	✓	✗	✗	49.1	73.5	78.2	62.9	73.5	74.5	62.2	48.3	78.6	68.6	51.8	81.5	66.9
TPDS [38]	✓	✗	✗	59.3	80.3	82.1	70.6	79.4	80.9	69.8	56.8	82.1	74.5	61.2	85.3	73.5
C-SFDA [13]	✓	✗	✗	60.3	80.2	82.9	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5
PSAT-GDA [39]	✓	✗	✓	73.1	88.1	89.2	82.1	88.8	88.9	83.0	72.0	89.6	83.3	73.7	91.3	83.6
LCFD-C-RN [40]	✓	✓	✗	60.1	85.6	86.2	77.2	86.0	86.3	76.6	61.0	86.5	77.5	61.4	86.2	77.6
LCFD-C-B32 [40]	✓	✓	✓	72.3	89.8	89.9	81.1	90.3	89.5	80.1	71.5	89.8	81.8	72.7	90.4	83.3
DIFO-C-RN [41]	✓	✓	✗	62.6	87.5	87.1	79.5	87.9	87.4	78.3	63.4	88.1	80.0	63.3	87.7	79.4
DIFO-C-B32 [41]	✓	✓	✓	70.6	90.6	88.8	82.5	90.6	88.8	80.9	70.1	88.9	83.4	70.5	91.2	83.1
CLIP-RN [30]*	-	✓	✗	51.7	85.0	83.7	69.3	85.0	83.7	69.3	51.7	83.7	69.3	51.7	85.0	72.4
LLaVA-34B [25]* (w/ STS)	-	✓	✓	78.3	93.7	89.5	87.0	93.7	89.5	87.0	78.3	89.5	87.0	78.3	93.7	87.2
InstBLIP-XXL [4]* (w/ STS)	-	✓	✓	82.0	91.6	88.8	82.2	91.6	88.8	82.2	82.0	88.8	82.2	82.0	91.6	86.2
ShrGPT4V-13B [2]* (w/ STS)	-	✓	✓	66.7	85.8	84.8	83.2	85.8	84.8	83.2	66.7	84.8	83.2	66.7	85.8	80.1
RCL (Ours)	✓	✗	✗	82.5	95.3	**93.3**	89.1	**95.3**	**92.7**	**89.3**	**82.4**	92.8	89.4	82.1	95.4	90.0
RCL-ViT (Ours)	✓	✗	✓	**83.1**	**95.7**	93.1	**89.2**	**95.3**	92.6	89.2	82.3	**92.9**	**90.0**	**83.2**	**95.5**	**90.2**

Table 3. Accuracy (%) on **DomainNet dataset**.

Method	SF	CP	ViT	C→P	C→R	C→S	P→C	P→R	P→S	R→C	R→P	R→S	S→C	S→P	S→R	Avg.
Source	-	✗	✗	42.6	53.7	51.9	52.9	66.7	51.6	49.1	56.8	43.9	60.9	48.6	53.2	52.7
DAPL-RN [7]	✗	✓	✗	72.4	87.6	65.9	72.7	87.6	65.6	73.2	72.4	66.2	73.8	72.9	87.8	74.8
ADCLIP-RN [15]	✗	✓	✗	71.7	88.1	66.0	73.2	86.9	65.2	73.6	73.0	68.4	72.3	74.2	89.3	75.2
SHOT [20]	✓	✗	✗	63.5	78.2	59.5	67.9	81.3	61.7	67.7	67.6	57.8	70.2	64.0	78.0	68.1
NRC [45]	✓	✗	✗	62.6	77.1	58.3	62.9	81.3	60.7	64.7	69.4	58.7	69.4	65.8	78.7	67.5
GKD [34]	✓	✗	✗	61.4	77.4	60.3	69.6	81.4	63.2	68.3	68.4	59.5	71.5	65.2	77.6	68.7
AdaCon [1]	✓	✗	✗	60.8	74.8	55.9	62.2	78.3	58.2	63.1	68.1	55.6	67.1	66.0	75.4	65.4
CoWA [17]	✓	✗	✗	64.6	80.6	60.6	66.2	79.8	60.8	69.0	67.2	60.0	69.0	65.8	79.9	68.6
PLUE [23]	✓	✗	✗	59.8	74.0	56.0	61.6	78.5	57.9	61.6	65.9	53.8	67.5	64.3	76.0	64.7
TPDS [38]	✓	✗	✗	62.9	77.1	59.8	65.6	79.0	61.5	66.4	67.0	58.2	68.6	64.3	75.3	67.1
LCFD-C-RN [40]	✓	✓	✗	75.4	88.2	72.0	75.8	88.3	72.1	76.1	75.6	71.2	77.6	75.9	88.2	78.0
LCFD-C-B32 [40]	✓	✓	✓	77.2	88.0	75.2	78.8	88.2	75.8	79.1	77.8	74.9	79.9	77.4	88.0	80.0
DIFO-C-RN [41]	✓	✓	✗	73.8	89.0	69.4	74.0	88.7	70.1	74.8	74.6	69.6	74.7	74.3	88.0	76.7
DIFO-C-B32 [41]	✓	✓	✓	76.6	87.2	74.9	80.0	87.4	75.6	80.8	77.3	75.5	80.5	76.7	87.3	80.0
LLaVA-34B [25]* (w/ STS)	-	✓	✓	84.4	91.0	83.7	85.5	91.0	83.7	85.5	84.4	83.7	85.5	84.4	91.0	86.1
InstBLIP-XXL [4]* (w/ STS)	-	✓	✓	82.5	89.0	83.0	86.7	89.0	83.0	86.7	82.5	83.0	86.7	82.5	89.0	85.3
ShrGPT4V-13B [2]* (w/ STS)	-	✓	✓	79.7	87.9	79.2	79.9	87.9	79.2	79.9	79.7	79.2	79.9	79.7	87.9	81.7
RCL (Ours)	✓	✗	✗	87.6	92.8	87.9	89.2	92.7	87.8	89.6	87.7	87.6	89.4	87.5	92.7	89.4
RCL-ViT (Ours)	✓	✗	✓	**88.1**	**93.3**	**88.0**	**89.7**	**93.3**	**88.0**	**89.7**	**88.0**	**87.8**	**89.7**	**88.1**	**93.3**	**89.7**

5.2.3. Sensitivity to the capability of MLLMs.

Figure 5 compares two settings: (1) weaker MLLMs with lower zero-shot performance and (2) the strongest MLLMs with the highest ensemble accuracy. Labels are determined by majority vote, with ties assigned randomly. *RCL consistently surpasses individual MLLMs and their ensemble*, with the gap most pronounced in Setting 1 (4.3% increase), where weaker MLLMs struggle. Even in Setting 2, where MLLMs perform well, RCL achieves superior results over their ensemble (1.4% increase). *These results confirm RCL's ability to consistently enhance MLLM performance, especially when MLLMs have low individual performance.*

Table 4. Accuracy (%) on **VisDA-C** dataset.

Method	SF	CP	ViT	plane	bcyle	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
Source	-	✗	✗	60.4	22.5	44.8	73.4	60.6	3.28	81.3	22.1	62.2	24.8	83.7	4.81	45.3
DAPL-RN [7]	✗	✓	✗	97.8	83.1	88.8	77.9	97.4	91.5	94.2	79.7	88.6	89.3	92.5	62.0	86.9
PADCLIP-RN [15]	✗	✓	✗	96.7	88.8	87.0	82.8	97.1	93.0	91.3	83.0	95.5	91.8	91.5	63.0	88.5
ADCLIP-RN [33]	✗	✓	✗	98.1	83.6	91.2	76.6	98.1	93.4	96.0	81.4	86.4	91.5	92.1	64.2	87.7
SHOT [20]	✓	✗	✗	95.0	87.4	80.9	57.6	93.9	94.1	79.4	80.4	90.9	89.8	85.8	57.5	82.7
NRC [45]	✓	✗	✗	96.8	91.3	82.4	62.4	96.2	95.9	86.1	90.7	94.8	94.1	90.4	59.7	85.9
GKD [34]	✓	✗	✗	95.3	87.6	81.7	58.1	93.9	94.0	80.0	80.0	91.2	91.0	86.9	56.1	83.0
AaD [47]	✓	✗	✗	97.4	90.5	80.8	76.2	97.3	96.1	89.8	82.9	95.5	93.0	92.0	64.7	88.0
AdaCon [1]	✓	✗	✗	97.0	84.7	84.0	77.3	96.7	93.8	91.9	84.8	94.3	93.1	94.1	49.7	86.8
CoWA [17]	✓	✗	✗	96.2	89.7	83.9	73.8	96.4	**97.4**	89.3	**86.8**	94.6	92.1	88.7	53.8	86.9
SCLM [36]	✓	✗	✗	97.1	90.7	85.6	62.0	97.3	94.6	81.8	84.3	93.6	92.8	88.0	55.9	85.3
ELR [48]	✓	✗	✗	97.1	89.7	82.7	62.0	96.2	97.0	87.6	81.2	93.7	94.1	90.2	58.6	85.8
PLUE [23]	✓	✗	✗	94.4	91.7	89.0	70.5	96.6	94.9	92.2	88.8	92.9	95.3	91.4	61.6	88.3
TPDS [38]	✓	✗	✗	97.6	91.5	89.7	83.4	97.5	96.3	92.2	82.4	96.0	94.1	90.9	40.4	87.6
C-SFDA [13]	✓	✗	✗	97.6	88.8	86.1	72.2	97.2	94.4	92.1	84.7	93.0	90.7	93.1	63.5	87.8
PSAT-GDA [39]	✓	✗	✓	97.5	92.4	89.9	72.5	98.2	96.5	89.3	55.6	95.7	**98.2**	95.3	54.8	86.3
DIFO-C-RN [41]	✓	✓	✗	97.7	87.6	90.5	83.6	96.7	95.8	94.8	74.1	92.4	93.8	92.9	65.5	88.8
DIFO-C-B32 [41]	✓	✓	✓	97.5	89.0	90.8	83.5	97.8	97.3	93.2	83.5	95.2	96.8	93.7	65.9	90.3
LLaVA-34B [25]* (w/ STS)	-	✓	✓	99.4	97.3	94.8	83.9	98.9	95.8	95.9	80.9	92.7	98.8	97.4	68.9	92.1
InstBLIP-XXL [4]*(w/ STS)	-	✓	✓	99.2	89.6	82.0	69.8	97.9	91.0	97.5	84.3	73.6	99.3	96.7	60.0	86.7
ShrGPT4V-13B [2]*(w/ STS)	-	✓	✓	99.2	94.7	90.8	87.9	98.3	92.1	97.3	68.0	96.3	95.6	96.8	68.2	90.4
RCL (Ours)	✓	✗	✗	**99.5**	**96.1**	**92.6**	**89.4**	**99.1**	97.1	**97.0**	85.8	**96.6**	98.1	**97.3**	**70.0**	**93.2**

Table 5. Impact of RCL components (RKT, SMKE, MMR).

RCL			Office-Home				
RKT	SMKE	MMR	→A	→C	→P	→R	Avg.
✓	✗	✗	82.8	73.3	89.3	88.1	83.3
✓	✓	✗	88.5	80.9	95.1	92.5	89.3
✓	✗	✓	87.7	80.2	93.3	92.0	88.3
✓	✓	✓	**89.3**	**82.3**	**95.3**	**92.9**	**90.0**

Table 6. Performance of RCL with and without MLLMs on the Office-Home dataset. RCL (w/o MLLMs) uses three baseline methods (TPDS, LCFD-C-B32, and DIFO-C-B32).

Method	→C	→P	→R	→A	Avg.
TPDS	59.1	81.7	81.7	71.6	73.5
LCFD-C-B32	72.2	90.2	89.7	81.0	83.3
DIFO-C-B32	70.4	90.8	88.8	82.3	83.1
RCL (w/o MLLMs)	71.9	90.7	89.2	81.7	83.4
RCL (w/ MLLMs)	**82.3**	**95.3**	**92.9**	**89.3**	**90.0**

5.2.4. Number of MLLMs.

Figure 6 shows the impact of MLLM count on RCL performance, comparing using a single MLLM to four. We use the strongest models individually (ShareGPT-13B, InstructBLIP-XXL, LLaVA-34B), combining them for three, and adding BLIP2-XXL for four. *A single MLLM yields the lowest accuracy* (see Figure 7) since RKT-only learning is constrained by the MLLM-STS zero-shot upper bound. In contrast, *RCL surpasses each MLLM by integrating insights and adapting knowledge*. Comparing three vs. four MLLMs, *we find no significant difference*, indicating RCL efficiently leverages multiple MLLMs without requiring a large ensemble.

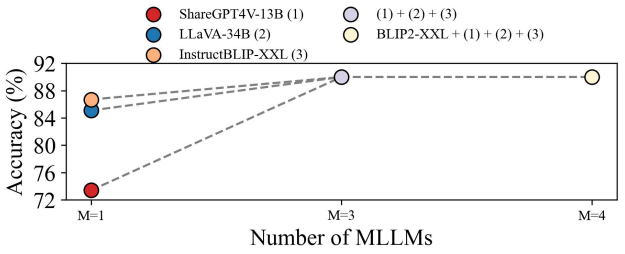


Figure 6. Impact of number of MLLMs on RCL performance, the performance of using single MLLM is with RKT only.

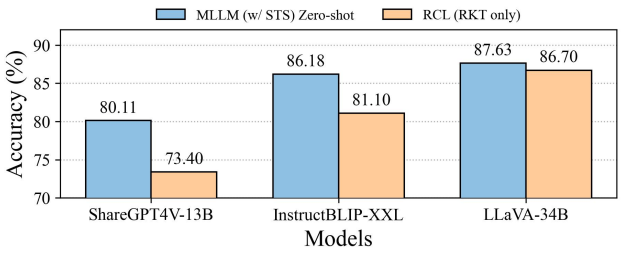


Figure 7. Distillation from single MLLM to RCL cannot surpass the teacher model

5.2.5. Knowledge transfer to a smaller backbone.

We investigate transferring to a smaller backbone for the scalability of SFDA. As shown in Table 7, RCL achieves similar

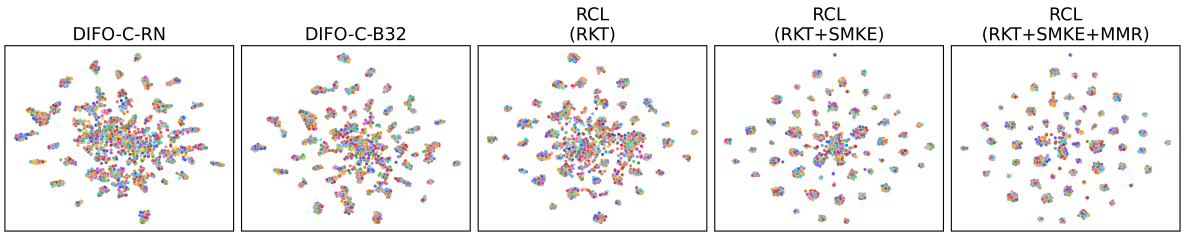


Figure 8. t-SNE feature distribution for A→C in Office-Home. DIFO-C-B32 uses ViT-B32; others use ResNet-50.

Table 7. Ablation study on the choice of backbone (BB).

Method	BB	Office-Home				
		→A	→C	→P	→R	Avg.
DIFO-C-RN	RN50	79.3	63.1	87.7	87.5	79.4
DIFO-C-B32	RN50	82.3	70.4	90.8	88.3	83.1
RCL (Ours)	RN18	89.1	81.5	95.1	92.6	89.6
RCL (Ours)	RN50	**89.3**	**82.3**	**95.3**	**92.9**	**90.0**

performance with ResNet18 as with ResNet50, *maintaining a +6.4% accuracy advantage over state-of-the-art methods*. With twice the speed (2 GFLOPS v.s. 4 GFLOPS), *RCL effectively distills pre-trained knowledge to a smaller backbone*, making it ideal for large-scale inference across diverse tasks. This demonstrates RCL's capability to efficiently leverage MLLM knowledge and adapt it to a lightweight model suitable for real-world deployment.

Table 8. Average Latency (ms / sample) comparison for inference. RCL uses RN50 backbone.

Model	Avg. Latency ↓
LLaVA-34B (w/ STS)	∼2850
ShrGPT4v-13B (w/ STS)	∼1890
InsBLIP-XXL (w/ STS)	∼2740
RCL	∼**5**

5.2.6. Latency of deployed model.

Table 8 illustrates the inference latency comparison of MLLMs and our proposed RCL. While LLaVA-34B, InstructBLIP-XXL, and ShareGPT4v-13B require latency over 1800ms per sample, RCL achieves a significantly faster inference speed of 5ms per sample. When combining all three MLLMs for ensembling, over 7000ms latency is required. This over 378x improvement in computational efficiency validates our approach of distilling MLLM knowledge into a compact model rather than relying on direct MLLM inference, enabling practical deployment while achieving even better performance. All SFDA methods using ResNet-50 have comparable inference latency. RCL adds complexity only during initial pseudo-label generation with MLLMs but maintains efficient training and inference post-deployment while achieving superior performance. In contrast, DIFO [41] incorporates additional models (e.g., CLIP-RN, CLIP-ViT) in every training iteration, increasing computational overhead throughout training.

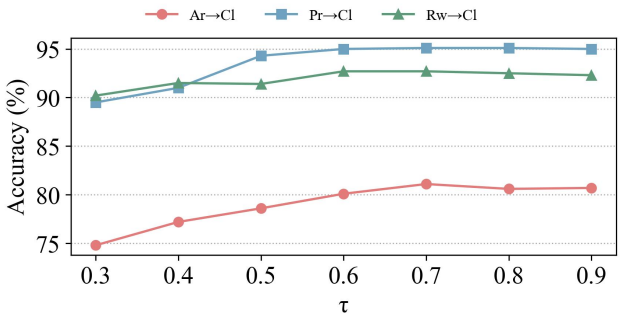


Figure 9. The effect of τ in SMKE.

5.2.7. Sensitivity of hyper-parameters.

Figure 9 illustrates the effect of the confidence threshold τ in SMKE. It shows how varying τ values influence model performance by balancing self-correction with MLLM guidance. Higher τ values increase reliance on high-confidence pseudo-labels from the target model, while lower values depend more on MLLM-generated pseudo-labels. The figure highlights the optimal τ that enhances adaptation performance by effectively leveraging both reliable and less reliable pseudo-labels. For the MMR, Table 9 shows that $\tau = 0.95$ provided a slight edge in accuracy, ensuring the model effectively utilizes its predictions while still considering MLLM guidance. In Table 10, we evaluate λ values from 0 to 2.0. The results show that $\lambda = 0.5$ consistently yielded the highest average accuracy across multiple domain adaptation tasks, suggesting it strikes a good balance between supervised learning and regularization.

Table 9. The effect of confidence threshold τ in MMR.

τ	1.0	0.95	0.85	0.75	0.65
Ar→Cl	82.3	82.5	82.2	82.0	81.9
Ar→Pr	95.4	95.3	95.3	95.3	95.3
Ar→Rw	93.1	93.3	93.3	93.3	93.4

Table 10. Effect of different λ values in MMR.

λ	→C	→P	→R	→A	Avg.
0	82.0	95.3	92.8	89.1	89.8
0.5	82.3	95.3	92.9	89.3	**90.0**
1	82.3	95.3	92.8	89.2	89.9
1.5	82.2	95.2	92.8	89.2	89.9
2	82.3	95.2	92.8	89.1	89.9

5.2.8. MMR Complexity and Effectiveness.

MMR demonstrates enhanced effectiveness for challenging adaptation tasks, with performance gains correlating to the proportion of unreliable samples. For example, the (→ Clipart) achieves the most significant improvement of +1.4% (80.9% to 82.3% in Table 5), despite having the highest proportion of unreliable (R=0) samples (see Fig. 10). These results indicate MMR's particular utility when dealing with large portions of unreliable labels, which aligns with real-world scenarios where MLLMs possess divergent knowledge spaces. MMR enables comprehensive data utilization through its ability to learn from all samples, rather than being limited to only high-confidence instances. Our lightweight inference model ensures deployment efficiency, achieving a practical latency of 5ms per sample.

Regarding the effectiveness of MMR components, we have conducted detailed ablation studies (Tables 5 and 10). These results demonstrate that both multi-hot masking and consistency loss contribute to the performance. Note that the MMR cannot be applied without Multi-hot Masking since the unreliable samples cannot be used by traditional pseudo-labeling methods.

6. Conclusion

We introduce a novel approach for adapting foundation knowledge MLLMs to significantly enhance SFDA, transforming their zero-shot capabilities into a structured adaptation process. Since MLLMs are primarily designed for text generation and may produce inconsistent outputs, we first introduce STS to align their zero-shot predictions with target classes, enabling their use as reliable pseudo-labelers in SFDA. Building on this, we propose RCL, which systematically leverages multiple MLLMs to refine pseudo-labels and improve adaptation. RCL employs curriculum learning to dynamically segment pseudo-labels by reliability, progressively refining knowledge and adaptation. This enables a self-correcting learning process that distills MLLM knowledge into a lightweight model, making it more practical for real-world deployment. While our focus is on using MLLMs for SFDA enhancement, this work serves as a broader demonstration of how foundation MLLM knowledge can enhance vision tasks and beyond.

Limitation. Although RCL's generalizability may still be affected by biases inherent in pre-trained MLLMs, our consensus-based approach attempts to mitigate this by aggregating knowledge from multiple MLLMs, reducing dependence on any single model's bias.

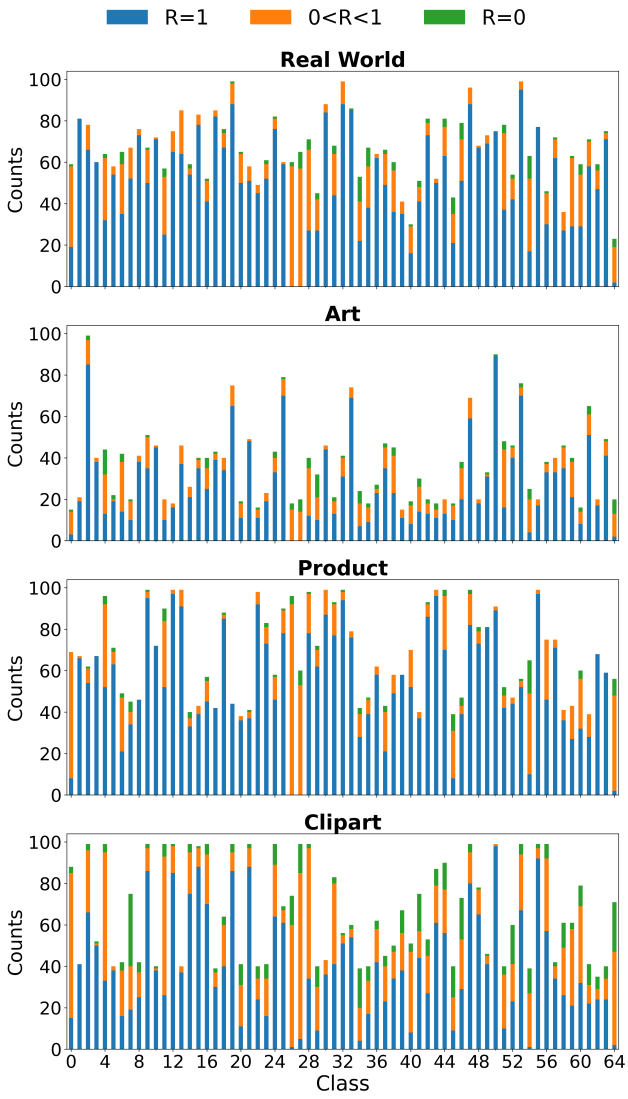


Figure 10. Per-class counts of MLLM pseudolabels on Office-Home.

References

[1] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022. 8, 9

[2] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793*, 2023. 4, 7, 8, 9

[3] Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu,

and Yueting Zhuang. Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10185–10192. IEEE, 2022. 2

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 7, 8, 9

[5] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7212–7222, 2022. 3

[6] Zixiang Ding, Guoqing Jiang, Shuai Zhang, Lin Guo, and Wei Lin. How to trade off the quantity and capacity of teacher ensemble: Learning categorical distribution to stochastically employ a teacher for distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17915–17923, 2024. 3

[7] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1, 8, 9

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. ALIGN: Towards Bridging Alignment for Large-scale Multi-modal Understanding. In *International Conference on Machine Learning (ICML)*, 2021. 3

[10] Jingjing Jiang, Ke Chen, and Hao Yang. Neighborhood Reconstructed Clustering for Source-Free Domain Adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[11] Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull. Mtkd: Multi-teacher knowledge distillation for image super-resolution. In *European Conference on Computer Vision*, pages 364–382. Springer, 2024. 3

[12] SeongKu Kang, Wonbin Kweon, Dongha Lee, Jianxun Lian, Xing Xie, and Hwanjo Yu. Distillation from heterogeneous models for top-k recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 801–811, 2023. 3

[13] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24120–24131, 2023. 2, 8, 9

[14] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. 7

[15] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *ICCV*, 2023. 1, 3, 8, 9

[16] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2691–2701, 2024. 3

[17] Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *ICML*, 2022. 8, 9

[18] Lei Li, Yankai Lin, Xuancheng Ren, Guangxiang Zhao, Peng Li, Jie Zhou, and Xu Sun. From mimicking to integrating: knowledge integration for pre-trained language models. *arXiv preprint arXiv:2210.05230*, 2022. 3

[19] Xiang Li, Xin Sun, Yilu Wu, and Hongxun Xu. Model Adaptation: Unsupervised Domain Adaptation Without Source Data. *arXiv preprint arXiv:2006.09785*, 2020. 3

[20] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1, 7, 8, 9

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[22] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7640–7650, 2023. 2

[23] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding Pseudo-Labels With Uncertainty Estimation for Source-Free Unsupervised Domain Adaptation. In *CVPR*, 2023. 7, 8, 9

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 1, 3, 4

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023. 1, 2, 4, 7, 8, 9

[26] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 6

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 6

[28] Cuong Pham, Tuan Hoang, and Thanh-Toan Do. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6435–6443, 2023. 3

[29] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric

dynamic prototype strategy for source-free domain adaptation. In *European conference on computer vision*, pages 165–182. Springer, 2022. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3, 8

[31] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020. 4

[32] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010. 6

[33] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. AD-CLIP: Adapting Domains in Prompt Space Using CLIP. In *ICCV Workshop*, 2023. 1, 8, 9

[34] S Tang, Yuji Shi, Zhiyuan Ma, Jian Li, Jianzhi Lyu, Qingdu Li, and Jianwei Zhang. Model adaptation through hypothesis transfer with gradual knowledge distillation. In *IROS*, 2021. 8, 9

[35] Song Tang, Yan Yang, Zhiyuan Ma, Norman Hendrich, Fanyu Zeng, Shuzhi Sam Ge, Changshui Zhang, and Jianwei Zhang. Nearest neighborhood-based deep clustering for source data-absent unsupervised domain adaptation. *arXiv:2107.12585*, 2021. 3

[36] Song Tang, Yan Zou, Zihao Song, Jianzhi Lyu, Lijuan Chen, Mao Ye, Shouming Zhong, and Jianwei Zhang. Semantic consistency learning on manifold for source data-free unsupervised domain adaptation. *Neural Networks*, 152, 2022. 8, 9

[37] Song Tang, Yan Zou, Zihao Song, Jianzhi Lyu, Lijuan Chen, Mao Ye, Shouming Zhong, and Jianwei Zhang. Semantic consistency learning on manifold for source data-free unsupervised domain adaptation. *Neural Networks*, 152:467–478, 2022. 3

[38] Song Tang, An Chang, Fabian Zhang, Xiatian Zhu, Mao Ye, and Changshui Zhang. Source-Free Domain Adaptation via Target Prediction Distribution Searching. *International Journal of Computer Vision*, pages 1–19, 2023. 8, 9

[39] Song Tang, Yuji Shi, Zihao Song, Mao Ye, Changshui Zhang, and Jianwei Zhang. Progressive source-aware transformer for generalized source-free domain adaptation. *IEEE Transactions on Multimedia*, 2023. 7, 8, 9

[40] Song Tang, Wenxin Su, Mao Ye, Jianwei Zhang, and Xiatian Zhu. Unified Source-Free Domain Adaptation. *arXiv preprint arXiv:2403.07601*, 2024. 8

[41] Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-Free Domain Adaptation with Frozen Multimodal Foundation Model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 7, 8, 9, 10

[42] Jingjing Wang, Ke Chen, and Shilei Yang. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. *ACM Multimedia*, 2021. 3

[43] Thomas Westfechtel, Dexuan Zhang, and Tatsuya Harada. Combining inherent knowledge of vision-language models with unsupervised domain adaptation through self-knowledge distillation. *arXiv preprint arXiv:2312.04066*, 2023. 3

[44] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. One teacher is enough? pre-trained language model distillation from multiple teachers. *arXiv preprint arXiv:2106.01023*, 2021. 3

[45] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *NeurIPS*, 2021. 8, 9

[46] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8978–8987, 2021. 3

[47] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *NeurIPS*, 2022. 8, 9

[48] Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *ICLR*, 2023. 8, 9

[49] Zhen Zhang, Lin Xu, and Yi Yang. Triple-Class GAN for Unsupervised Domain Adaptation. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[50] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 2024. 3