COCA: Classifier-Oriented Calibration via Textual Prototype for Source-Free Universal Domain Adaptation

Xinghong Liu[1,2,3] and Yi Zhou*[1,2]

[1] School of Computer Science and Engineering, Southeast University, Nanjing Jiangsu 211189, China {xhoml158,yizhou.szcn}@gmail.com
[2] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3] Postal Savings Bank of China, Beijing 100032, China

Abstract. Universal domain adaptation (UniDA) aims to address domain and category shifts across data sources. Recently, due to more stringent data restrictions, researchers have introduced source-free UniDA (SF-UniDA). SF-UniDA methods eliminate the need for direct access to source samples when performing adaptation to the target domain. However, existing SF-UniDA methods still require an extensive quantity of labeled source samples to train a source model, resulting in significant labeling costs. To tackle this issue, we present a novel plug-and-play **C**lassifier-**O**riented **CA**libration (COCA) method. COCA, which exploits textual prototypes, is designed for the source models based on few-shot learning with vision-language models (VLMs). It endows the VLM-powered few-shot learners, which are built for closed-set classification, with the unknown-aware ability to distinguish common and unknown classes in the SF-UniDA scenario. Crucially, COCA is a new paradigm to tackle SF-UniDA challenges based on VLMs, which focuses on classifier instead of image encoder optimization. Experiments show that COCA outperforms state-of-the-art UniDA and SF-UniDA models. The code is available at https://github.com/XHomL/COCA.

Keywords: Source-Free universal domain adaptation · Transfer learning · Few-Shot learning.

1 Introduction

Training deep neural networks (DNNs) on custom datasets can achieve outstanding performance when the models are employed on i.i.d. datasets. However, the trained DNNs are likely to underperform if the test dataset (target domain) exhibits a large domain shift compared to the training dataset (source domain). To address the performance degradation in unlabeled target domains caused by domain shift, researchers have studied unsupervised domain adaptation (UDA) [12,41,43,16,33] methods. Vanilla UDA methods are designed for the

* Corresponding author

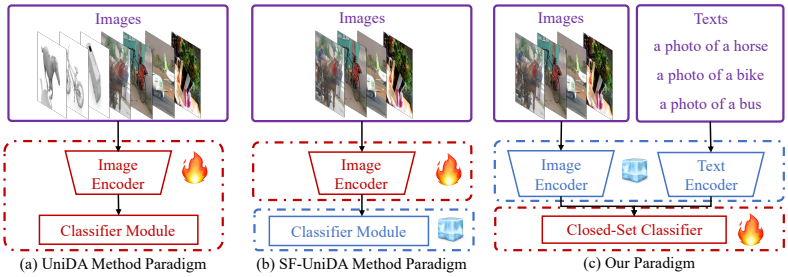
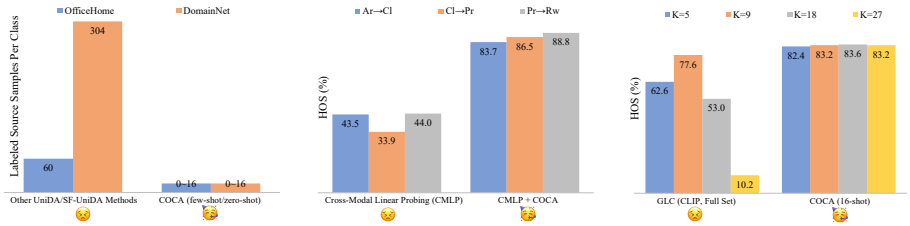


Fig. 1: (a) UniDA methods optimize both the image encoder and classifier module. (b) SF-UniDA methods freeze the classifier module and adapt the image encoder to the target domain. (c) We leverage and freeze image and text encoders while adapting the closed-set classifier to the target domain.

closed-set scenario. In real-world situations, category shift may arise, including open-set domain adaptation (OSDA) [36,25,45,17,26], partial domain adaptation (PDA) [14,23,4], and open-partial domain adaptation (OPDA) [6,19]. Thus, universal domain adaptation (UniDA) [34,20,35,18] is introduced to address such uncertain domain and category shifts, meaning it must handle arbitrary situations that may arise in OSDA, PDA, or OPDA. More recently, due to stricter data restrictions, source-free universal domain adaptation (SF-UniDA) [30] has been proposed. When performing SF-UniDA, it assumes that only the trained source model is available, rather than providing source samples. However, the existing SF-UniDA method [30] still requires training the source model using numerous labeled source samples, leading to substantial labeling costs.

Recently, the emergence of vision-language models (VLMs) such as CLIP [31], inspires us that labeling costs of source samples can be minimized by utilizing the strong representation ability of these models. The VLMs have become the new foundational engine for few-shot learning [42,44,13,24]. However, existing VLM-powered few-shot learners are primarily designed for closed-set scenarios and i.i.d. domain distribution. They experience substantial performance degradation when confronted with domain and category shifts. In this paper, we propose a plug-and-play method to calibrate the few-shot learning models, such as linear probe CLIP [31], CLIP Adapter [13], and cross-modal linear probing [24], to tackle the SF-UniDA challenge. Furthermore, we investigate the zero-shot learning problem within the context of SF-UniDA tasks. Specifically, we adapt zero-shot classifiers such as single linear layer [31] or the adapter module [13] to target domains. Traditional approaches in UniDA and SF-UniDA have predominantly concentrated on transferring specific knowledge from the source to the target domain. This is because traditional image encoders such as ResNet [15], which are pre-trained on the ImageNet [9] dataset, do not inherently encapsulate knowledge of both the source and target domains. We argue that the image and text encoders within the VLMs, having undergone pre-training on extensive image-text pair datasets, intrinsically encapsulate knowledge pertinent



(a) (b) (c)

Fig. 2: (a) Our method requires far fewer labeled source samples per class than traditional UniDA/SF-UniDA models. (b) Our plug-and-play method successfully adapts the VLM-powered few-shot learner [24] to new target domains. (c) COCA exhibits more robustness against variations in the hyperparameter K for K-means and outperforms the earlier SF-UniDA model GLC [30].

to both domains. As depicted in Fig. 1, in contrast to conventional UniDA or SF-UniDA methods that primarily concentrate on image encoder optimization to transfer source domain knowledge, we present a new paradigm focusing on classifier optimization to tackle the SF-UniDA challenge.

In order to adapt the closed-set classifier to new target domains, we present a novel plug-and-play classifier-oriented calibration (COCA) method, which exploits textual prototypes, to endow the VLM-powered few-shot learners with unknown-aware ability. To overcome domain and category shifts and enable the source model—specifically, the VLM-powered few-shot learner—to accurately distinguish common and unknown class samples, we introduce a novel autonomous calibration via textual prototype (ACTP) module. ACTP utilizes the K-means [27] clustering, along with image and text features, to implement self-training to adapt the close-set classifier to new unlabeled target domains. Furthermore, to encourage the closed-set classifier to exploit context information in images and enhance the mutual information, we propose the mutual information enhancement by context information (MIECI) module. Our experiments reveal that MIECI favorably influences model performance.

We illustrate our research motivation in Fig. 2a. For the SF-UniDA challenge, our approach can be applied to the few-shot/zero-shot learning problems, significantly reducing labeling costs of source samples compared to previous UniDA or SF-UniDA methods. As illustrated in Fig. 2b, COCA endows VLM-powered few-shot learners with the unknown-aware ability to accurately distinguish common and unknown class samples within target domains. Moreover, we have identified two issues: (1) The conventional UniDA and SF-UniDA methods [20,30], which utilize prototypes derived from image features—termed image prototypes, unavoidably incorporate domain information into these prototypes. This hinders the mutual information $I(\mathcal{X}, \mathcal{V})$ between the image set $\mathcal{X}$ and the prototype set $\mathcal{V}$ to be minimized. Consequently, the model performances are suboptimal. (2) The efficacy of the preceding model GLC [30], which employs image proto-

types, heavily relies on the empirical selection of appropriate hyperparameter K for K-means clustering. This is because the setting of K critically affects the quality of the image prototypes. To address the two issues, we propose a novel approach wherein text features are employed as prototypes—termed textual prototypes—to minimize the mutual information $I(\mathcal{X}, \mathcal{V})$, consequently enhancing the model performance. This approach, grounded in textual prototypes, guarantees more stable performance. Our approach performs better and is insensitive to variations in the hyperparameter K as shown in Fig. 2c.

We conduct experiments on three public benchmarks, each offering sufficient labeled source samples. Previous UniDA and SF-UniDA methods are fully trained using the source sample sets, but we only use few source shots to train the source model in our approach. The results indicate that COCA consistently surpasses state-of-the-art methods over all the benchmarks, even though our source model is trained on few source shots. To summarize, our contributions are highlighted as follows:

1. To the best of our knowledge, we are the first to explore few-shot and zero-shot learning problems in the UniDA/SF-UniDA scenario. Specifically, we propose a plug-and-play method for the VLM-powered few-shot learners to adapt them to new target domains and to endow them with unknown-aware ability. Crucially, we present a new paradigm, which focuses on classifier instead of image encoder optimization, to tackle the SF-UniDA challenge.
2. To overcome domain and category shifts and enable the closed-set source models to accurately distinguish common and unknown class samples, we present a novel autonomous calibration via textual prototype (ACTP) module. We introduce the mutual information enhancement by context information (MIECI) module to encourage the classifier to exploit context information in image features and enhance the mutual information.
3. Experiments conducted on three public benchmarks under various category-shift settings show the substantial superiority of our approach. Moreover, our study reveals that VLMs have encapsulated knowledge of both the source and target domains, enabling VLM-powered models to autonomously adapt to new target domains.

2 Related Work

Universal Domain Adaptation. Li *et al.* [20] defined the universal domain adaptation (UniDA) scenario measuring the robustness of a model to various category shifts. They leveraged domain consensus knowledge to enhance target clustering and discover private categories. OVANet [35] employs a one-vs-all classifier for each source class, determining the "known" or "unknown" class through its output. Conventional UniDA methods require direct access to source samples for domain adaptation. In response to burgeoning data protection policies, the source-free universal domain adaptation (SF-UniDA) is proposed in [30]. However, GLC [30] still requires an extensive quantity of labeled source samples to

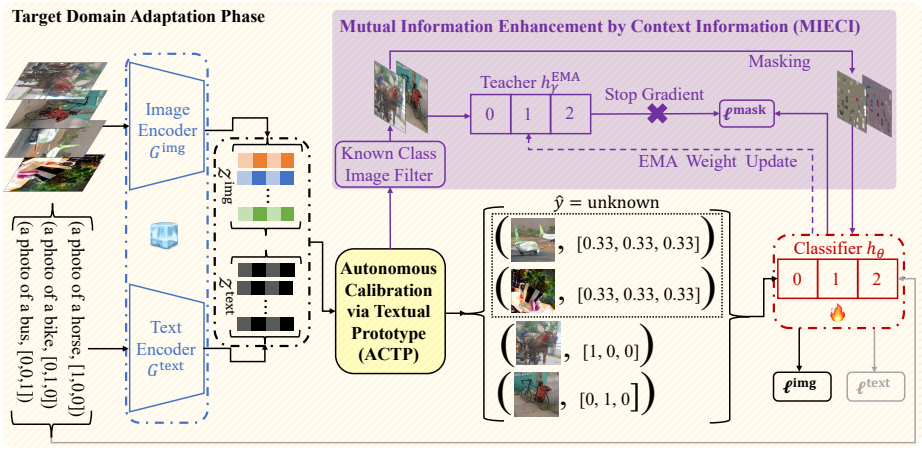


Fig. 3: Overview of the COCA method. COCA adapts the closed-set classifier h_θ to the target domain to tackle the SF-UniDA challenge.

develop a source model, resulting in significant labeling costs. Moreover, We observe that the existing UniDA and SF-UniDA paradigms concentrating on image encoder optimization are ill-suitable for VLMs.

Few-Shot Learning. In the context of few-shot visual classification, a classifier is initially pretrained on a set of base classes [9] to learn a good feature representation and then finetuned on a limited number of novel class samples. In recent advancements, some few-shot learning methods [13,24] based on CLIP [31] have been proposed. However, these methods, specifically designed for i.i.d. datasets and the closed-set scenario, exhibit performance deterioration due to domain and category shifts, making them unsuitable to be directly applied in UniDA/SF-UniDA. Furthermore, it is unfeasible to label few-shot target samples in UniDA/SF-UniDA since the target class set is uncertain. *E.g.*, in OSDA/OPDA scenarios, the quantity of unknown classes remains uncertain, and in OPDA/PDA, certain classes might be absent in the target domain.

3 Methodology

Preliminaries. We have a labeled source domain $\mathcal{D}^s = \{(x^s, y^s) : y^s \in \mathcal{C}^s\}$, where $\mathcal{C}^s$ denotes the source class set. Each class name in the source domain is associated with its respective ground truth label, denoted as y_c. For instance, the class name `horse` carries the ground truth label $y_c = [1, 0, 0]$ in Fig. 3. We also have an unlabeled target domain $\mathcal{D}^t = \{(x^t)\}$ with a domain distribution that differs from that of $\mathcal{D}^s$. Assuming $\mathcal{C}^t$ is the target class set, the relationship between $\mathcal{C}^s$ and $\mathcal{C}^t$ in the UniDA scenario can be categorized into $\mathcal{C}^s \subset \mathcal{C}^t$ (OSDA), $\mathcal{C}^t \subset \mathcal{C}^s$ (PDA), and $\mathcal{C}^s \cap \mathcal{C}^t \neq \emptyset, \mathcal{C}^s \not\subset \mathcal{C}^t, \mathcal{C}^t \not\subset \mathcal{C}^s$ (OPDA). The classes in $\mathcal{C}^s$ are referred to as known classes, while the classes in $\mathcal{C} = \mathcal{C}^s \cap \mathcal{C}^t$ are

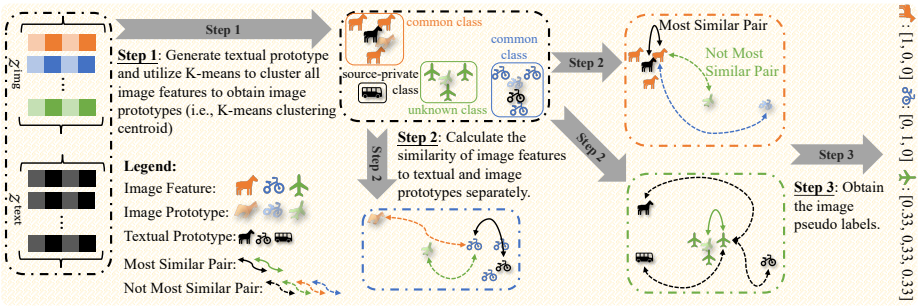


Fig. 4: Pipeline of the ACTP module. `horse` and `bike` are the common classes, `bus` is the source-private class, and `airplane` is the unknown class. In Step 3, ACTP generates pseudo labels via Eq. (6).

termed common classes. The source- and target-private class sets are defined as $\overline{\mathcal{C}^s} = \mathcal{C}^s \setminus \mathcal{C}^t$ and $\overline{\mathcal{C}^t} = \mathcal{C}^t \setminus \mathcal{C}^s$, and the classes in $\overline{\mathcal{C}^t}$ are unknown classes.

Overview. The overview of the classifier-oriented calibration (COCA) approach is depicted in Fig. 3. For the few-shot learning problem, we use the VLM-powered few-shot learner trained on the source domain as the source model. In the zero-shot learning problem, the classifier's weights are initialized with the text features. At the target domain adaptation phase, we adapt the closed-set classifier h_θ of the source model to the target domain. To enable the classifier to distinguish common and unknown class samples in the target domain, we propose the autonomous calibration via the textual prototype (ACTP) module that exploits the close feature distance between text features and image prototypes. We introduce the mutual information enhancement by context information (MIECI) module, which includes a teacher classifier h_γ^{EMA}, to encourage the classifier h_θ to exploit context information in images and enhance the mutual information.

3.1 Autonomous Calibration via Textual Prototype

To enable the closed-set classifier h_θ to distinguish common and unknown class samples, we introduce the autonomous calibration via textual prototype (ACTP) module. The pipeline of the ACTP module is depicted in Fig. 4.

Positive and Negative Prototypes. Leveraging the close feature distance between text features and image prototypes generated by VLMs, we can identify positive and negative prototypes for known classes using the class names. To generate positive prototypes for known classes, we input `a photo of a {CLS}` to the text encoder to obtain a text feature set $\mathcal{Z}^{\text{text}} = \{z_c^{\text{text}}\}_{c=1}^{|\mathcal{C}^s|}$, where $|\mathcal{C}^s|$ is the number of source classes and the class tokens `{CLS}` are replaced by specific class names in the source domain, such as `horse`, `bike`, or `bus`. The prediction probability that a target image x_i belongs to a known class c is computed as:

$$p(y=c|x_i) = \frac{\exp(\cos(z_c^{\text{text}}, z_i^{\text{img}})/T)}{\sum_{j=1}^{|\mathcal{C}^s|} \exp(\cos(z_j^{\text{text}}, z_i^{\text{img}})/T)}, \tag{1}$$

where T is a temperature parameter, $\cos(\cdot,\cdot)$ denotes cosine similarity, z_i^{img} indicates the image feature of target image x_i, and z_j^{text} represents the j-th element of the text feature set $\mathcal{Z}^{\text{text}}$. Conversely, as we lack information on the target-domain class name, generating negative prototypes via class names is impractical. Therefore, to locate negative prototypes for a known class c, we employ K-means [27] to cluster all target image features and generate an image prototype (cluster centroid) set $\mathcal{V}^{\text{img}} = \{v_k^{\text{img}}\}_{k=1}^{K}$, where K is the K-means hyperparameter. The formal definition is as follows:

$$\{v_k^{\text{img}}\}_{k=1}^{K} = \text{K-means}(\{z_i^{\text{img}}\}_{i=1}^{N}), \tag{2}$$

where N is the number of target samples. We will later discuss optimal K value determination. The K value sensitivity analysis can be found in the experiment. After clustering, there are K image prototypes. We assume there are one positive image prototype p^c and $K-1$ negative image prototypes $\{n_k^c\}_{k=1}^{K-1}$ for a known class c. p^c and $\{n_k^c\}_{k=1}^{K-1}$ for a known class c are determined by:

$$\begin{cases} p^c = v_j^{\text{img}}, \text{ if } \cos(z_c^{\text{text}}, v_j^{\text{img}}) = \max\{\cos(z_c^{\text{text}}, v_k^{\text{img}})\}_{k=1}^{K} \\ \{n_k^c\}_{k=1}^{K-1} = \{v_k^{\text{img}}\}_{k=1}^{K} / \{v_j^{\text{img}}\}. \end{cases} \tag{3}$$

In this paper, we use the text feature z_c^{text} generated by the text encoder instead of the positive image prototype p^c as the positive prototype, and we employ the negative image prototypes $\{n_k^c\}_{k=1}^{K-1}$ as the negative prototypes. According to the information bottleneck theory [1], we need to maximize the objective function

$$R_{\text{IB}}(\omega) = I(\mathcal{Z}, \mathcal{Y}; \omega) - \beta I(\mathcal{Z}, \mathcal{X}; \omega), \tag{4}$$

where R_{IB} is the information bottleneck, $\mathcal{X}$ is the input image set, $\mathcal{Y}$ is the output label set, and I denotes the mutual information of input/output sets with the feature set $\mathcal{Z}$. Here, the parameters ω of encoders in VLMs are held constant, causing the value of R_{IB} to be contingent upon the features within $\mathcal{Z}$. The image feature set $\mathcal{Z}^{\text{img}}$ derived from the target image set $\mathcal{X}$ inherently encapsulates domain-specific information. In contrast, the text feature set $\mathcal{Z}^{\text{text}}$ exhibits similarities to the image prototype set but lacks domain-related information, *i.e.*, $I(\mathcal{Z}^{\text{img}}, \mathcal{X}; \omega) > I(\mathcal{Z}^{\text{text}}, \mathcal{X}; \omega)$, which further implies $I(\mathcal{V}^{\text{img}}, \mathcal{X}; \omega) > I(\mathcal{Z}^{\text{text}}, \mathcal{X}; \omega)$. For another thing, the image prototype set $\mathcal{V}^{\text{img}}$ resulting from a limited set of images may not adequately capture the essence of a class. *E.g.*, the distribution of images for subclasses such as pony and zebra within the `horse` class in DomainNet is imbalanced, rendering the image prototype inadequate in representing the `horse` class. Consequently, we deduce that $I(\mathcal{Z}^{\text{text}}, \mathcal{Y}; \omega) > I(\mathcal{V}^{\text{img}}, \mathcal{Y}; \omega)$. This leads us to conclude that:

$$R_{\text{IB}}(\mathcal{Z}^{\text{text}}) > R_{\text{IB}}(\mathcal{V}^{\text{img}}). \tag{5}$$

Furthermore, we observe that the quality of the positive image prototype p^c heavily relies on the empirical determination of appropriate hyperparameters K for clustering. If the hyperparameter K is set smaller than the oracle K, it poses

the risk of merging two distinct classes into a single positive image prototype. If K is larger than the oracle K, the positive image prototype may represent a subclass concept rather than the intended class concept. On the other hand, the text feature z_c^{text}, derived from text, remains unaffected by K. Thus, it becomes evident that utilizing the text feature $z_c^{\text{text}} \in \mathcal{Z}^{\text{text}}$ as the positive prototype stands as a more rational choice than employing the image prototype $p^c \in \mathcal{V}^{\text{img}}$.

Self-Training. Given the positive prototype z_c^{text} and the set of negative prototypes $\{n_k^c\}_{k=1}^{K-1}$, the ACTP module generates a pseudo label $\hat{y}_i$ for a target sample x_i to implement self-training. The formal definition is as follows:

$$\hat{y}_i = \begin{cases} \text{one-hot}(c), & \text{if } \exists p(y=c|x_i) \geq \max\{\cos(z_c^{\text{text}}, n_k^c)\}_{k=1}^{K-1} \\ \texttt{unknown}, & \text{else} \end{cases} \tag{6}$$

where $p(y=c|x_i)$ is the probability in Eq. (1). As shown in Fig. 3, for a target sample x_i with pseudo label `unknown`, ACTP assigns a uniform encoding $\hat{y}_i = [1/|\mathcal{C}^s|, ..., 1/|\mathcal{C}^s|] \in \mathbb{R}^{1\times|\mathcal{C}^s|}$ to increase the uncertainty for x_i. Hence, the model can distinguish common and unknown class samples at the inference phase.

Image Loss and Text Loss. The image cross-entropy loss with pseudo labels $\hat{y}_i$ is defined as follows:

$$\ell^{\text{img}} = -\frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \log(\sigma(h_\theta(z_i^{\text{img}}))), \tag{7}$$

where σ and h_θ denote the softmax function and the closed-set classifier respectively. For another thing, given a text feature derived from a known (source) class name such as `horse`, its ground truth label can be determined. The text cross-entropy loss with ground truth labels y_c is defined as follows:

$$\ell^{\text{text}} = -\frac{1}{|\mathcal{C}^s|}\sum_{c=1}^{|\mathcal{C}^s|} y_c \log(\sigma(h_\theta(z_c^{\text{text}}))). \tag{8}$$

Optimal K Determination. To determine the optimal K for K-means, various methods have been introduced in [32,3,8]. Analysis of the effects of these methods on COCA can be found in the supplementary material. In summary, COCA performs stably across these methods. In this paper, we employ the Silhouette metric [32] to estimate the value of K. The introduction to the Silhouette score can be found in the supplementary material. We consider the potential values of K as $[1/3|\mathcal{C}^s|, 1/2|\mathcal{C}^s|, |\mathcal{C}^s|, 2|\mathcal{C}^s|, 3|\mathcal{C}^s|]$ following [30]. The model selects the optimal K from this list using the Silhouette value. Since we freeze the image and text encoders, the image and text features do not change. Hence, we estimate the K value only at the first epoch.

3.2 Mutual Information Enhancement by Context Information

In order to encourage the closed-set classifier to exploit the context information in images and enhance the mutual information, we introduce the mutual information enhancement by context information (MIECI) module.

Masked Image and EMA Teacher. We use a patch mask M [16], which is randomly sampled from a uniform distribution $\mathcal{U}(0,1)$, to mask out images:

$$M_{mw+1:(m+1)w,nw+1:(n+1)w} = \mathbb{1}_{(v>r)} \text{ with } v \sim \mathcal{U}(0,1), \tag{9}$$

where w is the patch size, r is the mask ratio, and $m, n \in \{0, ..., W/w - 1\}$ are the patch indices, W is the input image size. The masked image x_i^M is obtained by performing element-wise multiplication of the mask M and target image x_i:

$$x_i^M = M \odot x_i. \tag{10}$$

We use the exponential moving average (EMA) teacher [39] as the classifier h_γ^{EMA}. Its weights are updated via the weights of h_θ with a smoothing factor α:

$$\gamma_{iter+1} \leftarrow \alpha\gamma_{iter} + (1-\alpha)\theta_{iter}, \tag{11}$$

where *iter* denotes iteration. The original image x_i is fed to the teacher classifier h_γ^{EMA} and the masked image x_i^M to the classifier h_θ. The probability $p(y|x_i;\gamma)$ generated by the teacher classifier h_γ^{EMA} for x_i is considered as a soft label to compute the mask loss ℓ^{mask} to optimize the classifier h_θ. The formal definition of $p(y|x_i;\gamma)$ is:

$$p(y|x_i;\gamma) = \sigma(h_\gamma^{\text{EMA}}(G^{\text{img}}(x_i))). \tag{12}$$

Additionally, since classifiers $h_\gamma^{\text{EMA}}, h_\theta$ are closed-set, we need to select the target samples belonging to known classes. In this paper, we use the pseudo labels $\hat{y}_i$ generated by the ACTP module to assist in selecting known class samples.

Mask Loss. The mask loss $\mathcal{L}^{\text{mask}}$ in the MIECI module is defined as:

$$\ell^{\text{mask}} = \mathbb{E}(-\mathbb{1}_{(\hat{y}_i \in \mathcal{C}^s)} p(y|x_i;\gamma) \log(\sigma(h_\theta(G^{\text{img}}(x_i^M))))), \tag{13}$$

where $\mathbb{E}$ indicates expectation. The mask loss ℓ^{mask} encourages the classifier h_θ to produce a similar probability distribution, akin to the EMA teacher h_γ^{EMA}, when provided with a known class masked image x_i^M. It enhances the capacity of h_θ to harness contextual cues effectively. Prompting the model to utilize contextual information enhances $p(y|z;\theta)$, thereby increasing the mutual information $I(\mathcal{Z}^{\text{img}}, \mathcal{Y}; \theta)$, particularly in domains with limited texture information.

3.3 Model Optimization

The overall training loss of our approach can be written as:

$$\ell = \ell^{\text{img}} + \ell^{\text{text}} + \ell^{\text{mask}}. \tag{14}$$

Decision Boundary Adaptation. The general objective of DA is to establish a decision boundary by minimizing classification loss of source samples and explicitly design a loss term measuring domain divergence to make the fixed decision boundary suitable for target samples. This objective aims to transfer

the source domain knowledge to the target domain. (1) In contrast to DA, SF-UniDA tasks do not allow direct access to source samples, making it impossible to estimate the source data distribution. As a result, explicitly designing a loss term for measuring domain divergence is not feasible. Consequently, in this work, we consider to adapt the decision boundary, *i.e.*, the classifier. The core idea is if we properly adapt the decision boundary to target domains, the classification error will be low without the need to explicitly design a domain shift loss term. (2) We contend that the knowledge of both the source and target domains has been encapsulated within the VLMs, given its pretraining on large datasets. Therefore, the explicit design of a domain divergence measurement term for optimizing the image and text encoders to transfer source domain knowledge becomes redundant. The focus, instead, should be shifted toward guiding the classifier to establish a more appropriate decision boundary. Given the above two reasons, we propose this new paradigm as shown in Fig.1, which concentrates on classifier optimization. We exploit the close feature distance between image prototypes and text features to generate positive and negative prototypes, thereby adapting the classifier's decision boundary to a new target domain. In the experiment, we demonstrate that our new paradigm is applicable not only to VLM-powered few-shot learners but also to zero-shot classifiers such as single linear layer [31] or the adapter module [13], suggesting that the image and text encoders within the VLMs have encapsulated knowledge of both the source and target domains.

3.4 Inference

Since h_θ is a closed-set classifier, we utilize the normalized Shanon Entropy [37] to measure the uncertainty of target samples $U(x_i) \in [0, 1]$ to distinguish common and unknown class samples:

$$U(x_i) = -\frac{1}{\log |\mathcal{C}^s|} \sigma(h_\theta(z_i^{\text{img}})) \log(\sigma(h_\theta(z_i^{\text{img}}))). \tag{15}$$

At the inference phase, we separate common and unknown class samples as:

$$y(x_i) = \begin{cases} \text{argmax}(h_\theta(z_i^{\text{img}})), & \text{if } U(x_i) < \tau \\ \texttt{unknown}. & \text{else.} \end{cases} \tag{16}$$

The inference phase is depicted in Fig.5. We set $\tau = 0.55$ for all our experiments. Its sensitivity analysis can be found in the supplementary material.

4 Experiments

4.1 Datasets and Evaluation Metric

Dataset. We utilize the following three public benchmarks: **OfficeHome** [40], **VisDA-2017** [29], and **DomainNet** [28], to evaluate the effectiveness of our approach. OfficeHome consists of four domains: Art (Ar), Clipart (Cl), Product

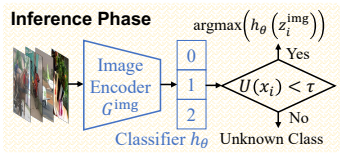


Fig. 5: COCA at the inference phase.

Table 1: Details regarding the split of classes and the average number of source samples per class provided to methods.

Dataset	Class Split ($\mathcal{C}/\overline{\mathcal{C}^s}/\overline{\mathcal{C}^t}$)			Number of Source Samples Per Class	
	OPDA	OSDA	PDA	Others	Ours
OfficeHome	10/5/50	25/0/40	25/40/0	60 (full set)	[0, 1, 4, 16]
VisDA-2017	6/3/3	6/0/6	6/6/0	12,700 (full set)	[0, 1, 4, 16]
DomainNet	150/50/145	-	-	304 (full set)	[0, 1, 4, 16]

Table 2: HOS (%) comparison in **OPDA** on OfficeHome. **SF** denotes support for source-free learning, while **FSL** indicates support for few-shot learning on the source domain. Best in **bold**.

Method		SF	FSL	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**
OSBP [36]	ResNet50	✗	✗	39.6	45.1	46.2	45.7	45.2	46.8	45.3	40.5	45.8	45.1	41.6	46.9	44.5
DCC [20]		✗	✗	58.0	54.1	58.0	74.6	70.6	77.5	64.3	73.6	74.9	81.0	75.1	80.4	70.2
DCC+SPA [18]		✗	✗	59.3	79.5	81.5	74.7	71.7	82.0	68.0	74.7	75.8	74.5	75.8	81.3	74.9
OVANet [35]		✗	✗	62.8	75.5	78.6	70.7	68.8	75.0	71.3	58.6	80.5	76.1	64.1	78.9	71.8
GATE [7]		✗	✗	63.8	70.5	75.8	66.4	67.9	71.7	67.3	61.5	76.0	70.4	61.8	75.1	69.0
UniOT [6]		✗	✗	67.3	80.5	86.0	73.5	77.3	84.3	75.5	63.3	86.0	77.8	65.4	81.9	76.6
UMAD [22]		✓	✗	61.1	76.3	82.7	70.7	67.7	75.7	64.4	55.7	76.3	73.2	60.4	77.2	70.1
GLC [30]		✓	✗	64.3	78.2	89.8	63.1	81.7	89.1	77.6	54.2	88.9	80.7	54.2	85.9	75.6
CoDE [38]		✓	✗	65.8	77.4	87.6	73.5	69.1	80.2	77.6	62.0	84.1	84.3	68.4	83.1	76.1
OVANet [35]	ViT	✗	✗	55.8	77.6	86.2	72.6	69.9	81.1	76.3	47.4	84.7	79.9	53.0	79.6	72.0
UniOT [6]		✗	✗	63.8	88.2	90.2	75.0	81.0	84.6	78.9	61.3	87.6	82.4	63.7	88.3	78.4
UniAM [46]		✗	✗	72.0	87.1	90.7	80.3	82.4	79.8	85.0	68.4	89.0	85.4	72.1	86.1	81.7
GLC [30]		✓	✗	68.5	**89.8**	**91.0**	82.4	**88.1**	**89.4**	82.1	69.7	88.2	82.4	70.9	88.9	82.6
DCC [20]	CLIP	✗	✗	62.6	88.7	87.4	63.3	68.5	79.3	67.9	63.8	82.4	70.7	69.8	87.5	74.4
OVANet [35]		✗	✗	71.0	85.4	86.8	79.4	78.8	85.4	76.8	64.8	**89.1**	83.2	70.3	84.4	79.6
GLC [30]		✓	✗	79.4	88.9	90.8	76.3	84.7	89.0	71.5	72.9	85.7	78.2	79.4	**90.0**	82.2
Source Model (16-shot) [24]		✗	✓	43.5	31.6	39.5	52.0	33.9	43.1	55.7	51.9	44.0	49.9	54.4	27.6	43.9
+ COCA-w-p^c		✓	✓	83.6	85.2	87.2	81.9	84.1	86.7	78.0	82.4	86.7	82.7	82.8	84.1	83.8
+ COCA		✓	✓	**83.7**	86.6	89.0	**88.0**	86.5	88.9	**88.2**	**83.9**	88.8	**88.3**	**83.7**	86.6	**86.9**

(Pr), and Real-World (Rw). Following preceding studies [7,30], our experiments were conducted on three domains of DomainNet: Painting (P), Real (R), and Sketch (S). We conducted evaluations of our method in OPDA, OSDA, and PDA. A summary of class splits and the comparison of the number of source samples per class on average provided to different methods are detailed in Table 1. The classes are separated according to their alphabetical order. While other methods are fully trained on source samples, our approach was allotted merely few-shot source samples per class for source model training.

Evaluation Metric. We use the same evaluation metric as previous works [7,30] for a fair comparison. Specifically, we report *HOS* [2], also referred to as H-score [11], to evaluate the model performance in OPDA and OSDA scenarios.

4.2 Comparisons with State-of-the-Art Methods

We conducted extensive experiments across three category-shift scenarios: OPDA, OSDA, and PDA. In Sec. 4.2 and Sec. 4.3, the source model is the cross-modal linear probing model [24] based on CLIP(ViT-B/16). **COCA-w-p^c** represents that we utilize the image prototype p^c in Eq. (3) instead of the text feature z_c^{text} as the positive prototype. We also included the results from DCC, OVANet, and

Table 3: HOS (%) in **OPDA** on DomainNet and VisDA-2017.

Method		SF	FSL	DomainNet							VisDA
				P→R	P→S	R→P	R→S	S→P	S→R	**Avg**	
OSBP [36]	ResNet50	✗	✗	33.6	30.6	33.0	30.6	30.5	33.7	32.0	27.3
DCC [20]		✗	✗	56.9	43.7	50.3	43.3	44.9	56.2	49.2	43.0
DCC+SPA [18]		✗	✗	59.1	52.7	47.6	45.4	46.9	56.7	51.4	-
OVANet [35]		✗	✗	56.0	47.1	51.7	44.9	47.4	57.2	50.7	53.1
GATE [7]		✗	✗	57.4	48.7	52.8	47.6	49.5	56.3	52.1	56.4
UniOT [6]		✗	✗	59.3	51.8	47.8	48.3	46.8	58.3	52.0	57.3
UMAD [22]		✓	✗	59.0	44.3	50.1	42.1	32.0	55.3	47.1	58.3
GLC [30]		✓	✗	63.3	50.5	54.9	50.9	49.6	61.3	55.1	73.1
CoDE [38]		✓	✗	62.7	47.1	51.3	43.4	48.3	60.9	52.3	74.5
OVANet [35]	ViT	✗	✗	65.0	42.5	54.7	37.7	40.5	58.9	49.9	49.6
UniOT [6]		✗	✗	72.4	49.3	59.5	47.4	56.9	69.4	59.1	63.3
UniAM [46]		✗	✗	73.9	52.3	60.9	51.4	60.0	70.7	61.5	65.2
GLC [30]		✓	✗	67.6	51.1	55.9	46.6	53.8	66.0	56.8	80.3
DCC [20]	CLIP	✗	✗	61.1	38.8	51.8	49.3	49.1	60.3	52.2	61.2
OVANet [35]		✗	✗	65.4	53.7	56.3	53.1	55.9	67.2	58.6	72.0
GLC [30]		✓	✗	74.4	63.4	60.0	62.9	52.0	74.3	64.5	77.6
Source Model (16-shot) [24]		✗	✓	46.5	51.8	56.3	53.9	32.7	33.2	45.7	32.2
+ COCA-w-p^c		✓	✓	78.9	**69.2**	68.5	**69.0**	67.6	77.6	71.8	79.8
+ COCA		✓	✓	**80.8**	**69.2**	**69.6**	**69.0**	**69.4**	**80.5**	**73.1**	**83.2**

Table 4: HOS (%) comparison in **OSDA** on OfficeHome and VisDA-2017.

Method		SF	FSL	OfficeHome													VisDA
				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**	
OSBP [36]	ResNet50	✗	✗	55.1	65.2	72.9	64.3	64.7	70.6	63.2	53.2	73.9	66.7	64.5	72.3	64.7	52.3
STA_{max} [25]		✗	✗	55.8	54.0	68.3	57.4	60.4	66.8	61.9	53.2	69.5	67.1	54.5	64.5	61.1	65.0
DCC [20]		✗	✗	56.1	67.5	66.7	49.6	66.5	64.0	55.8	53.0	70.5	61.6	57.2	71.9	61.7	59.6
OVANet [35]		✗	✗	58.9	66.0	70.4	62.2	65.7	67.8	60.0	52.6	69.7	68.2	59.1	67.6	64.0	66.1
UADAL [17]		✗	✗	**76.9**	56.6	63.0	70.8	77.4	63.2	72.1	**76.8**	60.6	73.4	64.2	69.5	68.7	61.3
GATE [7]		✗	✗	63.8	70.5	75.8	66.4	67.9	71.7	67.3	61.5	76.0	70.4	61.8	75.1	69.0	70.8
UMAD [22]		✓	✗	59.2	71.8	76.6	63.5	69.0	71.9	62.5	54.6	72.8	66.5	57.9	70.7	66.4	66.8
GLC [30]		✓	✗	65.3	74.2	79.0	60.4	71.6	74.7	63.7	63.2	75.8	67.1	64.3	77.8	69.8	72.5
OVANet [35]	ViT	✗	✗	55.5	71.1	76.9	64.6	67.4	75.1	64.4	47.2	76.7	71.4	53.5	70.5	66.2	57.4
GLC [30]		✓	✗	68.4	81.7	**84.5**	76.0	82.4	**83.8**	69.9	59.6	**84.6**	73.3	66.8	83.9	76.2	81.6
DCC [20]	CLIP	✗	✗	62.9	73.3	78.4	49.8	69.2	75.0	59.3	61.5	80.9	68.1	62.5	80.0	68.4	66.2
OVANet [35]		✗	✗	65.0	73.6	77.5	71.1	73.9	79.0	65.7	54.6	80.3	73.2	61.1	77.1	71.0	70.7
GLC [30]		✓	✗	72.1	79.7	83.3	55.5	81.3	77.9	52.1	65.9	78.2	69.0	71.3	83.9	72.5	83.4
Source Model (16-shot) [24]		✗	✓	33.3	13.5	17.8	36.0	18.5	23.0	37.4	48.6	28.1	18.3	11.7	8.8	24.6	55.6
+ COCA-w-p^c		✓	✓	68.3	**85.5**	78.9	73.4	**86.8**	79.0	73.6	67.9	78.5	75.4	73.0	**85.7**	77.2	70.7
+ COCA		✓	✓	75.6	84.5	82.5	**79.7**	84.3	82.5	**79.6**	74.5	82.5	**80.0**	**75.7**	84.4	**80.5**	**86.3**

GLC models with the ViT-B/16 [10] pre-trained on ImageNet [9] and with the CLIP(ViT-B/16), to ensure a more fair comparison.

Results for OPDA. Our first experiment focuses on the most challenging setting, *i.e.*, OPDA. Results for OfficeHome are shown in Table 2, and those for VisDA-2017 and DomainNet are summarized in Table 3. As presented in Table 2 and Table 3, our approach outperforms all previous methods, despite these methods utilizing the full set of source samples for model training. The source model, *i.e.*, cross-modal linear probing [24], utilizes the uncertainty in Eq. (16) to distinguish common and unknown class samples. The source model exhibits subpar performance in the OPDA scenario due to its inability to distinguish between common and unknown class samples. In the most challenging benchmark, *i.e.*, DomainNet, our approach surpasses all previous works by a dramatically wide margin. It implies the potential of COCA on large-scale DA datasets.

Results for OSDA. Subsequent experiments are conducted on OSDA. The corresponding results for OfficeHome and VisDA are shown in Table 4. As ev-

Table 5: Accuracy (%) comparison in **PDA** on OfficeHome and VisDA-2017.

Method		SF	FSL	OfficeHome														VisDA
				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**	
ETN [5]	ResNet50	✗	✗	59.2	77.0	79.5	62.9	65.7	75.0	68.3	55.4	84.4	75.7	57.7	84.5	70.4	59.8
BA3US [23]		✗	✗	60.6	83.2	88.4	71.8	72.8	83.4	75.5	61.6	86.5	79.3	62.8	86.1	76.0	54.9
DANCE [34]		✗	✗	53.6	73.2	84.9	70.8	67.3	82.6	70.0	50.9	84.8	77.0	55.9	81.8	71.1	73.7
DCC [20]		✗	✗	54.2	47.5	57.5	83.8	71.6	86.2	63.7	65.0	75.2	85.5	**78.2**	82.6	70.9	72.4
OVANet [35]		✗	✗	34.1	54.6	72.1	42.4	47.3	55.9	38.2	26.2	61.7	56.7	35.8	68.9	49.5	34.3
GATE [7]		✗	✗	55.8	75.9	85.3	73.6	70.2	83.0	72.1	59.5	84.7	79.6	63.9	83.8	74.0	75.6
SHOT-P [21]		✓	✗	64.7	85.1	90.1	75.1	73.9	84.2	76.4	64.1	90.3	80.7	63.3	85.5	77.8	74.2
GLC [30]		✓	✗	55.9	79.0	87.5	72.5	71.8	82.7	74.9	41.7	82.4	77.3	60.4	84.3	72.5	76.2
OVANet [35]	ViT	✗	✗	34.7	61.3	76.8	49.8	51.5	60.6	43.9	25.9	70.1	63.1	32.5	72.9	53.6	32.0
GLC [30]		✓	✗	63.2	80.7	86.5	76.0	77.9	84.1	74.5	56.8	84.7	79.8	57.4	83.0	75.4	84.0
DCC [20]	CLIP	✗	✗	59.4	78.8	83.2	62.0	78.6	79.3	64.2	44.4	82.9	76.5	70.7	84.6	72.1	79.8
OVANet [35]		✗	✗	39.2	55.7	72.8	61.9	63.4	71.6	47.6	30.0	73.2	61.0	40.1	70.5	57.3	41.4
GLC [30]		✓	✗	**77.8**	82.8	89.5	68.7	81.8	86.4	74.3	**75.3**	86.3	79.0	78.1	**87.2**	80.6	84.4
Source Model (16-shot) [24]		✗	✓	71.6	84.9	88.4	80.7	81.2	87.8	79.7	73.3	87.6	**85.7**	74.1	86.1	81.8	85.6
+ COCA		✓	✓	69.1	**87.7**	**92.4**	**83.9**	**86.7**	**90.7**	**83.8**	68.6	**90.5**	84.3	69.7	**87.2**	**82.9**	**89.1**

Table 6: HOS (%) with respect to K in **OPDA**. COCA and COCA-w-p^c are based on the source model (16-shot)[24].

Model	OfficeHome ($\lvert\mathcal{C}^t\rvert = 60$)						VisDA-2017 ($\lvert\mathcal{C}^t\rvert = 9$)						DomainNet ($\lvert\mathcal{C}^t\rvert = 295$)					
	K=8	K=15	K=30	K=45	K=60	K=75	K=5	K=9	K=18	K=27	K=36	K=45	K=100	K=200	K=400	K=600	K=800	K=1000
GLC (ViT)	75.5	77.4	80.5	82.0	74.1	68.0	73.7	80.3	47.8	24.1	23.2	21.1	52.0	56.8	47.0	37.4	25.2	24.5
GLC (CLIP)	79.3	80.7	82.5	81.4	78.8	76.5	62.6	77.6	53.0	10.2	9.7	9.6	62.0	64.5	53.8	46.6	41.1	46.3
COCA-w-p^c	70.7	73.0	78.8	83.8	84.6	82.8	78.8	79.8	63.4	44.8	36.2	22.4	66.5	70.0	71.8	72.3	72.0	72.0
COCA	**85.1**	**85.9**	**86.7**	**86.9**	**87.3**	**87.4**	**82.4**	**83.2**	**83.6**	**83.2**	**83.7**	**83.3**	**73.0**	**73.1**	**73.0**	**72.8**	**72.7**	**72.6**

idenced in Table 4, our method surpasses all prior models, despite the source model training based on few-shot source samples. These results indicate that our proposed plug-and-play approach can aid the closed-set few-shot learner in differentiating between common and unknown class samples within the target domain, consequently lowering the total labeling cost.

Results for PDA. Lastly, we evaluate the effectiveness of COCA on PDA, where the class set $\mathcal{C}^t$ of the target domain is a subset of the source domain. The shown results in Table 5 demonstrate that our proposed method surpasses previous approaches, even those [5,23] specifically designed for PDA.

4.3 Ablation Studies

Hyperparameter Sensitivity. The comparison experiments of K values are shown in Table 6. COCA, exploiting textual prototypes, exhibits consistent performance across various K values and is better and more stable than COCA-w-p^c and the best-performed GLC [30], both utilizing image-based positive prototypes. COCA-w-p^c and GLC are sensitive to the choice of K values. In cases where K values are inappropriate, the image-based positive prototypes may fail to accurately represent corresponding classes. Specifically, if K values are smaller than the oracle K, distinct classes might erroneously merge into a single cluster. Conversely, if K values are larger than the oracle K, different subclasses within the same class may be separated into distinct clusters, causing the image-based positive prototypes to represent subclasses instead of the intended classes. The

Table 7: HOS (%) with respect to various source models in **OPDA**.

Source Model	Linear Probe CLIP [31]						CLIP-Adapter [13]						Cross-Modal Linear Probing [24]			
	0-shot	+COCA	1-shot	+COCA	16-shot	+COCA	0-shot	+COCA	1-shot	+COCA	16-shot	+COCA	1-shot	+COCA	4-shot	+COCA
OfficeHome	61.1	86.9	57.7	86.7	51.7	86.7	73.7	87.1	71.4	87.0	58.5	86.9	52.9	86.9	57.7	87.0
VisDA-2017	59.2	83.0	41.1	83.6	44.0	83.4	71.0	83.0	46.0	83.7	43.4	83.5	41.1	82.9	37.3	82.8
DomainNet	57.8	72.5	54.1	72.6	45.3	73.1	69.3	71.3	67.5	72.1	47.4	73.1	54.1	72.8	50.0	72.7

experiment results and the analysis in Eq. (5) indicate that **the textual prototypes are better suited as the positive prototypes**.

Results on Varying Source Models. Results on varying source models in OPDA are shown in Table 7. With respect to the diverse tasks in OPDA, the outcomes reveal that the shots used in source model training do not significantly affect our method performance. We attribute this observation to the phenomenon of catastrophic forgetting. Specifically, since the few-shot source samples cannot be accessed at the target domain adaptation phase, the knowledge derived from a massive number of target samples with pseudo labels covers the knowledge from few-shot source samples. In contrast to traditional UniDA methods, our approach's success is not predicated on knowledge extracted from a profusion of source samples. Instead, its effectiveness hinges upon the quality of self-training. Therefore, we choose the textual prototypes instead of the image prototypes as the positive ones since they demonstrate their advantages in SF-UniDA scenarios. We posit that the knowledge learned from the massive number of labeled source samples is no longer a crucial factor for the success of VLMs in addressing the SF-UniDA challenge. It means the labeling cost for source samples in future works can be dramatically reduced compared with traditional SF-UniDA methods when utilizing VLMs. Furthermore, the outcomes from diverse source models [31,13,24] validate that our plug-and-play approach is compatible with multiple few-shot learning frameworks, ensuring consistent performance. The results from "zero-shot linear probe CLIP + COCA" and "zero-shot CLIP-Adapter + COCA" indicate that **VLMs have encapsulated knowledge of both the source and target domains**. This suggests that we should **focus on classifier optimization to adapt the VLM-powered models to new target domains**.

5 Conclusion

In this paper, we present a novel plug-and-play method, called COCA, that endows VLM-powered few-shot learners with the unknown-aware ability to tackle the SF-UniDA challenge. Our paradigm shifts the focus to study classifier optimization for SF-UniDA since we realize that VLMs have encapsulated knowledge of the source and target domains, to some extent. We hope that our method will inspire more SF-UniDA research.

Acknowledgments. This study was funded by the National Natural Science Foundation of China (grant number 62106043) and the Natural Science Foundation of Jiangsu Province (grant number BK20210225).

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: Int. Conf. Learn. Represent. (2017)
2. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: Eur. Conf. Comput. Vis. pp. 422–438 (2020)
3. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974)
4. Cao, Z., Ma, L., Long, M., Wang, J.: Partial adversarial domain adaptation. In: Eur. Conf. Comput. Vis. pp. 135–150 (2018)
5. Cao, Z., You, K., Long, M., Wang, J., Yang, Q.: Learning to transfer examples for partial domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2985–2994 (2019)
6. Chang, W., Shi, Y., Tuan, H., Wang, J.: Unified optimal transport framework for universal domain adaptation. In: Adv. Neural Inform. Process. Syst. pp. 29512–29524 (2022)
7. Chen, L., Lou, Y., He, J., Bai, T., Deng, M.: Geometric anchor correspondence mining with uncertainty modeling for universal domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16134–16143 (2022)
8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (1979)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255 (2009)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
11. Fu, B., Cao, Z., Long, M., Wang, J.: Learning to detect open classes for universal domain adaptation. In: Eur. Conf. Comput. Vis. pp. 567–583 (2020)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.S.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**, 59:1–59:35 (2016)
13. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters (2021)
14. Gu, X., Yu, X., yang, y., Sun, J., Xu, Z.: Adversarial reweighting for partial domain adaptation. In: Adv. Neural Inform. Process. Syst. pp. 14860–14872 (2021)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)
16. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: MIC: Masked image consistency for context-enhanced domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
17. Jang, J., Na, B., Shin, D.H., Ji, M., Song, K., Moon, I.c.: Unknown-aware domain adversarial learning for open-set domain adaptation. In: Adv. Neural Inform. Process. Syst. pp. 16755–16767 (2022)
18. Kundu, J.N., Bhambri, S., Kulkarni, A.R., Sarkar, H., Jampani, V., R, V.B.: Subsidiary prototype alignment for universal domain adaptation. In: Adv. Neural Inform. Process. Syst. pp. 29649–29662 (2022)
19. Kundu, J.N., Venkat, N., Babu, R.V., et al.: Universal source-free domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4544–4553 (2020)

20. Li, G., Kang, G., Zhu, Y., Wei, Y., Yang, Y.: Domain consensus clustering for universal domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9757–9766 (2021)
21. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: Int. Conf. Mach. Learn. pp. 6028–6039 (2020)
22. Liang, J., Hu, D., Feng, J., He, R.: Umad: Universal model adaptation under domain and category shift (2021)
23. Liang, J., Wang, Y., Hu, D., He, R., Feng, J.: A balanced and uncertainty-aware approach for partial domain adaptation. In: Eur. Conf. Comput. Vis. pp. 123–140 (2020)
24. Lin, Z., Yu, S., Kuang, Z., Pathak, D., Ramanan, D.: Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19325–19337 (2023)
25. Liu, H., Cao, Z., Long, M., Wang, J., Yang, Q.: Separate to adapt: Open set domain adaptation via progressive separation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2927–2936 (2019)
26. Liu, X., Zhou, Y., Zhou, T., Qin, J.: Self-paced learning for open-set domain adaptation. Journal of Computer Research and Development **60**(8), 1711–1726 (2023)
27. MacQueen, J.: Classification and analysis of multivariate observations. In: Berkeley Symp. Math. Statist. Probability. pp. 281–297 (1967)
28. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Int. Conf. Comput. Vis. pp. 1406–1415 (2019)
29. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge (2017)
30. Qu, S., Zou, T., Röhrbein, F., Lu, C., Chen, G., Tao, D., Jiang, C.: Upcycling models under domain and category shift. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20019–20028 (2023)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
32. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
33. Roy, S., Siarohin, A., Sangineto, E., Bulo, S.R., Sebe, N., Ricci, E.: Unsupervised domain adaptation using feature-whitening and consensus loss. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9471–9480 (2019)
34. Saito, K., Kim, D., Sclaroff, S., Saenko, K.: Universal domain adaptation through self supervision. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 16282–16292 (2020)
35. Saito, K., Saenko, K.: Ovanet: One-vs-all network for universal domain adaptation. In: Int. Conf. Comput. Vis. pp. 9000–9009 (2021)
36. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: Eur. Conf. Comput. Vis. pp. 153–168 (2018)
37. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27**(3), 379–423 (1948)
38. Shen, M., Lu, Y., Hu, Y., Ma, A.J.: Collaborative learning of diverse experts for source-free universal domain adaptation. In: ACM MM. p. 2054–2065 (2023)
39. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Adv. Neural Inform. Process. Syst. (2017)

40. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5018–5027 (2017)
41. Wang, J., Chen, Y., Feng, W., Yu, H., Huang, M., Yang, Q.: Transfer learning with dynamic distribution adaptation. ACM Trans. Intell. Syst. Technol. **11**(1), 6:1–6:25 (2020)
42. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7959–7971 (2022)
43. Zheng, Z., Yang, Y.: Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int. J. Comput. Vis. **129**(4), 1106–1120 (2021)
44. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vis. **130**(9), 2337–2348 (2022)
45. Zhou, Y., Bai, S., Zhou, T., Zhang, Y., Fu, H.: Delving into local features for open-set domain adaptation in fundus image analysis. In: MICCAI. pp. 682–692 (2022)
46. Zhu, D., Li, Y., Yuan, J., Li, Z., Kuang, K., Wu, C.: Universal domain adaptation via compressive attention matching. In: Int. Conf. Comput. Vis. pp. 6974–6985 (2023)