Towards Dynamic-Prompting Collaboration for Source-Free Domain Adaptation

Mengmeng Zhan , Zongqian Wu , Rongyao Hu , Ping Hu , Heng Tao Shen , Xiaofeng Zhu*
School of Computer Science and Engineering,
University of Electronic Science and Technology of China

Abstract

In domain adaptation, challenges such as data privacy constraints can impede access to source data, catalyzing the development of source-free domain adaptation (SFDA) methods. However, current approaches heavily rely on models trained on source data, posing the risk of overfitting and suboptimal generalization.This paper introduces a dynamic prompt learning paradigm that harnesses the power of large-scale vision-language models to enhance the semantic transfer of source models. Specifically, our approach fosters robust and adaptive collaboration between the source-trained model and the vision-language model, facilitating the reliable extraction of domain-specific information from unlabeled target data, while consolidating domain-invariant knowledge. Without the need for accessing source data, our method amalgamates the strengths inherent in both traditional SFDA approaches and vision-language models, formulating a collaborative framework for addressing SFDA challenges. Extensive experiments conducted on three benchmark datasets showcase the superiority of our framework over previous SOTA methods.

1 Introduction

Domain adaptation (DA) [Wang *et al.*, 2022b; Nejjar *et al.*, 2023] refers to the process of adapting a model trained on a labeled source domain to an unlabeled target domain. Traditional methods for this task assume access to labeled source data during the adaptation process. However, in the real world, practical constraints such as privacy and security often limit our ability to access source data directly. Consequently, the field of source-free domain adaptation (SFDA) [Roy *et al.*, 2022; Tang *et al.*, 2023] has gained significant attention. SFDA aims to adapt a source-trained model to an unlabeled target domain without the need for direct access to source data.

Many SFDA methods prioritize transferring domain-invariant knowledge from source-trained models to the target domain by iteratively fine-tuning models through self-supervised techniques, such as the use of pseudo-labels. In an ideal scenario, when the source and target domains exhibit a high degree of similarity, the source-trained model can efficiently capture domain-specific information from the target data and learn accurate classification boundaries. However, practical situations often deviate from this ideal, and the target data may significantly differ from the source hypothesis. This can lead to unreliable pseudo-labels and introduce bias into the learning process. Existing efforts aim to ameliorate these issues by employing techniques such as local structural adjustments [Qu *et al.*, 2022], entropy-based strategies [Litrico *et al.*, 2023], and historical consistency measures [Huang *et al.*, 2021]. Nevertheless, the inherent domain discrepancy can persist, resulting in an accumulation of irreparable bias and a loss of domain-invariant information throughout training [Fang *et al.*, 2022; Yu *et al.*, 2023].

Recently, large-scale pre-trained vision-language models, such as CLIP [Radford *et al.*, 2021], have demonstrated exceptional generalization capabilities with invariant feature encoding across diverse visual domains. However, foundational models like CLIP struggle to effectively represent domain-specific knowledge, a crucial factor for achieving success in downstream tasks [Jia *et al.*, 2022]. To address this limitation, prompt learning has emerged as a solution, involving the encoding of domain-related context as learnable parameters

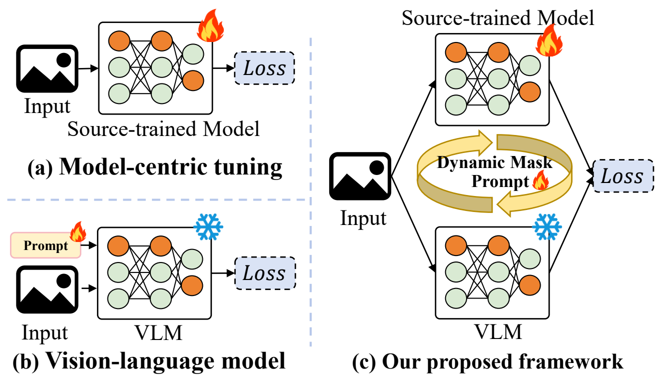

Figure 1: Conceptual comparison between (a) traditional model-centric tuning method, (b) large vision-language model, and (c) our proposed framework. Our method amalgamates the strengths of both paradigms to achieve effective source-free domain adaptation.

*Corresponding authors (seanzhuxf@gmail.com).

(referred to as prompts) at the input end [Zhou *et al.*, 2022b]. This approach also offers a viable solution to challenges like domain adaptation [Fahes *et al.*, 2023]. Nevertheless, optimizing these learnable prompts often demands careful data curation and annotation, which can be difficult SFDA, where labeled data is scarce, and extracting domain-specific information poses significant challenges.

Motivated by the aforementioned insights, our work aims to tackle SFDA by complementing the traditional methodology with a large vision-language model. Our goal is to extract domain-specific information from unlabeled target data reliably while consolidating the domain-invariant knowledge embedded within the source model. However, the integration of these two distinct modeling paradigms into a collaborative SFDA framework presents significant challenges. The first challenge is how to facilitate effective interaction between these two models. A well-designed integration of both paradigms can result in mutual reinforcement and improved performance. Conversely, a suboptimal solution may lead to undesired outcomes. The second challenge is how to ensure reliable learning from pseudo-labels. As the pseudo-labels in SFDA can be susceptible to noise and errors.

To address these challenges, we propose a collaborative framework centered around an innovative dynamic prompting method. As illustrated in Figure 1 (c), our approach fosters collaboration between a source-trained model and a vision-language model, such as CLIP, not only during the later stages of loss functions but also through joint prompting mechanisms at earlier feature and input stages. This unique design helps reconcile the disparities between the two distinct paradigms and promotes the exchange of complementary information within each component. Furthermore, to mitigate the disruption caused by unreliable pseudo-labels, we introduce a dynamic mask prompting (DMP) mechanism. When presented with a pseudo-labeled image, DMP is trained to analyze the corresponding class activation maps to identify noisy and distracting elements within the image. These identified pixels are then suppressed by replacing them with visual prompts. Consequently, subsequent model tuning can focus more on task-related information and be less susceptible to noise and background interference. We summarize the contributions of this work as follows:

- We present a collaborative framework in which a traditionally source-trained model dynamically collaborates with large vision-language models to address the SFDA task.
- We recognize the adverse effects of incorrect pseudo-labels in SFDA scenarios and propose a dynamic mask prompting mechanism to enhance the learning process.
- We conduct extensive experiments on multiple datasets, and the results show that our method outperforms previous state-of-the-arts in SFDA.

2 Related Works

2.1 Source-Free Domain Adaptation

Previous SFDA methods can be divided into two directions: model-centric and data-centric. Model-centric methods assume the optimal target hypothesis to be closely related to the source hypothesis. Therefore, by exploring the outputs of the source training model, models can be fine-tuned using a self-training scheme [Liang *et al.*, 2020]. BMD [Qu *et al.*, 2022] proposed a dynamic multi-centre pseudo-labelling strategy for updating pseudo-labels during domain adaptation. JMDS [Lee *et al.*, 2022] proposed to mitigate the effect of noisy samples by using confidence scores as sample weights. With the success of Transformer in vision tasks, DSiT [Sanyal *et al.*, 2023] proposed a fully ViT-based solution for SFDA and achieved outstanding performance. Though achieving promising performance, this type of method relies on task-specific knowledge learned on the source domain, which can result in biased pseudo-labels and potentially suffer from the loss of domain-invariant information in the progressive learning process [Zhang *et al.*, 2023a].

Data-centric models aim to enhance the pre-trained task-specific knowledge by reconstructing data for the missing source domain, based on which existing DA methods can be easily extended to SFDA scenarios [Li *et al.*, 2020]. However, with only the unlabelled target data, it is challenging to effectively represent the task-specific information revealed by the source data [Chen *et al.*, 2022; Tang *et al.*, 2020]. Recently, large-scale pre-trained vision-language models have become a popular paradigm for transferring pre-trained models into downstream tasks. In domain adaptation *i.e.*, AutoLabel [Zara *et al.*, 2023] utilized CLIP to discover candidate target categories, enhancing the capture of discriminative information for video domain adaptation. PADCLIP [Lai *et al.*, 2023] introduced adaptive biasing learning to address the issue of noisy labels in domain adaptation. However, the lack of target annotations limits their application in SFDA.

2.2 Prompt Learning

Prompt learning is becoming a popular way to enhance large-scale pretrained models with domain-specific knowledge [Zhou *et al.*, 2022b]. Hard prompt learning employs discrete tokens in the vocabulary as prompts, while soft prompt learning introduces learnable parameters in the text embedding and tunes these parameters [Jia *et al.*, 2022; Jiang *et al.*, 2024]. To address vision tasks, CoOp [Zhou *et al.*, 2022b] designed class-specific prompts through back-propagation. DAPL [Ge *et al.*, 2023] proposed ad-hoc prompting to learn disentangled domain and category representations. ADCLIP [Singha *et al.*, 2023] introduced a domain-agnostic prompt learning strategy for CLIP and achieved state-of-the-art results. Although these methods showing promising performance in domain adaptation, prompt learning with large vision language models requires labeled training data, and the limited capacity of prompts can be difficult to model complex domain-specific knowledge of target domains.

3 Methodology

Problem Definition. The task of SFDA involves a labeled source dataset $\mathbf{D}^s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ and an unlabeled target dataset $\mathbf{D}^t = \{x_i^t\}_{i=1}^{n_t}$, where x and y denote the image and label, respectively. Typically, $\mathbf{D}^s$ and $\mathbf{D}^t$ follow distinct distributions, and $\mathbf{D}^s$ is only available during the pre-training

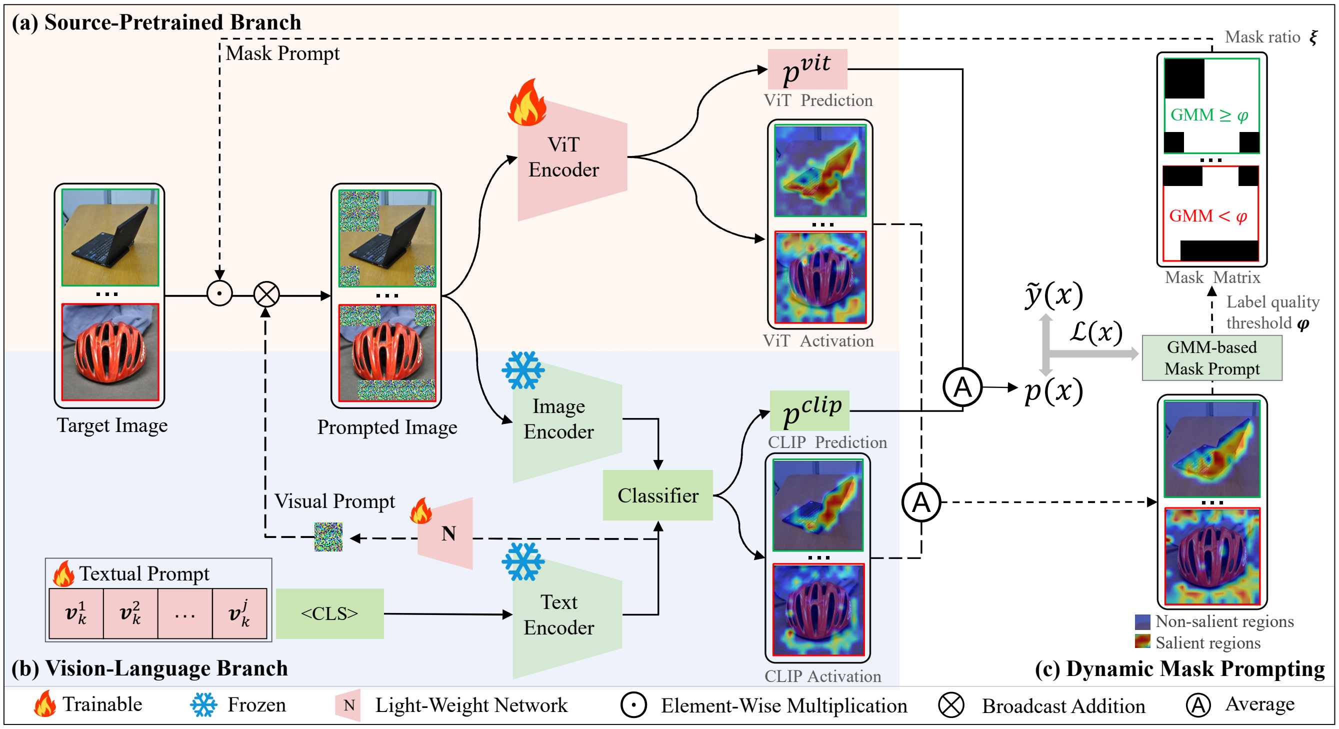

Figure 2: The framework of our proposed model comprises three main modules: (a) Source-Pretrained Branch (light yellow block) explores domain-specific knowledge in the target data. (b) Vision-Language Branch (light blue block) leverages a large-scale Visual Language Model (VLM) to extract domain-invariant task knowledge. (c) Dynamic Mask Prompting (DMP) customizes dynamic visual mask prompts for each image to suppress noisy visual content. During training, we begin by averaging the predictions (p^{vit} and p^{clip}) from both branches to generate the pseudo-label $\tilde{y}$ and compute the loss. The quality of the pseudo-label is assessed using a Gaussian Mixture Model (GMM) applied to the loss values. Subsequently, the DMP module analyzes activation maps corresponding to pseudo-labels, identifying and masking out distracting image patches. The masked areas are then filled with visual prompts. Finally, the prompted images and pseudo-labels are used to train the target model. In the testing phase, the outputs of both branches are averaged to generate the final predictions.

stage of the source-trained model f_θ. The primary objective of SFDA is to adapt the model f_θ to the target domain using solely $\mathbf{D}^t$. This paper mainly focuses on the general SFDA setting, wherein both domains share the same set of K classes (closed-set setting). Nevertheless, in our experiments, we also explore partial and open-set scenarios.

Overview. An overview of our method is depicted in Figure 2. The framework comprises three key components: the source-pretrained branch, the vision-language branch, and the dynamic mask prompting mechanism. In the source-pretrained branch, we adhere to the traditional model-centric SFDA approach, progressively fine-tuning the entire model based on pseudo-labels to encode domain-specific knowledge from the target domain. The vision-language branch leverages the CLIP model and employs learnable textual prompts to extract task-related domain-invariant information. The dynamic mask prompting module plays a pivotal role in our method. It identifies and enhances the reliability of pseudo-labels by introducing dynamic visual prompts. This module facilitates collaboration between the two branches, enhancing the overall performance of our approach. This comprehensive design allows our method to effectively acquire domain-related knowledge from the target data while ensuring the preservation of domain-shared information related to classification tasks. In the following, we will provide detailed descriptions of the source-model-based protocol, introduce the vision-language model-based paradigm, and propose the dynamic mask prompt mechanism.

3.1 Source-Pretrained Branch

The source-pretrained branch is primarily dedicated to adjusting model parameters to align with the target data, aiming to learn domain-specific knowledge relevant to the target domain. Given the absence of annotations for the target data, we employ a common self-training strategy used in SFDA methods. We start with an initial classification model f_θ (*e.g.*, ViT [Dosovitskiy *et al.*, 2020]) and tuning it on the target dataset $\mathbf{D}^t$ by optimizing the cross-entropy loss on the target image x^t:

$$\mathcal{L}_m = -\tilde{y}^t \log p_m\left(\hat{x}^t\right), \tag{1}$$

where $\tilde{y}^t$ represents the pseudo-label generated based on the average of the predictions from both branches, $p_m(\cdot)$ denotes the predicted probability of the target model, and $\hat{x}^t$ refers to the prompted input image, which will be comprehensively introduced in Section 3.3.

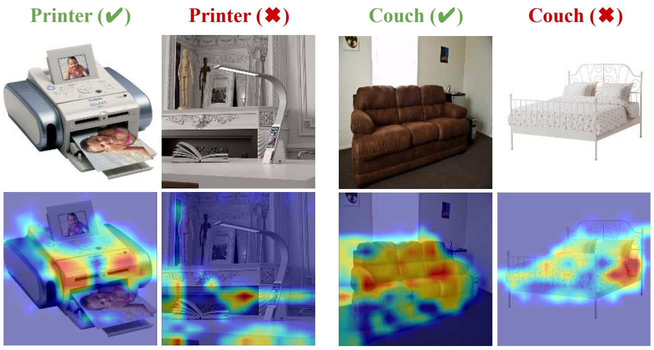

Figure 3: Visualization of activation maps. The first row shows the pseudo-label for each image, with green indicating correct pseudo-labels and red representing incorrect pseudo-labels. The second row displays the image and the third row showcases the activation maps corresponding to the pseudo-labels. As observed, incorrect pseudo-labels can misguide the model focus on unrelated or confusing image regions.

3.2 Vision-Language Branch

We leverage the exceptional generalization capabilities of CLIP [Radford *et al.*, 2021] to extract domain-invariant task information (*i.e.*, category semantics) from the target data. CLIP consists of a vision encoder f_v and a text encoder f_t. The vision encoder f_v transforms the image x^t into a visual representation $f_v(x^t) \in \mathbb{R}^d$. Simultaneously, a set of class embedding $f_t(w_k) \in \mathbb{R}^d$ is generated by feeding template prompts w_k (*e.g.*, "A photo of a $[CLS]$") into text encoder f_t. Hence, the classification probability of x^t is defined as:

$$p_d\left(x^t\right) = \frac{\exp(\mathrm{sim}(f_v(x^t), f_t(w_k))/\tau)}{\sum_{i=1}^{c} \exp(\mathrm{sim}(f_v(x^t), f_t(w_i))/\tau)}, \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ represents cosine similarity, and τ is a temperature factor. Following CoOp [Zhou *et al.*, 2022b], we convert the class name k into a text embedding $w_k \in \mathbb{R}^d$ and enhance this embedding with a set of learnable prompts $\{v_i\}_{i=1}^{T}$. Formally, the input for the text encoder becomes $V_k = \{v_1, v_2, \ldots, v_T, w_k\}$. By normalizing the representations for each class k, we can calculate the classification probability using Eq. (2). During training, we optimize the learnable parameters on the target data by minimizing the cross-entropy loss:

$$\mathcal{L}_d = -\tilde{y}^t \log p_d\left(\hat{x}^t\right), \tag{3}$$

where $\tilde{y}^t$ denotes the pseudo-label of image x^t, generated based on the average of the predictions from the source-trained model and CLIP model. $\hat{x}^t$ represents the prompted input image, which will be introduced in the next section.

3.3 Dynamic Mask Prompting

Due to the domain discrepancy, pseudo-labels can be prone to limitations such as noise and errors. Learning directly from such pseudo supervision may result in suboptimal performance and limited generalization on the target domain. To ensure the reliable transfer of domain-shared knowledge from the source model to the unlabeled target data, we start by analyzing the visual content that supports these pseudo-labels. As depicted in Figure 3, effective pseudo-labels correspond to relevant regions within the image, whereas incorrect pseudo-labels often get distracted by unrelated or confusing areas. This observation motivates us to enhance the quality of learning by suppressing distracting responses in the images. We introduce a novel mechanism called Dynamic-Mask Prompting (DMP), as illustrated in Figure 2 (c). In DMP, we initially obtain a pseudo-label for an image based on predictions from both the source-pretrained branch and the vision-language branch. Subsequently, we generate corresponding class activation maps, which are guided by a Gaussian Mixture Model (GMM) [Permuter *et al.*, 2006] to estimate regions with a negative impact. These negative regions are then masked out from the input image. The masked pixels are replaced with visual prompts, further enhancing the interaction between the two branches.

Mask Prompt Generation. To identify image regions that negatively impact learning, we begin by analyzing the corresponding class activation maps of pseudo-labels. Given the input image x^t and the pseudo-label, we employ Grad-CAM [Selvaraju *et al.*, 2017] to generate the activation map for the model in the source-pretrained branch. Specifically, we first obtain feature tokens $\mathbf{F}_v \subseteq \mathbb{R}^{b\times b\times d}$, where b represents the patch size (*e.g.*, 16). Then, we compute the activation map as follows:

$$\mathbf{A}_{vit} = \mathrm{GradCAM}(\mathbf{F}_v, \frac{\partial \tilde{y}}{\partial \mathbf{F}_v}), \tag{4}$$

where $\mathrm{GradCAM}(\cdot)$ is the activation map generation function [Selvaraju *et al.*, 2017].

For CLIP, we utilize the similarity between the text encoding of the pseudo-label and the visual features of the image to obtain the class attention map. Specifically, we compute the similarity between the features of image tokens $\mathbf{F}_i \subseteq \mathbb{R}^{b\times b\times d}$ and the features of texts $\mathbf{F}_t \subseteq \mathbb{R}^{b\times b\times d}$ with l_2 normalization along the feature channel dimension, as shown in the equation:

$$\mathbf{A}_{clip} = \mathrm{norm}\left(\frac{\mathbf{F}_i}{\|\mathbf{F}_i\|_2} \cdot \left(\frac{\mathbf{F}_t}{\|\mathbf{F}_t\|_2}\right)^{\mathbb{T}}\right). \tag{5}$$

Then, we combine both activation maps by averaging,

$$\mathbf{A}_u = \frac{(\mathbf{A}_{vit} + \mathbf{A}_{clip})}{2}. \tag{6}$$

As a result, the activation map reflects the correlation between pixel locations and the pseudo-label, revealing how the visual contents will be activated during learning. When the pseudo-label is correct, such activations assist the model in concentrating on the most relevant regions. However, in cases of an incorrect pseudo-label, the high-response areas can divert the model's attention away from informative content, resulting in less effective knowledge transfer.

The absence of annotations poses challenges in assessing the quality of pseudo-labels. Guided by our empirical observation that correct pseudo-labels yield lower loss values and vice versa, we train a Gaussian Mixture Model (GMM) that takes the loss values as input. By applying a thresholding strategy with a hyperparameter φ we can roughly distinguish between correct and incorrect pseudo-labels. Based on the prediction scores, we devise an adaptive strategy to suppress the most irrelevant visual content during model tuning. For samples predicted as incorrect, we mask out the ξ highest activated pixel locations based on their activation map $\mathbf{A}_u$. Conversely, for samples predicted as correct, we mask out the ξ lowest activated pixel locations, as these regions are most unrelated to the task.

Mask Prompt Padding. After obtaining the masked images, simply filling the masked patches with constant values and passing them into an image encoder can lead to undesired effects during model training. This approach neglects the semantic information on the textual side, which is essential for effective learning. To address this, we introduce a dynamic visual prompting strategy that interacts with the semantic information from the textual side. Specifically, we define a set of learnable visual prompts $\mathbf{g}$ while simultaneously extracting domain-shared task-specific information from the text encoder through a lightweight MLP network:

$$\hat{\mathbf{g}} = \mathbf{g} + \mathrm{h}\left(f_t\left(w_k\right)\right), \tag{7}$$

where $\mathrm{h}(\cdot)$ represent the lightweight MLP network, $\hat{\mathbf{g}}$ is the visual prompt. The masked pixels are then filled with the visual prompt $\hat{\mathbf{g}}$. These prompted images are subsequently fed into the image encoder for pseudo-label-based self-training.

3.4 Optimization

Throughout the training process, we jointly optimize the textual prompts, visual prompts, and the parameters θ of the source-trained model within a collaborative framework. The pseudo-labels are derived from the averaged predictions of both branches. Additionally, to prevent the model from converging to a trivial solution where it predicts all data as a specific class, we incorporate the prediction diversity loss $\mathcal{L}_{div}$ into our framework:

$$\mathcal{L}_{div} = \textstyle\sum_{k=1}^{K} \mathrm{KL}\left(\bar{p}_f^k\left(\hat{x}^t\right) || \, q\right), \tag{8}$$

where $\bar{p}_f^k\left(\hat{x}^t\right)$ represents the empirical label distribution of the k-th class, and q is the uniform distribution defined as $q = 1/K$, where K represents the number of classes. $\mathrm{KL}\left(\cdot \,||\, \cdot\right)$ denotes the Kullback–Leibler divergence. Hence, together with Eqs.(1) and (3), the overall loss is presented as:

$$\mathcal{L} = \mathcal{L}_m + \mu_1 \mathcal{L}_d + \mu_2 \mathcal{L}_{div}, \tag{9}$$

where μ_1 and μ_2 are non-negative parameter weights. Eq. (9) simultaneously collaborates with two independent modelling paradigms to achieve optimal performance in SFDA.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed approach on three standard benchmarks in domain adaptation including Office-31 [Saenko *et al.*, 2010], Office-Home [Venkateswara *et al.*, 2017], and DomainNet [Peng *et al.*, 2019]. Office-31 contains 4,652 images for 31 classes collected from three distinct domains, which are Amazon (A), DSLR (D), and Webcam (W). Office-Home is a medium-scale dataset consisting of images for everyday objects, and can be divided into four domains, *i.e.*, Artistic (Ar), ClipArt (Cl), Product (Pr), and Real-world (Rw), with 65 classes for each. DomainNet is a large-scale benchmark with six domains, each with 345 classes. Following the setting of [Litrico *et al.*, 2023], we select four domains: ClipArt (C), Real (R), Painting (P), and Sketch (S), with 126 classes as the SFDA benchmark. All reported results are the average of three independent runs.

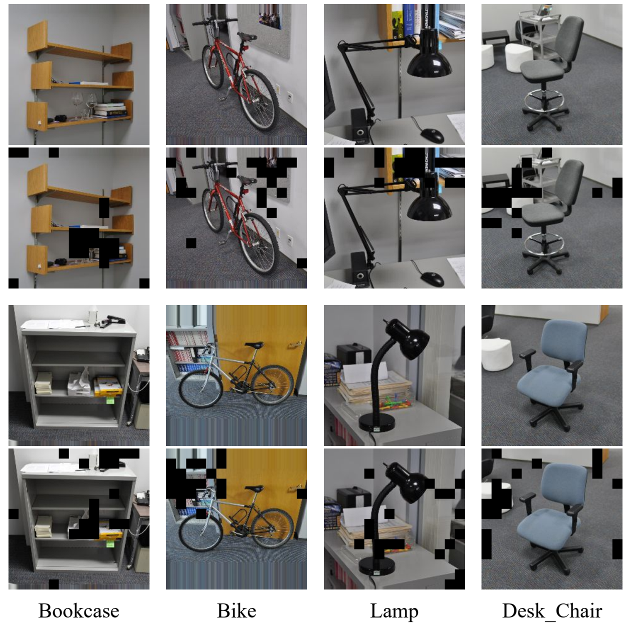

Figure 4: Visualization of masked patches in our method. Black squares indicate masked patches. As shown, the unrelated or distracting visual contents are successfully suppressed.

Implementation Details. As introduced in the methodology, our proposed method contains two network branches. In the Source-Pretrained Branch, we follow the experimental settings of [Sanyal *et al.*, 2023] and utilize ViT-B/16 (input size 224 × 224, patch size 16×16, resulting in 14×14 patches per input). For Vision-Language Branch, we adopt the experimental settings of CoCoOp [Zhou *et al.*, 2022a] and use CLIP based on ViT-B/16 as the vision encoder. We optimize training objectives via the Stochastic Gradient Descent (SGD) [Zinkevich *et al.*, 2010] optimizer, given a mini-batch size of 16, the momentum of 0.9, and weight decay ratio of 1×10^{-4}, respectively.

4.2 Experimental Results

Source-Free Domain Adaptation. Tables 1, 2, and 3 report the performance of our method and previous approaches on SFDA. As indicated in the tables, we compare with traditional CNN-based methods, Transformer-based methods, and vision-language-based models. Results demonstrate that our source-free method outperforms the previous state-of-

Methods	SF	VLM	Office-Home												
			Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
TVT† [Yang *et al.*, 2023]	✗	✗	74.8	86.8	89.4	82.7	87.9	88.2	79.8	71.9	90.1	85.4	74.6	90.5	83.5
DAPL [Ge *et al.*, 2023]	✗	✓	70.6	90.2	91.0	84.9	89.2	90.9	84.8	70.5	90.6	84.8	70.1	90.8	84.0
ADCLIP [Singha *et al.*, 2023]	✗	✓	70.9	92.5	92.1	85.4	92.4	92.5	86.7	74.3	93.0	86.9	72.6	93.8	86.1
SHOT [Liang *et al.*, 2020]	✓	✗	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
NRC [Yang *et al.*, 2021]	✓	✗	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
AaD [Yang *et al.*, 2022]	✓	✗	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
JMDS [Lee *et al.*, 2022]	✓	✗	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
CRS [Zhang *et al.*, 2023b]	✓	✗	63.5	82.1	85.0	73.0	82.7	82.4	69.5	62.9	82.6	74.2	65.7	87.3	75.9
SHOT† [Liang *et al.*, 2020]	✓	✗	67.1	83.5	85.5	76.6	83.4	83.7	76.3	65.3	85.3	80.4	66.7	83.4	78.1
DIPE† [Wang *et al.*, 2022a]	✓	✗	66.0	80.6	85.6	77.1	83.5	83.4	75.3	63.3	85.1	81.6	67.7	89.6	78.2
DSiT† [Sanyal *et al.*, 2023]	✓	✗	69.2	83.5	87.3	80.7	86.1	86.2	77.9	67.9	86.6	82.4	68.3	89.8	80.5
Ours	✓	✓	**71.1**	**87.1**	**91.3**	**86.3**	**90.9**	**91.6**	**86.6**	**74.1**	**91.8**	**87.6**	**75.0**	**91.8**	**85.4**

Table 1: Source-Free Domain Adaptation (SFDA) on Office-Home benchmark. “SF” and “VLM” indicate source-free adaptation and visual language model. “†” indicates Transformer-based methods. Bold numbers indicate the best results among SFDA methods.

Method	SF	Office-31						
		A→D	A→W	D→A	D→W	W→A	W→D	Avg.
CDTrans† [Xu *et al.*, 2021]	✗	97.0	96.7	81.1	99.0	81.9	100.0	92.6
TVT† [Yang *et al.*, 2023]	✗	96.4	96.4	84.9	99.4	86.1	100.0	93.8
SHOT [Liang *et al.*, 2020]	✓	94.0	90.1	74.7	98.4	74.3	99.9	88.6
NRC[Yang *et al.*, 2021]	✓	96.0	90.8	75.3	99.0	75.0	**100.0**	89.4
AaD [Yang *et al.*, 2022]	✓	96.4	92.1	75.0	**99.1**	76.5	**100.0**	89.9
JMDS [Lee *et al.*, 2022]	✓	94.4	95.2	76.2	98.5	77.6	**100.0**	90.3
CRS [Zhang *et al.*, 2023b]	✓	96.6	95.5	76.9	**99.1**	78.3	**100.0**	91.1
SHOT† [Liang *et al.*, 2020]	✓	95.3	94.3	79.4	99.0	80.2	**100.0**	91.4
DIPE† [Wang *et al.*, 2022a]	✓	94.8	95.5	77.5	98.5	77.1	**100.0**	90.5
DSiT† [Sanyal *et al.*, 2023]	✓	**98.0**	**97.2**	81.7	**99.1**	81.8	**100.0**	93.0
Ours	✓	96.9	97.0	**83.9**	98.2	**83.7**	**100.0**	**93.3**

Table 2: Source-Free Domain Adaptation (SFDA) on Office-31 benchmark. “SF” indicate source-free adaptation. “†” indicates Transformer-based methods. Bold numbers indicate the best results among SFDA methods.

the-art SFDA methods, and achieves similar performance to the source-dependent models. Specifically, it surpassed the prior best method $\mathrm{DSiT}^{\dagger}$ by 0.3% Office-31, by 4.9% Office-Home, and method $\mathrm{AaD}^{\dagger}$ by 12.9% on DomainNet, thereby establishing a new state-of-the-art for SFDA. Besides, Compared to source-dependent DA approaches, our proposed method is equally competitive. Specifically, our method improves over the VLM-based DAPL by 1.4% on Office-Home. The reasons can be summarized as follows. First, our proposed approach unites the traditional source-pretrained model and the vision-language model in a collaborative framework, complementing to enhance their advantages and address their shortcomings. The domain-specific knowledge encoded in the tuned source pre-trained model helps to unleash from VLM the reliable domain-invariant task information as well as the generalization ability to fill the domain gap. Moreover, benefiting from the proposed dynamic mask prompt learning, our proposed approach mitigates the noise and erroneous in pseudo-label-based self-training.

Open-set/Partial-set SFDA. We also evaluate our method with more practical settings like open-set and partial-set scenarios. In the open-set setting, the target domain contains some classes that are agnostic in the source domain. In contrast, in the partial-set scenario, the target domain contains only some of the classes in the source domain. We compare the performance for the open-set and partial-set settings in Table 4. In both scenarios, the proposed method achieves the best performance. Compared with the previous best method CRS, ours improves by 5.2% on average. This shows the superior performance of our methods. In addition, compared to the sub-optimal method AaD, our method improves by 6.35% on average, which demonstrates the ability to better enhance the generalization capacity in vision-language models.

Ablation Analysis. In this part, we analyze the impact of different components in the proposed framework and report the ablation results in Table 5. Specifically, “SPB” and “VLB” denote applications of the Source-Pretrained Branch and the Vision-Language Branch, respectively. “DMP” indicates the dynamic mask prompt learning. As shown in the Table 5, each component makes a unique contribution to the final performance. We observe that the SPB-only baseline outperforms the VLB-only baseline on easier datasets (*e.g.*, Office-31), and underperforms on more complex datasets (*e.g.*, Office-Home, DomainNet). This may be because the VLB model is less effective in encoding domain-specific knowledge under small domain gaps, whereas the SPB model struggles to extract domain-invariant task information under large domain discrepancies. Thus, both separate branches have their inherent disadvantages. The fusion of VLB and SPB allows for the integration of their respective strengths. As shown, the simple combination “SPB +VLB” is 1.2% better than “SPB” on Office-31 and 1.7% better than “VLB” on Office-Home. By further encouraging the interaction between the two paradigms, our dynamic mask prompting mechanism leads to a more adaptive and robust model, thereby enhancing performance across datasets of varying scales and complexities, *e.g.*, boosting the performance by

Method	SF	VLM	DomainNet												
			Cl→Pn	Cl→Rl	Cl→Sk	Pn→Cl	Pn→Rl	Pn→Sk	Rl→Cl	Rl→Pn	Rl→Sk	Sk→Cl	Sk→Pn	Sk→Rl	Avg.
DAPL [Ge *et al.*, 2023]	✗	✓	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
ADCLIP [Singha *et al.*, 2023]	✗	✓	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
SHOT [Liang *et al.*, 2020]	✓	✗	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 [Yang *et al.*, 2021]	✓	✗	62.6	77.1	58.3	62.9	81.3	60.7	64.7	69.4	58.7	69.4	65.8	78.7	67.5
AdaCon [Chen *et al.*, 2022]	✓	✗	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
JMDS [Lee *et al.*, 2022]	✓	✗	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
SHOT† [Liang *et al.*, 2020]	✓	✗	64.8	82.3	63.1	68.9	84.0	62.7	72.3	70.6	61.7	74.0	69.2	83.6	71.4
AaD† [Yang *et al.*, 2022]	✓	✗	66.8	81.0	63.8	70.4	84.0	65.4	74.6	72.1	63.8	76.4	71.2	82.8	72.7
Ours	✓	✓	**82.5**	**89.6**	**82.1**	**89.7**	**91.2**	**80.9**	**86.1**	**82.9**	**81.4**	**87.2**	**84.8**	**89.2**	**85.6**

Table 3: Source-Free Domain Adaptation (SFDA) on Domain-Net benchmark. "SF" and "VLM" indicate source-free adaptation and visual language model. "†" indicates Transformer-based methods. Bold numbers indicate the best results among SFDA methods.

Partial-set DA	Avg.	Open-set DA	Avg.
SHOT [Liang *et al.*, 2020]	79.3	SHOT [Liang *et al.*, 2020]	72.8
HCL [Huang *et al.*, 2021]	79.6	HCL [Huang *et al.*, 2021]	72.6
JMDS [Lee *et al.*, 2022]	83.2	CoWA [Lee *et al.*, 2022]	73.2
AaD [Yang *et al.*, 2022]	79.7	AaD [Yang *et al.*, 2022]	71.8
CRS [Zhang *et al.*, 2023b]	80.6	CRS [Zhang *et al.*, 2023b]	73.2
Ours	**87.6**	Ours	**76.6**

Table 4: Partial-set SFDA and Open-set SFDA on Office-Home benchmarks. Bold numbers indicate the best results among SFDA methods.

SPB	VLB	DMP	Office-31	Office-Home	DomainNet	Avg.
✓	-	-	91.1	75.9	71.1	79.4
-	✓	-	81.7	82.4	82.3	82.1
✓	✓	-	92.3	84.1	83.9	86.8
✓	-	✓	91.6	77.6	72.4	80.5
-	✓	✓	82.3	83.0	82.7	82.7
✓	✓	✓	**93.3**	**85.4**	**85.6**	**88.8**

Table 5: Ablation study of different components of the proposed method. "SPB" and "VLB" denote the source-pretrained branch tuning and the vision-language branch, respectively.

1% and 1.3% on Office-31 and Office-Home, respectively, compared to the simple fusion strategy. The effectiveness of applying dynamic mask prompt learning to individual or both branches is confirmed. In Figure 4, we also visualize the masks generated in the dynamic mask prompting component. As shown, the background and distracting content are successively suppressed, hence enabling the model to better learn from the target data.

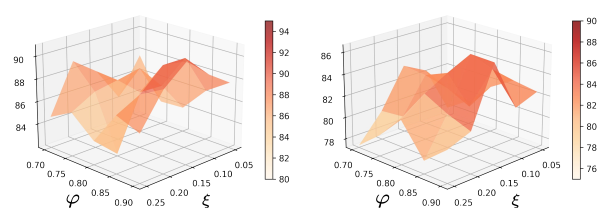

Figure 5: Classification results of our method under different parameter settings (*i.e.*, φ , ξ) for different transfer tasks in Office-Home.

Hyper-parameter Analysis. We investigate the impact of hyper-parameters in our proposed method, *i.e.*, the threshold φ for GMM and the percentage ξ for selecting on activation maps. We conduct the classification task on Office-Home (*i.e.*, Cl-Ar, Cl-Pr) by varying the value of φ in the range of [0.7,0.9] and the value of ξ in the range of [0.05,0.25]. The results are illustrated in Figure 5. The accuracy decreases when the value of φ becomes small (*i.e.*, 0.7). This is because the initial pseudo-labels of the model are of good quality, and when setting at a lower φ, it is easy to confuse the correct labels leading to poor training results. If the value of ξ is too large or small, it is easy to get poor results. Too many mask patches may lose discriminative information, while too few mask patches will contain noisy information. This also shows the effectiveness of our proposed dynamic masked prompt.

5 Conclusion

In this paper, we present a collaborative framework for source-free domain adaptation, which exploits the inherent advantages of traditional model-centric protocol and large-scale vision-language model to complement each other. To enhance the interaction between the two paradigms, we further introduce a novel dynamic mask prompting mechanism that adaptively suppresses noisy and distracting visual content during training. Extensive experiments have been conducted to analyze the proposed approach, and the results demonstrate that our method outperforms previous state-of-the-art of source-free domain adaptation.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFA1004100, the National Natural Science Foundation of China under Grant 62306066, and the Sichuan Province Innovative Talent Funding Project for Postdoctoral Fellows under Grant BX202313.

References

[Chen *et al.*, 2022] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, pages 295–305, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fahes *et al.*, 2023] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Poda: Prompt-driven zero-shot domain adaptation. In *ICCV*, pages 18623–18633, 2023.

[Fang *et al.*, 2022] Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. Source-free unsupervised domain adaptation: A survey. *arXiv preprint arXiv:2301.00265*, 2022.

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

[Huang *et al.*, 2021] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *NeurIPS*, volume 34, pages 3635–3649, 2021.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.

[Jiang *et al.*, 2024] Xin Jiang, Hao Tang, Junyao Gao, Xiaoyu Du, Shengfeng He, and Zechao Li. Delving into multimodal prompting for fine-grained visual classification. In *AAAI*, pages 2570–2578, 2024.

[Lai *et al.*, 2023] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *ICCV*, pages 16155–16165, 2023.

[Lee *et al.*, 2022] Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *ICML*, pages 12365–12377, 2022.

[Li *et al.*, 2020] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, pages 9641–9650, 2020.

[Liang *et al.*, 2020] 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*, pages 6028–6039, 2020.

[Litrico *et al.*, 2023] Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *CVPR*, pages 7640–7650, 2023.

[Nejjar *et al.*, 2023] Ismail Nejjar, Qin Wang, and Olga Fink. Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *CVPR*, pages 11744–11754, 2023.

[Peng *et al.*, 2019] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[Permuter *et al.*, 2006] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4):695–706, 2006.

[Qu *et al.*, 2022] 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 *ECCV*, pages 165–182, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Roy *et al.*, 2022] Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *ECCV*, pages 537–555, 2022.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[Sanyal *et al.*, 2023] Sunandini Sanyal, Ashish Ramayee Asokan, Suvaansh Bhambri, Akshay Kulkarni, Jogendra Nath Kundu, and R Venkatesh Babu. Domain-specificity inducing transformers for source-free domain adaptation. In *ICCV*, pages 18928–18937, 2023.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[Singha *et al.*, 2023] Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *ICCV*, pages 4355–4364, 2023.

[Tang *et al.*, 2020] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *ACMMM*, pages 610–618, 2020.

[Tang *et al.*, 2023] Longxiang Tang, Kai Li, Chunming He, Yulun Zhang, and Xiu Li. Consistency regularization for generalizable source-free domain adaptation. In *ICCV*, pages 4323–4333, 2023.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[Wang *et al.*, 2022a] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *CVPR*, pages 7151–7160, 2022.

[Wang *et al.*, 2022b] Zengmao Wang, Chaoyang Zhou, Bo Du, and Fengxiang He. Self-paced supervision for multi-source domain adaptation. In *IJCAI*, 2022.

[Xu *et al.*, 2021] Tongkun Xu, Weihua Chen, WANG Pichao, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. In *ICLR*, 2021.

[Yang *et al.*, 2021] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *NeurIPS*, volume 34, pages 29393–29405, 2021.

[Yang *et al.*, 2022] Shiqi Yang, Shangling Jui, Joost van de Weijer, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *NeurIPS*, volume 35, pages 5802–5815, 2022.

[Yang *et al.*, 2023] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *ICCV*, pages 520–530, 2023.

[Yu *et al.*, 2023] Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *arXiv preprint arXiv:2302.11803*, 2023.

[Zara *et al.*, 2023] Giacomo Zara, Subhankar Roy, Paolo Rota, and Elisa Ricci. Autolabel: Clip-based framework for open-set video domain adaptation. In *CVPR*, pages 11504–11513, 2023.

[Zhang *et al.*, 2023a] Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Rethinking the role of pre-trained networks in source-free domain adaptation. In *ICCV*, pages 18841–18851, 2023.

[Zhang *et al.*, 2023b] Yixin Zhang, Zilei Wang, and Weinan He. Class relationship embedded learning for source-free unsupervised domain adaptation. In *CVPR*, pages 7619–7629, 2023.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

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

[Zinkevich *et al.*, 2010] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. In *NeurIPS*, volume 23, 2010.