Beyond Retraining: Training-Free Unknown Class Filtering for Source-Free Open Set Domain Adaptation of Vision–Language Models

Yongguang Li[1], Jindong Li[4], Qi Wang[1,2*], QianLi Xing[3], Runliang Niu[1], Shengsheng Wang[3], Menglin Yang[4]

[1]School of Artificial Intelligence, Jilin University
[2]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China
[3]College of Computer Science and Technology, Jilin University
[4]The Hong Kong University of Science and Technology (Guangzhou)
[1,2,3] {ygli25, niurl19}@mails.jlu.edu.cn, {qiwang, qianlixing, wss}@jlu.edu.cn
[4]jli839@connect.hkust-gz.edu.cn, menglin.yang@outlook.com

Abstract

Vision-language models (VLMs) have gained widespread attention for their strong zero-shot capabilities across numerous downstream tasks. However, these models assume that each test image's class label is drawn from a predefined label set and lack a reliable mechanism to reject samples from emerging unknown classes when only unlabeled data are available. To address this gap, open-set domain adaptation methods retrain models to push potential unknowns away from known clusters. Yet, some unknown samples remain stably anchored to specific known classes in the VLM feature space due to semantic relevance, which is termed as *Semantic Affinity Anchoring (SAA)*. Forcibly repelling these samples unavoidably distorts the native geometry of VLMs and degrades performance. Meanwhile, existing score-based unknown detectors use simplistic thresholds and suffer from threshold sensitivity, resulting in sub-optimal performance. To address aforementioned issues, we propose VLM-OpenXpert, which comprises two training-free, plug-and-play inference modules. SUFF performs SVD on high-confidence unknowns to extract a low-rank "unknown subspace". Each sample's projection onto this subspace is weighted and softly removed from its feature, suppressing unknown components while preserving semantics. BGAT corrects score skewness via a Box–Cox transform, then fits a bimodal Gaussian mixture to adaptively estimate the optimal threshold balancing known-class recognition and unknown-class rejection. Experiments on 9 benchmarks and three backbones (CLIP, SigLIP, ALIGN) under Source-Free OSDA settings show that our training-free pipeline matches or outperforms retraining-heavy state-of-the-art methods, establishing a powerful lightweight inference calibration paradigm for open-set VLM deployment.

Introduction

Recent advances in large-scale vision-language models (VLMs) such as CLIP (Radford et al. 2021) have demonstrated powerful zero-shot capabilities across diverse downstream tasks without task-specific fine-tuning, making them foundational for real-world applications (Zhang et al. 2024). However, such zero-shot predictions inherently assume a closed-world scenario, where test samples must belong to a predefined label space. This assumption is frequently violated as unlabeled classes inevitably emerge in real-world scenarios (Yu, Irie, and Aizawa 2025). Misclassifying these unknown samples as known classes can lead to severe consequences. For instance, autonomous driving systems may misclassify unseen construction machinery as normal vehicles, posing serious safety hazards (Yang et al. 2024). Moreover, real-world deployments typically lack labeled data and involve significant domain shifts, presenting critical challenges for enhancing unknown-class rejection in VLMs.

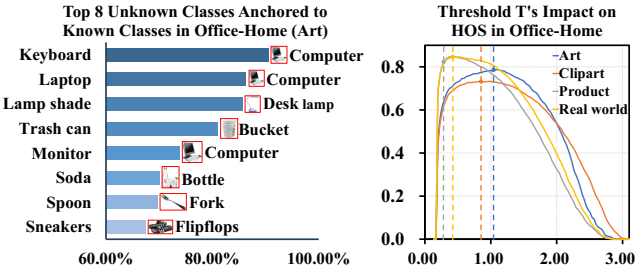


Figure 1: Analysis of CLIP on Office-Home. Left: top-8 unknown classes with the highest anchor rates and their anchored known classes. Right: HOS (%) on four domains versus threshold; the dashed line marks the optimum.

Against this backdrop, models must retain discriminative power on known classes while reliably rejecting unknown samples using only unlabeled target data. This challenge has given rise to Open-Set Unsupervised Domain Adaptation (OSDA) and its variant, Source-Free OSDA (SF-OSDA) (Fang et al. 2020). Conventional single-modal OSDA methods jointly train on labeled source and unlabeled target data to transfer knowledge and enhance unknown-class detection. The general idea of such methods is to enforce a repulsion loss that pushes potential unknown samples away from known-class clusters (Liu et al. 2019). In contrast, SF-OSDA dispenses with source data at adaptation time, relying solely on a source-pretrained model while still applying the same separation objective (Wan et al.

*Corresponding author.

2024). In the VLM era, many works have directly transplanted this "retraining + separation" paradigm into vision–language settings by using VLM backbones, aiming to establish clear decision boundaries between known and unknown classes (Liang et al. 2024; Min et al. 2023).

However, this paradigm implicitly relies on the global separability assumption, that unknown classes should be semantically distant from all known classes. Due to inherent semantic affinity, many unknown classes naturally anchor to specific known-class clusters. We term this phenomenon as *Semantic Affinity Anchoring (SAA)*. For example, in Office-Home (Venkateswara et al. 2017), nearly all samples of the unknown class "keyboard" fall into the known "computer" cluster (Figure 1(left)), and over 50% of unknown classes exhibit SAA in CLIP. As SAA reflects the true semantic relationships captured by vision–language pretraining, naively pushing anchored samples away risks distorting the model's semantic structure and harming performance. In addition to the geometric mismatch, VLM-based unknown sample detection also suffers from threshold sensitivity issue caused by existing score-based strategies. As shown in Figure 1(right), optimal thresholds vary significantly across datasets. Prior approaches often adopt simplistic strategies such as fixed values (Yu, Irie, and Aizawa 2025), mean scores (Bucci, Loghmani, and Tommasi 2020), k-means clustering (Wan et al. 2024), or heuristic rules, which fail to model score distributions effectively and lead to suboptimal performance.

To address such issues, we propose VLM-OpenXpert, a training-free and label-free source-free OSDA framework of vision–language models, featuring two plug-and-play modules. Specifically, the first module, SUFF (SVD-Based Unknown-Class Feature Filtering), which adopts a directional filtering strategy instead of global repulsion. It begins by selecting high-confidence unknown samples using an uncertainty-based criterion, explicitly including those hard-to-detect cases affected by Semantic Affinity Anchoring (SAA). Then, it performs SVD (Eckart and Young 1936) on the centralized features of these samples to construct a low-rank subspace, where unknowns exhibit higher projection energy than knowns. Finally, all samples are softly attenuated based on their projection in this subspace, effectively suppressing unknown-class features while preserving the semantic geometry of the model and mitigating the effects of SAA.

Next, we propose BGAT (Box-Cox GMM-Based Adaptive Thresholding) to replace brittle fixed or heuristic thresholds. BGAT first applies a Box-Cox transformation on confidence scores to stabilize scale and skewness, then fits a bimodal Gaussian mixture in the transformed space. The midpoint between the two modes yields a near-optimal, data-adaptive threshold. These two modules augment VLMs with reliable open-set detection capabilities at inference time with minimal computational overhead and without requiring any labeled samples or retraining. Experimental results demonstrate that our VLM-OpenXpert, despite being entirely training-free, outperforms representative training-intensive OSDA methods. Also, it can serve as a plug-and-play enhancement to boost the performance of existing trainable Source-Free OSDA methods. **Our main contributions are summarized as follows**:

- We identify and validate ***Semantic Affinity Anchoring (SAA)*** phenomenon, where unknown samples being semantically attracted to known classes. We further show that excessively pushing unknowns away from knowns can distort VLM's inherently generalized feature space, thereby degrading its performance.
- We propose SUFF, a training-free and plug-and-play module that suppresses unknown-class features via selective filtering, alleviating SAA while preserving VLM semantic geometry. Also, we introduce BGAT, a lightweight adaptive thresholding strategy that estimates near-optimal thresholds, enhancing the discriminative ability of VLM on identifying known classes while reliably rejecting unknown samples.
- Experiments on 9 benchmarks and 3 VLM backbones (e.g., CLIP, SigLIP, ALIGN) show that our training-free method matches or outperforms training-heavy representative OSDA methods, establishing a strong training-free baseline for VLM-based open-set domain adaptation.

Related Work

Vision-language Models. Vision–Language Models (VLMs) (Zhang et al. 2024) learn a shared embedding space for images and texts by contrastively aligning large-scale image–caption pairs, enabling simple image–text similarity to power a wide range of downstream tasks without task-specific heads. Representative models such as CLIP, SigLIP, and ALIGN normally adopt a twin-tower architecture with separate vision and text encoders. Such models are trained on massive web-scale datasets and support zero-shot inference via promptable text queries. While this design offers strong cross-task generalization, it lacks a built-in mechanism to reject unseen classes at test time (Wang et al. 2023). To address this limitation, we propose an inference-time calibration procedure that can be seamlessly integrated into any frozen VLM. Our method requires no labels or additional training and equips the model with robust open-set recognition capability, enabling reliable rejection of unknown samples during deployment.

Unsupervised Open-Set Domain Adaptation. OSDA (Panareda Busto and Gall 2017; Li et al. 2025) aims to mitigate the decline in model performance caused by distribution discrepancies between the source and target domains, while assuming that the target domain contains unknown categories absent in the source domain. Traditional OSDA methods (Liu et al. 2019; Bucci, Loghmani, and Tommasi 2020) typically rely on training with both source and target domain data concurrently. However, practical applications may restrict direct access to source domain data during target domain training due to privacy concerns, thereby motivating the development of Source-Free Unsupervised Domain Adaptation (SF-OSDA) (Luo et al. 2023; Qu et al. 2024). Most existing SF-OSDA (Luo et al. 2023; Qu et al. 2024) approaches substitute source domain samples with models pre-trained on the source domain for target domain training. In recent years, capitalizing on

CLIP's remarkable zero-shot transfer capabilities, some SF-OSDA methods have begun to integrate CLIP with a set of known class names to train directly on unlabeled target domain data (Min et al. 2023; Liang et al. 2024).

Methodology

Given an unlabeled target dataset $D_t = \{x_i^t\}_{i=1}^{N_t}$ and a set of known categories $L_t = \{l_c\}_{c=1}^{C}$, where l_c represents the category name of the c-th class, some samples in D_t belong to L_t, while others do not. Our goal is to perform inference on D_t by assigning samples belonging to L_t to their respective known classes and classifying the rest as an unknown class $(C+1)$. The model's performance is evaluated on D_t.

As shown in Figure 2, VLM-OpenXpert uses a frozen VLM backbone for zero-shot inference to produce known-class predictions and compute per-sample scores. It then applies BGAT to estimate a threshold for identifying representative samples to construct the SUFF module, suppresses unknown-class features via SUFF, recomputes scores, and uses BGAT again to determine the threshold for final detection of unknown-class samples.

Box-Cox GMM-Based Adaptive Thresholding (BGAT) Module

Traditional open-set domain adaptation methods typically employ fixed thresholds (e.g., Max/2) (Yu, Irie, and Aizawa 2025) or global means (Bucci, Loghmani, and Tommasi 2020; Liu et al. 2019) to distinguish unknown-class samples, which lack theoretical foundation and oversimplify real-world distributions. To investigate a more universal and adaptive threshold calibration strategy, motivated by previous work (Jahan and Savakis 2024), we introduce a statistically grounded approach combining Box-Cox transformation (Box and Cox 1964) and Gaussian Mixture Models (GMM) (Bishop and Nasrabadi 2006).

Specifically, as shown in Figure 2, given an unlabeled target dataset D_t and a set of known class names $L_t = \{l_1, l_2, \ldots, l_C\}$, we perform zero-shot classification on all samples in D_t using a VLM (e.g., CLIP). The classification probabilities for known classes and the detection of unknown samples are computed as follows:

$$\hat{y}_{i,c} = \frac{\exp\left(\cos\left(X_{all}[i], X_{text}[c]\right)/t\right)}{\sum_{c'=1}^{C} \exp\left(\cos\left(X_{all}[i], X_{text}[c']\right)/t\right)}, \quad (1)$$

$$\text{Pred}(x_i) = \begin{cases} C+1, & \text{if Score}(\hat{y}_i) > T^*, \\ \arg\max_c \hat{y}_{i,c}, & \text{otherwise.} \end{cases} \quad (2)$$

In particular, we define:

$$X_{all} = \{E(x_i) \mid x_i \in D_t\}, \quad X_{text} = \{G(p(l_c)) \mid l_c \in L_t\}, \quad (3)$$

where E denotes the image encoder, G denotes the text encoder, and $p(l_c)$ is the class description sentence with the default prompt "a photo of a". The threshold is denoted by T^*, and we adopt entropy as the default sample score:

$$\text{S}(x_i) = H(\hat{y}_i) = -\sum_{c=1}^{C} \hat{y}_{i,c} \log(\hat{y}_{i,c}). \quad (4)$$

We assume that the score distributions for known-class and unknown-class samples follow Gaussian distributions with different means and variances. Under the common assumption of symmetric costs and comparable priors, the optimal threshold T^* admits the well-known closed form given by the intersection of the two densities. However, there are two issues in practical datasets that need to be addressed:

(i) *Distribution Skewness*: Taking the entropy of the samples as an example, the entropy of known classes is concentrated near 0 due to the high certainty of most known-class samples. This results in a significant left skew in distribution, violating the assumption of a Gaussian distribution; (ii) *Error in GMM Estimation*: During the GMM estimation process, errors arise, particularly when the entropy of the samples is highly concentrated, the variance estimation for known classes becomes notably inaccurate, significantly reducing the accuracy of intersection points.

Addressing Issue (i): We apply the Box–Cox (Box and Cox 1964) transformation; to ensure positivity, add a tiny $\epsilon > 0$ to the scores (still denoted S_i). The transform is:

$$S_i^{(\lambda)} = \begin{cases} \dfrac{S_i^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log S_i, & \lambda = 0, \end{cases} \quad (5)$$

where the optimized parameter λ^* is estimated using maximum likelihood (Fisher 1921) as:

$$\lambda^* = \arg\max_{\lambda} \sum_{i=1}^{N} \log \mathcal{N}\big(S_i^{(\lambda)};\, \mu_\lambda, \sigma_\lambda^2\big) \;+\; (\lambda - 1) \sum_{i=1}^{N} \log S_i. \quad (6)$$

Since this data transformation is monotonically increasing, it does not alter the relative ordering of the sample scores. However, it can alleviate the issue of sample scores being too close to known classes, thereby improving the accuracy of mean and variance estimation.

Addressing Issue (ii): After Box-Cox transformation, we fit a two-component GMM to $\{S_i^{(\lambda^*)}\}$ as:

$$p\Big(S^{(\lambda^*)}\Big) = \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2). \quad (7)$$

Parameters $\{\pi_k, \mu_k, \sigma_k\}_{k=1}^{2}$ are estimated via EM (Dempster, Laird, and Rubin 1977) and sorted as $(\mu_{\text{low}}, \sigma_{\text{known}}), (\mu_{\text{high}}, \sigma_{\text{unk}})$. However, the GMM estimation may introduce errors, particularly the underestimation of the known-class variance σ_{know}. Since the intersection point is highly sensitive to the variance , it may not be the optimal choice in practice.

To address this issue, we adopt a compromise approach by selecting the midpoint of the means, $T^*_{trans} = \frac{\mu_{know} + \mu_{unk}}{2}$, as the threshold. This choice is motivated by the fact that the mean estimated by GMM is less sensitive to errors than the variance, and using the midpoint can enhance the robustness of the threshold. Finally, the optimal threshold T^* is derived via inverse Box-Cox transformation as:

$$\begin{cases} \left(\lambda^* T^*_{\text{trans}} + 1\right)^{1/\lambda^*}, & \lambda^* \neq 0, \\ \exp\left(T^*_{\text{trans}}\right), & \lambda^* = 0. \end{cases} \quad (8)$$

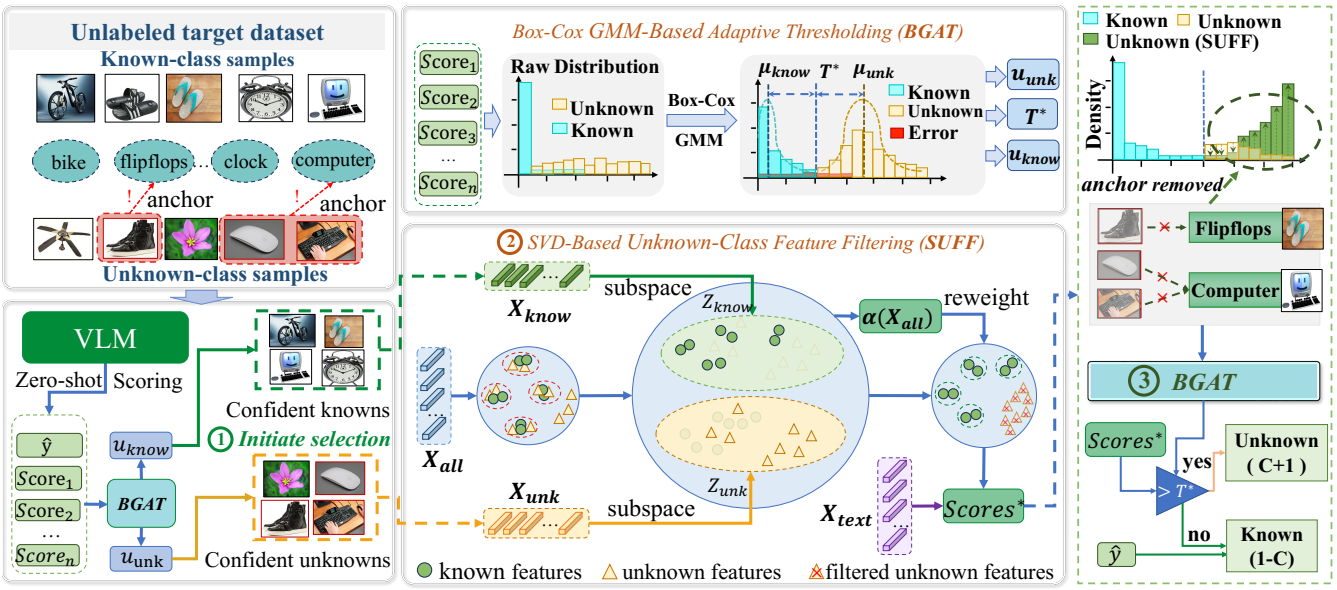


Figure 2: Overview of VLM-OpenXpert, which performs lightweight, training-free inference on unlabeled target data via entropy-based scoring, unknown-class feature filtering (SUFF), and adaptive thresholding (BGAT).

In addition, the estimated means $(\mu_{\text{know}}, \mu_{\text{unk}})$ of the known and unknown classes also need to be transformed back to the space before the Box-Cox transformation.

SVD-Based Unknown-Class Feature Filtering (SUFF) Module

Although the BGAT module estimates the optimal threshold T^*, the SAA phenomenon still causes some unknown-class samples to have low entropy and be misclassified as known-class samples. To resolve this, we propose an SVD-based Unknown Feature Filtering Module (SUFF), which filters out unknown-class features from X_{all} to mitigate the SAA phenomenon.

Specifically, we use the known-class mean μ_{know}, along with the unknown-class mean μ_{unk} and threshold T^*, both estimated by the BGAT module from VLM zero-shot scores, to filter high-confidence known- and unknown-class sample features. We filter samples with entropy less than or equal to $\frac{\mu_{\text{know}}+T^*}{2}$, forming the high-confidence known-class sample feature set $X_{\text{know}} \in \mathbb{R}^{N_{\text{know}} \times D}$; and we filter samples with entropy greater than $\frac{\mu_{\text{unk}}+T^*}{2}$, forming the high-confidence unknown-class sample feature set $X_{\text{unk}} \in \mathbb{R}^{N_{\text{unk}} \times D}$.

Next, we perform SVD on both known-class and unknown-class feature sets to construct their principal component subspaces as:

$$X_{\text{know}} = U_{\text{know}} \Sigma_{\text{know}} V_{\text{know}}^{\top}, \tag{9}$$

$$X_{\text{unk}} = U_{\text{unk}} \Sigma_{\text{unk}} V_{\text{unk}}^{\top}, \tag{10}$$

where $\Sigma_{\text{know}} = \text{diag}(\sigma_1^{\text{know}}, \ldots, \sigma_r^{\text{know}})$ and $\Sigma_{\text{unk}} = \text{diag}(\sigma_1^{\text{unk}}, \ldots, \sigma_r^{\text{unk}})$ are the singular value matrices. $V_{\text{know}} \in \mathbb{R}^{r \times D}$ and $V_{\text{unk}} \in \mathbb{R}^{r \times D}$ are the right singular matrices. The smallest dimension k is selected such that the cumulative variance contribution satisfies the following:

$$\min k \quad \text{s.t.} \quad \frac{\sum_{i=1}^{k} (\sigma_i)^2}{\sum_{j=1}^{r} (\sigma_j)^2} \geq \tau. \tag{11}$$

We retain the top k principal components to construct the projection matrices $W_{\text{know}} = V_{\text{know}}[:k, :]^{\top}$ and $W_{\text{unk}} = V_{\text{unk}}[:k, :]^{\top}$. Then, we project all sample features $X_{\text{all}} \in \mathbb{R}^{N \times D}$ into the constructed known class space Z_{know} and unknown class space Z_{unk} as follows:

$$X_{\text{all}}^{\text{know}} = Z_{know}(X_{\text{all}}) = X_{\text{all}} W_{\text{know}} W_{\text{know}}^{\top}, \tag{12}$$

$$X_{\text{all}}^{\text{unk}} = Z_{unk}(X_{\text{all}}) = X_{\text{all}} W_{\text{unk}} W_{\text{unk}}^{\top}. \tag{13}$$

Next, we analyze the proportion $\alpha(X_{\text{all}})$ of unknown class features in the samples based on the similarity $S(x_{\text{all}})$ before and after normalizing the sample feature mapping:

$$s_{\text{know}}(X_{all}) = \frac{X_{all} \cdot X_{\text{all}}^{\text{know}}}{\|X_{all}\| \|X_{\text{all}}^{\text{know}}\|}, \tag{14}$$

$$s_{\text{unk}}(X_{all}) = \frac{X_{all} \cdot X_{\text{all}}^{\text{unk}}}{\|X_{all}\| \|X_{\text{all}}^{\text{unk}}\|}, \tag{15}$$

$$\alpha(X_{all}) = \frac{\exp(s_{\text{unk}}(X_{all})/t)}{\exp(s_{\text{know}}(X_{all})/t) + \exp(s_{\text{unk}}(X_{all})/t)}, \tag{16}$$

where t is temperature coefficient. t is set to 1.0 by default and it controls the smoothness of probability distribution.

To remove the unknown-class features from the sample features, we perform a weighted subtraction operation:

$$X_{all}^{*} = X_{all} - \alpha(X_{all}) X_{\text{all}}^{\text{unk}}. \tag{17}$$

Finally, we recompute sample scores using the filtered features X_{all}^{*} and X_{text}, and re-estimate the threshold T^* via BGAT. Samples scoring above T^* are marked as unknown, while others retain their original predictions from Eq. 1, preserving known-class recognition and mitigating filtering-induced errors.

Method	E	VLM	TF	OfficeHome												AVG	VisDA
				C	P	R	A	P	R	A	C	R	A	C	P		Syn
				A			C			P			R				Real
OSDA (trains jointly on labeled source data and unlabeled target data)																	
USD	RN	×	×	60.1	62.6	67.8	61.1	56.3	59.1	70.0	65.2	71.1	76.3	68.9	56.3	64.6	69.4
OSMDA-CLIP	Vit-B-32	✓	×	74.1	74.7	76.6	71.9	69.7	70.4	81.9	82.2	81.8	83.8	84.0	69.7	76.7	83.4
UniOT-CLIP	Vit-B-16	✓	×	79.2	71.8	77.1	77.8	72.3	76.0	83.3	83.1	87.3	86.3	84.1	84.7	80.3	-
GLC-CLIP	Vit-B-16	✓	×	55.5	52.1	69.0	72.1	65.9	71.3	79.7	81.3	83.9	83.3	77.9	65.9	71.5	83.4
SF-OSDA (uses a model trained on labeled source data and performs unsupervised training on unlabeled target data)																	
LEAD	RN	×	×	61.0	65.5	64.8	60.7	59.8	57.7	70.8	68.6	75.8	76.5	70.8	59.8	66.0	74.2
DTDE	RN	×	×	-	-	-	-	-	-	-	-	-	-	-	-	70.3	80.4
(Co-learn++)-CLIP	RN	✓	×	60.4	56.0	64.0	54.9	51.0	58.8	77.6	72.7	83.6	78.4	75.9	51.0	65.4	-
COSDA	RN	×	×	65.0	67.8	70.3	70.5	63.7	64.1	74.9	72.2	79.1	79.9	76.1	63.7	70.6	-
UPUK	RN	×	×	66.4	67.6	67.8	55.8	55.1	59.4	76.7	73.1	74.4	78.4	77.6	55.1	67.3	72.3
COCA-CLIP	Vit-B-16	✓	×	79.7	79.6	80.0	75.6	74.5	75.7	84.5	84.3	84.4	82.5	82.5	74.5	79.8	86.3
COSDA-CLIP	Vit-B-32	✓	×	-	-	-	-	-	-	-	-	-	-	-	-	-	78.4
DIFO-CLIP	Vit-B-32	✓	×	68.2	67.2	71.9	64.5	62.1	65.3	86.2	79.3	84.4	87.9	86.1	62.1	73.8	-
Target-only Adaptation (adapts directly on the unlabeled target domain; notably, only our method performs inference without any training)																	
UOTA-CLIP	Vit-B-32	✓	×	75.9			75.1			84.2			86.2			80.4	85.3
UEO-CLIP	Vit-B-16	✓	×	72.1			71.1			79.5			79.5			75.6	82.6
ours-CLIP	RN	✓	✓	73.5			61.9			81.9			81.6			74.7	85.1
ours-CLIP	Vit-B-32	✓	✓	75.1			72.8			85.1			87.2			80.0	87.6
ours-CLIP	Vit-B-16	✓	✓	78.2			74.0			86.8			85.7			81.2	**89.0**
ours-SigLIP	Vit-B-16	✓	✓	**85.8**			82.9			90.0			89.2			87.0	88.3
ours-ALIGN	EfficientNet-B7	✓	✓	85.0			**83.2**			**90.9**			**91.3**			**87.6**	87.3

Table 1: HOS (%) on Office-Home and VisDA-2017. Bold denotes the best results and underline denotes the second best. For Office-Home, each column pair uses the upper domain as source and the lower as target. "-CLIP", "-SigLIP", and "-ALIGN" represent using the respective VLM backbone. (E: Image Encoder; TF: Training-Free).

Experiments

Experimental Settings

Datasets. We evaluate our method under the SF-OSDA experimental setting on three representative benchmark datasets: Office-Home (Venkateswara et al. 2017) (65 categories, 4 domains), VisDA-2017 (Peng et al. 2017) (12 categories, synthetic-to-real adaptation), and DomainNet (Peng et al. 2019) (345 categories, 6 domains). To further test the generalization ability of our model, we select 6 datasets from the Visual Task Adaptation Benchmark (VTAB-6).

Baselines. We select three categories of baseline methods for comparison: **OSDA/SF-OSDA methods** jointly train on source data (or a source-trained model) and unlabeled target data, including USD (Jahan and Savakis 2024), LEAD (Qu et al. 2024), DIFO (Tang et al. 2024), OSMDA (Liu et al. 2025), UniOT (Chang et al. 2022), GLC (Qu et al. 2023), COCA (Liu and Zhou 2024), DTDE (Yu et al. 2025), Co-learn++ (Zhang, Shen, and Foo 2025), COSDA (Wang et al. 2025), UPUK (Wan et al. 2024); **Unsupervised target-only adaptation methods** such as UOTA (Min et al. 2023) and UEO (Liang et al. 2024), which perform training and unknown class detection directly on the target domain; **CLIP-based zero-shot OOD methods** using Maximum Concept Matching (MCM) (Ming et al. 2022).

Implementation Details. We evaluate on three VLM backbones—CLIP (Radford et al. 2021), SigLIP (Zhai et al. 2023), and ALIGN (Jia et al. 2021). Image encoders follow the table headers; when unspecified, we default to ViT-B/16 (Han et al. 2022) for CLIP and SigLIP, and EfficientNet-B7 (Tan and Le 2019) for ALIGN. Unless otherwise noted, we set $\tau = 0.8$ and use a batch size of 32.

More details can be found in the extended version available online.

Method	E	VTAB-6	DomainNet	AVG
MCM-CLIP	ViT-L/14	59.9	67.8	63.9
MCM-SigLIP	ViT-B/16	66.3	66.5	66.4
MCM-ALIGN	EfficientNet-B7	59.3	66.6	62.9
UOTA-CLIP	ViT-L/14	-	71.1	-
UEO-CLIP	ViT-L/14	57	69.6	63.3
ours-CLIP	ViT-L/14	65.4	**72.9**	69.1
ours-SigLIP	ViT-B/16	**70.2**	70.8	**70.5**
ours-ALIGN	EfficientNet-B7	62.2	66.9	64.5

Table 2: HOS (%) on DomainNet and six visual task adaptation datasets.

Overall Performance

As shown in Table 1 and Table 2, our training-free approach surpasses previous OSDA, SF-OSDA, and target-only adaptation methods. Using the same backbone on Office-Home, it outperforms GLC, DIFO, and UEO by 9.7%, 7.4%, and 5.6%, respectively. Compared with the state-of-the-art UOTA, it achieves gains of 3.7% on VisDA-2017 and 1.8% on DomainNet, while matching UOTA on Office-Home. These results confirm the effectiveness of our method and show that an efficient, training-free inference strategy can

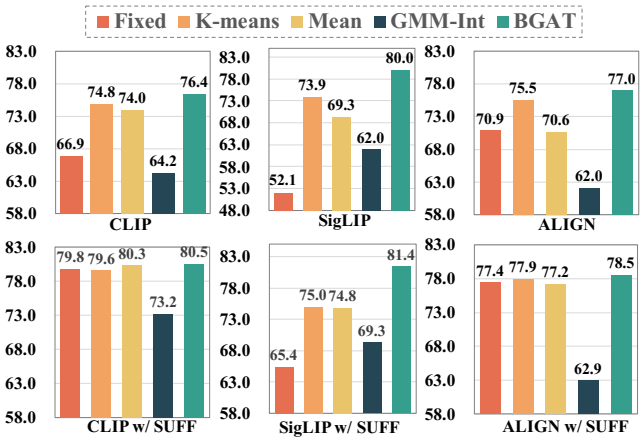


Figure 3: HOS(%) performance variation under different thresholding methods.

unlock the inherent open-set zero-shot capability of VLMs on unlabeled target data. Moreover, our approach consistently outperforms MCM on three different VLM backbones across all six Visual Task Adaptation benchmarks, demonstrating the robust improvements in unknown-class detection. Notably, our method adds only 2.5% inference time compared to direct zero-shot use of VLMs, while being tens of times faster than conventional training-intensive methods.

Ablation Study

Analysis of the SAA Phenomenon. To quantitatively validate how much unknown samples concentrate around a single known class, we propose the Anchor Rate at 1 (e.g., AR@1) metric, which is defined as the fraction of unknown classes whose strongest semantic affinity is with one known class (e.g., range 0–1). As shown in Table 3, all three VLMs exhibit high AR@1 values (0.46–0.58) on the widely used benchmarks, indicating that about half of unknown classes in these benchmarks anchor to specific known-class clusters, thereby confirming the widespread presence of the SAA phenomenon.

Effectiveness of SUFF. As shown in Table 3, SUFF reduces AR@1 by 0.23, 0.28, and 0.33 on CLIP, SigLIP, and ALIGN, respectively, directly demonstrating substantial mitigation of semantic affinity anchoring (SAA). This reduction translates into average HOS gains of 7.7%, 6.9%, and 4.3% across all thresholding strategies (Table 4), confirming improved known/unknown separability. Figures 4 and 6 further illustrate SUFF's dual effect: it separates unknown classes while preserving known-class separability and markedly reduces their clustering around specific known classes.

Effectiveness of BGAT. We compare five thresholding strategies across six tasks using three VLMs: a fixed threshold at half the maximum score (Fixed) (Yu, Irie, and Aizawa 2025), the mean of all scores (Mean) (Bucci, Loghmani, and Tommasi 2020), k-means clustering (K-Means) (Liang, Hu, and Feng 2020), the intersection point of two Gaussian densities (GMM-Int) (He, Wang, and Zhang 2025), and our proposed BGAT method. As shown in Figure 3, BGAT consistently achieves the best performance across all VLMs, both on raw features and SUFF-enhanced features, demonstrating its robustness and adaptability to feature variations.

Method	VTAB-6	Office-Home	VisDA-2017	AVG↓
CLIP	0.53	0.53	0.33	0.46
CLIP w/ SUFF	**0.44**	**0.24**	**0.00**	**0.23**(↓ 0.23)
SigLIP	0.56	0.67	0.50	0.58
SigLIP w/ SUFF	**0.32**	**0.24**	**0.33**	**0.30**(↓ 0.28)
ALIGN	0.57	0.54	0.50	0.54
ALIGN w/ SUFF	**0.32**	**0.14**	**0.17**	**0.21**(↓ 0.33)

Table 3: AR@1 on VTAB-6, Office-Home and VisDA-2017 for baseline vs. SUFF-augmented models (↓ lower is better.)

VLM	w/ SUFF	Fixed	k-means	BGAT	AVG
CLIP	×	66.9	74.0	76.4	72.4
	✓	**79.8**(+12.9%)	**80.3**(+6.2%)	**80.5**(+4.1%)	**80.1**(+7.7%)
SigLIP	×	52.1	69.3	80.0	67.1
	✓	**65.4**(+13.3%)	**74.8**(+5.5%)	**81.4**(+1.4%)	**73.9**(+6.9%)
ALIGN	×	70.9	70.6	77.0	72.8
	✓	**77.4**(+6.5%)	**77.2**(+6.6%)	**78.5**(+1.5%)	**77.7**(+4.3%)

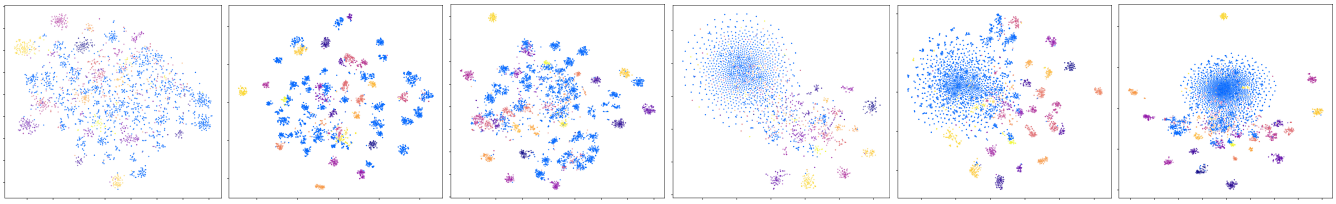
Table 4: HOS(%) before vs. after SUFF across different VLM backbones and threshold strategies.

Method	Art	Clipart	Product	Real	AVG
SHOT	58.1	50.8	59.9	62.3	57.8
w/ BGAT	59.2	54.8	65.7	68.8	62.1(+4.3%)
w/ BGAT + SUFF	**59.3**	**55.3**	**66.5**	**70.5**	**62.9**(+5.1%)
UEO-CLIP	68.3	63.7	65.6	70.4	67.0
w/ BGAT	72.1	71.1	79.5	80.7	75.9(+8.9%)
w/ BGAT + SUFF	**74.8**	**73.5**	**83.6**	**83.3**	**78.8**(+11.8%)

Table 5: HOS (%) for SHOT and UEO-CLIP enhanced with our modules on the four Office-Home subdomains.

Plug-and-Play Validation on UEO and SHOT. Table 5 demonstrates the effectiveness of our proposed BGAT and SUFF modules when integrated into the VLM-based method UEO and the representative single-modal SF-OSDA method SHOT (Liang, Hu, and Feng 2020). As shown in Table 5, adding BGAT improves UEO and SHOT by 8.9% and 4.3% in terms of HOS; with SUFF added, these gains rise to 11.8% and 5.1%. This confirms the generality of our approach across both multimodal and single-modal methods.

Effectiveness of Box-Cox. As shown in Figure 5(a–c), applying the Box–Cox transformation in BGAT consistently improves performance across five tasks and three VLM backbones, confirming its effectiveness in improving threshold estimation and overall accuracy. On average, the three VLMs gain 2.67%, with SigLIP benefiting the most due to its more concentrated known-class scores, which make Box-Cox more effective.



(a) Art (b) Product (c) Real (d) Art-S (e) Product-S (f) Real-S

Figure 4: t-SNE Visualizations of four domains in the Office-Home datasets. (a)-(c) represent the raw features obtained from CLIP for target dataset image samples, while (d)-(f) represent the features filtered by the SUFF module. Blue points indicate unknown class samples, and the points in other colors represent different known classes.

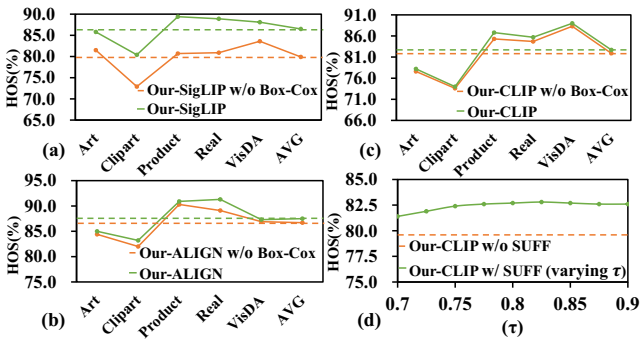
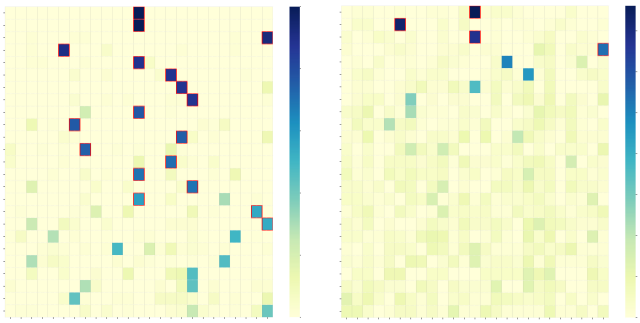


Figure 5: Effect of the Box-Cox transformation and SUFF on HOS (%) for Office-Home and VisDA datasets. Panels (a–c) show the impact of Box-Cox transformation and panel (d) shows the effect of varying τ.



(a) (b)

Figure 6: Unknown-class heat map in the Real-World domain of the Office-Home dataset for the ALIGN model: (a) before SUFF processing and (b) after SUFF processing. Darker colors indicate stronger tendencies.

Hyper-parameter Analysis

We analyze the effect of the variance retention ratio (τ) used in subspace construction (Eq. 11). Since the singular value spectrum decays rapidly and then flattens, small τ values may discard useful information, while large τ values tend to retain noise or redundancy. As a result, extreme values offer no meaningful benefit, and we explore a practical range of [0.7, 0.9]. As shown in Figure 5 (d), the average HOS remains consistently stable across five tasks, with a variance of just 0.19 $(\%)^2$, indicating that model performance is largely insensitive to the choice of τ within this range.

Case Study

Visualization of Semantic Affinity Anchoring. Figure 6 shows heat maps of the anchoring of unknown to known classes (a) before SUFF module and (b) after SUFF module, with anchored unknowns dropping from 18 to 4. This shows SUFF greatly reduces semantic affinity anchoring from unknowns to known classes. The few remaining anchors result from the limits in VLM's initial sample selection, making it difficult for the unknown-class subspace to cover the feature distribution of these unknown samples.

t-SNE Visualization. Figure 4 (a–c) shows the original VLM feature distribution, where unknown-class features are mixed with known-class clusters and hard to distinguish. After applying SUFF (Figure 4 (d–f)), unknown features are clearly separated, and the separability among known classes is also preserved. This confirms that SUFF improves known/unknown discrimination without harming known-class structure.

Conclusion

This work identifies and formally defines *Semantic Affinity Anchoring (SAA)*, a phenomenon common to many benchmark datasets. In SAA, certain unknown-class samples cluster near semantically related known classes in the VLM feature space, which makes them hard to detect. To tackle this problem, we propose VLM-OpenXpert, a completely training-free and label-free inference framework consisting of two plug-and-play modules. Specifically, SUFF filters unknown-class features through soft suppression in a low-rank subspace dominated by unknown-class feature energy, significantly reducing SAA-induced detection failures. BGAT combines a Box–Cox transformation with a bimodal Gaussian mixture model to adaptively estimate an optimal threshold that balances known class recognition and unknown class detection. With these two lightweight inference modules, VLM-OpenXpert surpasses or matches training-intensive state-of-the-art methods on unlabeled target domains, validating its effectiveness and establishing a new design paradigm for unsupervised open-set domain adaptation of vision-language models.

Acknowledgements

This work is supported by the National Key Research and Development Program of China(No.2023YFF0905400) and the National Natural Science Foundation of China (No.62206107 and No.62406127).

References

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Box, G. E.; and Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2): 211–243.

Bucci, S.; Loghmani, M. R.; and Tommasi, T. 2020. On the effectiveness of image rotation for open set domain adaptation. In *European conference on computer vision*, 422–438. Springer.

Chang, W.; Shi, Y.; Tuan, H.; and Wang, J. 2022. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35: 29512–29524.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22.

Eckart, C.; and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218.

Fang, Z.; Lu, J.; Liu, F.; Xuan, J.; and Zhang, G. 2020. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems*, 32(10): 4309–4322.

Fisher, R. A. 1921. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1: 3–32.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

He, W.; Wang, Z.; and Zhang, Y. 2025. Target Semantics Clustering via Text Representations for Robust Universal Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17132–17140.

Jahan, C. S.; and Savakis, A. 2024. Unknown sample discovery for source free open set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1067–1076.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Li, J.; Li, Y.; Fu, Y.; Liu, J.; Liu, Y.; Yang, M.; and King, I. 2025. CLIP-Powered Domain Generalization and Domain Adaptation: A Comprehensive Survey. *arXiv preprint arXiv:2504.14280*.

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

Liang, J.; Sheng, L.; Wang, Z.; He, R.; and Tan, T. 2024. Realistic unsupervised CLIP fine-tuning with universal entropy optimization. In *Proceedings of the 41st International Conference on Machine Learning*, 29667–29681.

Liu, B.; Xu, Y.; Xu, C.; Xu, X.; and He, S. 2025. Open-set mixed domain adaptation via visual-linguistic focal evolving. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, H.; Cao, Z.; Long, M.; Wang, J.; and Yang, Q. 2019. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2927–2936.

Liu, X.; and Zhou, Y. 2024. COCA: Classifier-Oriented Calibration via Textual Prototype for Source-Free Universal Domain Adaptation. In *Proceedings of the Asian Conference on Computer Vision*, 1671–1687.

Luo, Y.; Wang, Z.; Chen, Z.; Huang, Z.; and Baktashmotlagh, M. 2023. Source-free progressive graph learning for open-set domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 11240–11255.

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

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35: 35087–35102.

Panareda Busto, P.; and Gall, J. 2017. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, 754–763.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

Qu, S.; Zou, T.; He, L.; Röhrbein, F.; Knoll, A.; Chen, G.; and Jiang, C. 2024. Lead: Learning decomposition for source-free universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23334–23343.

Qu, S.; Zou, T.; Röhrbein, F.; Lu, C.; Chen, G.; Tao, D.; and Jiang, C. 2023. Upcycling models under domain and category shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20019–20028.

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

Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114. PMLR.

Tang, S.; Su, W.; Ye, M.; and Zhu, X. 2024. Source-free domain adaptation with frozen multimodal foundation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23711–23720.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5018–5027.

Wan, F.; Zhao, H.; Yang, X.; and Deng, C. 2024. Unveiling the unknown: Unleashing the power of unknown to known in open-set source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24015–24024.

Wang, H.; Li, Y.; Yao, H.; and Li, X. 2023. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1802–1812.

Wang, W.; Zhou, R.; Wang, J.; Zhou, Y.; Zhu, C.; Tang, R.; Han, B.; and Zhang, N. L. 2025. COSDA: Counterfactual-based Susceptibility Risk Framework for Open-Set Domain Adaptation. In *Forty-second International Conference on Machine Learning*.

Yang, Z.; Yue, J.; Ghamisi, P.; Zhang, S.; Ma, J.; and Fang, L. 2024. Open set recognition in real world. *International Journal of Computer Vision*, 132(8): 3208–3231.

Yu, Q.; Irie, G.; and Aizawa, K. 2025. Open-set domain adaptation with visual-language foundation models. *Computer Vision and Image Understanding*, 250: 104230.

Yu, Z.; Liao, Z.; Li, J.; Chen, Z.; and Zhu, L. 2025. Dynamic Target Distribution Estimation for Source-Free Open-Set Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22254–22262.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644.

Zhang, W.; Shen, L.; and Foo, C.-S. 2025. Source-free domain adaptation guided by vision and vision-language pre-training. *International Journal of Computer Vision*, 133(2): 844–866.