Text-Phase Synergy Network with Dual Priors for Unsupervised Cross-Domain Image Retrieval

Jing Yang[1,2] Hui Xue[1,2*] Shipeng Zhu[1,2] Pengfei Fang[1,2*]
[1]School of Computer Science and Engineering, Southeast University, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and
Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{yangjing.seu, hxue, shipengzhu, fangpengfei}@seu.edu.cn

Abstract

This paper studies unsupervised cross-domain image retrieval (UCDIR), which aims to retrieve images of the same category across different domains without relying on labeled data. Existing methods typically utilize pseudo-labels, derived from clustering algorithms, as supervisory signals for intra-domain representation learning and cross-domain feature alignment. However, these discrete pseudo-labels often fail to provide accurate and comprehensive semantic guidance. Moreover, the alignment process frequently overlooks the entanglement between domain-specific and semantic information, leading to semantic degradation in the learned representations and ultimately impairing retrieval performance. This paper addresses the limitations by proposing a Text-Phase Synergy Network with Dual Priors (TPSNet). Specifically, we first employ CLIP to generate a set of class-specific prompts per domain, termed as domain prompt, serving as a text prior that offers more precise semantic supervision. In parallel, we further introduce a phase prior, represented by domain-invariant phase features, which is integrated into the original image representations to bridge the domain distribution gaps while preserving semantic integrity. Leveraging the synergy of these dual priors, TPSNet significantly outperforms state-of-the-art methods on UCDIR benchmarks.

1. Introduction

Cross-domain image retrieval aims to bridge the distribution gaps between heterogeneous image domains (e.g., real images and sketches) and enable accurate matching by extracting domain-invariant semantic features. It has broad applicability across various fields, including intelligent security [18], medical image analysis [12], mobile product image search [33], to name but a few. Most existing methods address this task by training a retrieval model using extensive annotated data [8, 15, 43]. However, in real-world scenarios, acquiring large-scale, high-quality labeled data is often prohibitively expensive. To mitigate the cost of manual annotation, an emerging solution is to learn directly from unlabeled data, a task known as unsupervised cross-domain image retrieval (UCDIR).

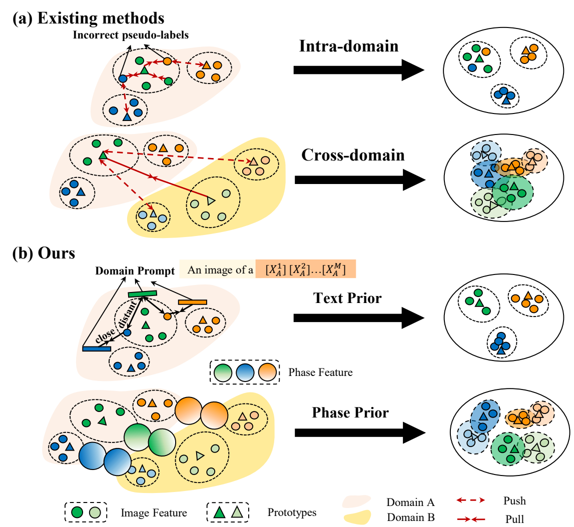

Figure 1. Comparison between (a) existing methods and (b) our proposed TPSNet. Existing methods rely on inaccurate pseudo-labels for intra-domain and cross-domain learning, often causing semantic loss. In contrast, TPSNet leverages text and phase dual priors to extract domain-invariant semantic features.

In UCDIR, the absence of label information poses a significant challenge to learn domain-invariant semantic features for effective cross-domain retrieval. To address this issue, recent approaches commonly adopt a two-stage learning paradigm: intra-domain representation learning followed by cross-domain feature alignment. This is typically achieved by first applying a clustering algorithm to gen-

[*]Co-corresponding authors

erate pseudo-labels, and then constructing class prototypes for each cluster to facilitate the learning process. However, it introduces two key issues, as illustrated in Figure 1. In the intra-domain representation learning stage, contrastive learning is reliant on pseudo-labels to pull together images from the same category and push apart those from different categories, thereby capturing semantic structure in each domain of data. Nevertheless, the resulting discrete pseudo-labels often fail to provide accurate and comprehensive semantic supervision. Such weak supervision leads to improper pairwise constraints and less discriminative class prototypes, which in turn hinder prototype-based alignment and degrade the quality of learned feature representations. In the cross-domain feature alignment stage, techniques such as adversarial domain adaptation [39], statistical distribution alignment [13], and cross-domain contrastive learning [20] are commonly used to reduce domain discrepancies from the perspective of the spatial domain. However, these alignment strategies may degrade semantic information due to the entangled nature of domain-specific and semantic features [10], especially in cases where pseudo-labels and prototypes are inaccurate or unreliable.

The recent success of vision-language models (VLMs), such as CLIP [30], has demonstrated promising capabilities in learning cross-modal semantic representations through large-scale image-text alignment [17, 19]. These textual representations provide richer semantic priors than discrete pseudo-labels, and can guide visual feature learning via multi-modal interaction [2, 28, 31]. Meanwhile, another line of research seeks to mitigate domain discrepancies by manipulating frequency components in the frequency domain, rather than relying on explicit alignment in the spatial domain. For example, studies such as FDA [44] and FUDA [40] obtain domain-invariant representations by replacing the low-frequency components of source images with those from target images, reducing inter-domain discrepancies while largely preserving semantic content. However, most existing approaches focus on manipulating high-frequency and low-frequency components. In contrast, the spectral decomposition perspective that separates amplitude and phase spectra remains underexplored, even though the phase spectrum is known to encode structural and semantic cues that are more robust to domain shifts [27].

Inspired by recent studies, we propose the *Text-Phase Synergy Network with Dual Priors* (TPSNet) for UCDIR, with its key components illustrated in Figure 1. We first introduce domain prompts, initialized as learnable embeddings for each cluster based on pseudo-labels generated via K-means clustering. The prompts are further tuned through an improved unsupervised CLIP-based contrastive learning scheme. It incorporates the semantic information encoded in the prompts to attain refined pseudo-labels. The resulting domain prompts then serve as a *text prior*, providing stronger semantic supervision for subsequent representation learning. Recognizing that the phase spectrum in the frequency domain conveys semantic signals invariant across data domains [14, 27], we additionally incorporate phase features as a *phase prior* to bridge domain discrepancies while preserving essential semantic content, often degraded by traditional alignment strategies. By jointly leveraging these *text-phase dual priors*, TPSNet effectively captures domain-invariant semantic features, thereby significantly enhancing performance on the UCDIR task. The contributions of this work can be summarized as follows:

- We introduce learnable domain prompts as a text prior to mitigate pseudo-label noise in UCDIR, enhancing semantic feature extraction via cross-modal alignment.
- We further incorporate a phase prior to alleviate semantic degradation in cross-domain alignment, enabling robust and domain-invariant representations through dual priors.
- Extensive experiments on the UCDIR datasets demonstrate the effectiveness of our method, which performs significantly better than the current state-of-the-arts and provides a new solution for the field of UCDIR.

2. Related Work

Cross-Domain Image Retrieval. Cross-domain image retrieval involves retrieving images where the query and gallery sets originate from different domains. Existing methods [15, 16, 24] predominantly rely on supervised learning with labeled data, which limits scalability in real-world applications. To address this limitation, recent research has focused on the setting of unsupervised cross-domain image retrieval (UCDIR), which seeks to learn domain-invariant semantic representations from unlabeled data. Representative approaches such as DD [13] and CODA [42] employ K-means clustering to generate pseudo-labels. DD further utilizes these pseudo-labels for intra-domain contrastive learning and minimizes the Distance-of-Distance loss to align distributions, while CODA leverages pseudo-labels to initialize both intra-domain and cross-domain classifiers, enabling the separate learning of intra-domain representations and cross-domain feature alignment. However, these methods decompose the UCDIR task into two independent stages, intra-domain representation learning and cross-domain alignment, overlooking the intrinsic correlation between them. To address this, ProtoOT [20] and ShieldIR [34] unify the two stages within an optimal transport framework, combining prototype alignment with contrastive learning. However, semantic degradation persists due to the entanglement of domain-specific and semantic features. SA-MoE [38] addresses this issue by emphasizing semantic modeling through semantically attentive fusion and contextual relevance, thereby preserving semantic integrity and underscoring the importance of semantics in UCDIR.

Unsupervised Domain Alignment Methods. Unsupervised domain alignment aims to bridge the distribution gaps between domains in the absence of labeled data. Within the UCDIR framework, effective domain alignment is critical for robust retrieval performance under domain discrepancies. To tackle this challenge, various alignment strategies have been proposed. Statistical alignment methods, such as MMD [22] and MDD [47], explicitly reduce domain discrepancies by aligning statistical properties (e.g., means and covariances) across domains. In contrast, adversarial learning approaches, including DANN [9] and CDAN [26], utilize domain discriminators to encourage feature extractors to learn domain-invariant representations through adversarial training. More recently, Optimal Transport (OT) has been applied to align domain distributions by computing an optimal mapping that minimizes transport cost [1, 4], providing a foundation for subsequent cross-domain contrastive learning that further mitigates domain discrepancies. Diverging from conventional alignment techniques, another line of work [40, 44] uses frequency domain information to reduce distribution discrepancies while preserving essential semantic content, offering a novel perspective for cross-domain alignment.

3. Method

3.1. Overview

We first formulate the problem of the UCDIR. Given two sets of unlabeled training images from different domains, denoted as $\mathcal{D}_A = \{x_i^A\}_{i=1}^{N_A}$ and $\mathcal{D}_B = \{x_j^B\}_{j=1}^{N_B}$, the goal of UCDIR is to learn a shared embedding space where semantically similar images from both domains are projected onto nearby representations. During testing, given a query image $x_i^A \in \mathcal{D}_A$ with class label y_i, the model is tasked with retrieving all semantically relevant images from $\mathcal{D}_B$ that share the same label y_i.

UCDIR is a challenging problem due to the stringent requirement of training a retrieval model in different data domains without relying on any annotations. To address this, we propose a novel approach named TPSNet, as illustrated in Figure 2. TPSNet comprises two key components: one Domain Prompt Generation Module (DPG) and one Text-Phase Dual Priors Network (TPDP). In DPG, class-specific prompts are optimized per domain using CLIP-based contrastive learning. Having the learned domain prompts at hand, TPDP adopts a dual-path architecture: the text branch leverages the learned domain prompts to provide enriched semantic supervision, while the image branch integrates phase features to support cross-domain alignment and preserve semantic content. By leveraging the synergy of text-phase dual priors, TPSNet effectively facilitates the extraction of domain-invariant semantic visual representations.

3.2. Domain Prompt Generation Module

The overall diagram of the proposed DPG is presented on the left of Figure 2. Following previous UCDIR methods, we first employ K-means clustering to generate discrete pseudo-labels for two domain of images. For each domain, we initialize C learnable class-specific prompts corresponding to the clusters identified by pseudo-labels. Each prompt is formulated as a textual template, e.g., "An image of a $[X]_A^1[X]_A^2...[X]_A^M$." for $\mathcal{D}_A$ and "An image of a $[X]_B^1[X]_B^2...[X]_B^M$." for $\mathcal{D}_B$. Here, $[X]^i$ is a learnable token that is randomly initialized, and M denotes the number of learnable tokens.

The initialized domain prompts and images are then fed into the pre-trained CLIP model as inputs, where a cross-modal contrastive learning strategy is employed to effectively update the domain prompts. Note that the parameters of both the image and text encoders are frozen, with only the token $[X]^i$ within the domain prompts is optimized. During this stage, the contrastive losses $\mathcal{L}_{i2t}^i$ and $\mathcal{L}_{t2i}^i$ are utilized, defined as follows:

$$\mathcal{L}_{i2t}^i = -\log \frac{\exp\left(s\left(I_i, T_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(s\left(I_i, T_j\right)/\tau\right)}, \tag{1}$$

and

$$\mathcal{L}_{t2i}^i = -\frac{1}{\left|P\left(y_i\right)\right|} \sum_{p \in P(y_i)} \log \frac{\exp\left(s\left(I_p, T_{y_i}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(s\left(I_j, T_{y_i}\right)/\tau\right)}, \tag{2}$$

where I_i and T_i denote the re-paired image and text embeddings based on cosine similarity scores, which allows the model to partially correct inaccurate pseudo-labels through contrastive alignment. $s(I_i, T_i) = \frac{I_i^\top T_i}{\|I_i\|\|T_i\|}$ and τ is the temperature. N is the batch size, $P(y_i)$ is the set of positive samples with respect to T_{y_i}, i.e., $P(y_i) = p \in \{1...N|y_p = y_i\}$. The symbol $|\cdot|$ denotes the Cardinality of this set. Then the proposed DPG is optimized via the following loss: $\mathcal{L}_{prompt} = \mathcal{L}_{i2t}^i + \mathcal{L}_{t2i}^i$.

By minimizing $\mathcal{L}_{prompt}$, the incrementally optimized domain prompts are leveraged as supervisory signals to refine the potentially inaccurate discrete pseudo-labels generated by the clustering process in DPG. This optimization yields domain prompts that encapsulate semantic information from the text representations, which are subsequently incorporated into the TPDP module to guide the extraction of semantic representations.

3.3. Text-Phase Dual Priors Network

The overall architecture of TPDP is illustrated in the top-right part of Figure 2. The proposed method adopts a dual-path design, consisting of a Text-Prior Semantic Feature Extraction (TPFE) module for the text path and a Phase-Prior Domain-Invariant Feature Extraction (PPFE) module

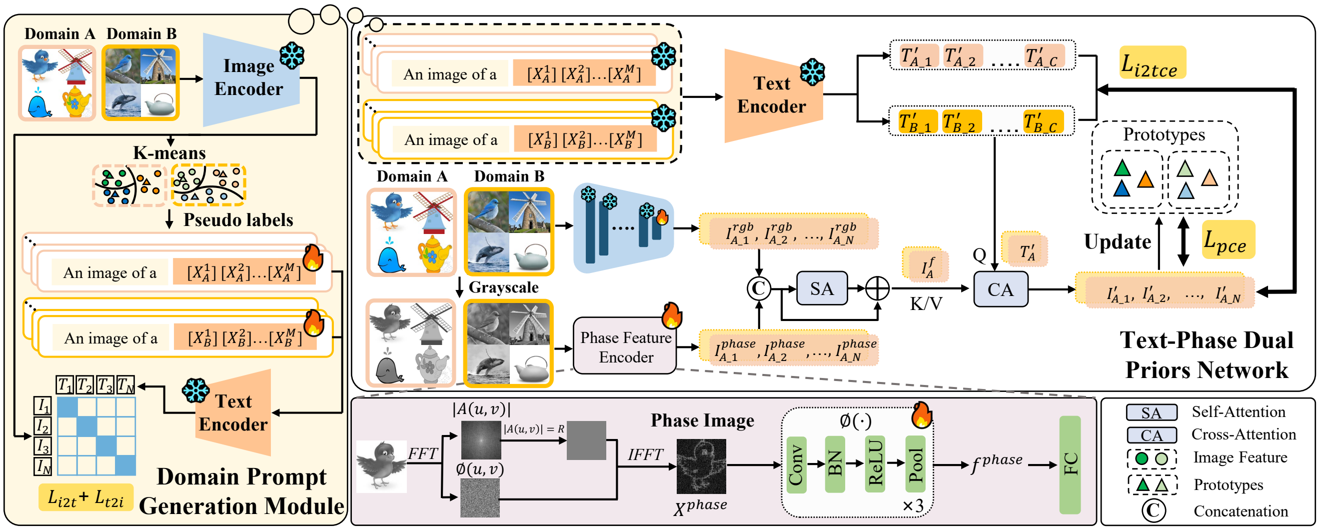

Figure 2. The pipeline of TPSNet. Left: domain prompt generation via the prompt learning paradigm. Top-right: text-phase dual prior construction with contrastive learning for unsupervised cross-domain image retrieval. Bottom: the detailed architecture of the proposed phase feature encoder.

for the image path. The following sections provide detailed descriptions of TPFE and PPFE, respectively.

Text-Prior Semantic Feature Extraction. In order to extract domain-invariant semantic features, we design the TPFE module utilizing the domain prompts learned from DPG. These domain prompts are passed through a shared text encoder to generate text features. For $\mathcal{D}_A$, the resulting text features are denoted as $T'_A = \{T'_{A\_1}, T'_{A\_2}, \ldots, T'_{A\_C}\}$, and for $\mathcal{D}_B$, as $T'_B = \{T'_{B\_1}, T'_{B\_2}, \ldots, T'_{B\_C}\}$, where C is the number of semantic categories. These text features, T'_A and T'_B, will serve as text prior to guide the semantic representation learning in subsequent process.

Phase-Prior Domain-Invariant Feature Extraction. In the PPFE module, images from two domains $\mathcal{D}_A$ and $\mathcal{D}_B$ are first processed by an image encoder to obtain RGB features I^{rgb}. Although I^{rgb} provides high-level semantics, it is inevitably influenced by domain-specific factors such as background and style, which limits its effectiveness for cross-domain retrieval. To further mitigate this issue, we further extract the corresponding phase features I^{phase} from grayscale images using a dedicated Phase Feature Encoder, as described in the following paragraph, and fuse them with I^{rgb}:

$$I^f = \text{Mean}\Big(\text{LayerNorm}\big(\text{Concat}(I^{rgb}, I^{phase}) + \text{SelfAttention}\big(\text{Concat}(I^{rgb}, I^{phase})\big)\big)\Big), \tag{3}$$

where Concat$(\cdot)$ and Mean$(\cdot)$ denote the concatenation and mean pooling, respectively. SelfAttention$(\cdot)$ and LayerNorm$(\cdot)$ refer to the self-attention and layer normalization. This operation yields the fused features I^f for two domains, effectively integrating semantic content from RGB with enhanced structural consistency from phase to fuse domain-invariant information while retaining rich semantics for robust representations.

Phase Feature Encoder. The architecture of the Phase Feature Encoder is illustrated at the bottom of Figure 2. In this process, we apply the Fast Fourier Transform (FFT) to the grayscale images to obtain the frequency representation $F(u,v) = |A(u,v)|e^{j\phi(u,v)}$, where $|A(u,v)|$ and $\phi(u,v)$ denote the amplitude spectrum and phase spectrum, respectively. To emphasize domain-invariant information, we retain the phase spectrum while replacing the amplitude spectrum with a constant value R, yielding the modified frequency signal $F'(u,v) = Re^{j\phi(u,v)}$. The final phase image X^{phase} is reconstructed by applying the Inverse Fast Fourier Transform (IFFT):

$$X^{phase} = \mathcal{F}^{-1}\{F'(u,v)\}, \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the IFFT operation. Then, we design a lightweight convolutional subnetwork to extract phase information f^{phase} from the phase image X^{phase}:

$$\phi(\cdot) = \text{AvgPool}_{2\times2}(\text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(\cdot)))), \tag{5}$$

$$f^{phase} = \phi\left(X^{phase}\right). \tag{6}$$

The f^{phase} is then passed through a fully connected layer, resulting the final phase features as:

$$I^{phase} = \text{FC}\left(\text{Flatten}(f^{phase})\right). \tag{7}$$

Synergy of Text-Phase Dual Priors. To effectively incorporate dual priors from text and phase information into cross-domain retrieval, we adopt a cross-attention mechanism to enable their synergistic integration. For each domain, the resulting text features T' serve as query vectors and the fused visual features I^f are used as key and value vectors. The cross-attention mechanism is adopted to fuse the two priors, as:

$$I' = \text{CrossAttention}(T'; I^f). \tag{8}$$

This yields the final visual features I'_A and I'_B for the two domains, which are enriched with the synergy of text-phase dual priors and retain domain-invariant semantic information. After obtaining the final visual features for two domains, we perform prototype updates at each training epoch for each domain individually:

$$\mathcal{P}_{y_i} \leftarrow m\mathcal{P}_{y_i} + (1-m)I'_i, \tag{9}$$

where $\mathcal{P}$ denotes the class prototypes initialized via K-means clustering, and α is a momentum-based update factor. As shown in Figure 2, prototype cross-entropy loss $\mathcal{L}_{pce}$ and image-to-text contrastive loss $\mathcal{L}_{i2tce}$ are employed in two domains to enhance representation learning:

$$\mathcal{L}^i_{pce} = -\log \frac{\exp\left(s\left(I'_i, \mathcal{P}_{y_i}\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(s\left(I'_i, \mathcal{P}_c\right)/\tau\right)}, \tag{10}$$

and

$$\mathcal{L}^i_{i2tce} = \sum_{j=1}^{C} -\sigma_j \log \frac{\exp\left(s\left(I'_i, T'_j\right)/\tau\right)}{\sum_{c=1}^{C} \exp\left(s\left(I'_i, T'_c\right)/\tau\right)}, \tag{11}$$

where C is the number of the categories. To further mitigate the impact of errors in pseudo-labels, we introduce a smoothing label $\sigma_j = (1-\epsilon)\cdot y_i + \epsilon/C$, where y_i is the one-hot encoding of the pseudo-label and ϵ is a constant. The overall loss function for TPDP is then computed by summing the individual losses:

$$\mathcal{L} = \alpha\mathcal{L}_{pce} + \beta\mathcal{L}_{i2tce}, \tag{12}$$

where α and β are hyperparameters balancing the importance of $\mathcal{L}_{pce}$ and $\mathcal{L}_{i2tce}$, respectively.

4. Experiments

4.1. Datasets and Setting

Datasets. We evaluate the proposed TPSNet on two widely used datasets for cross-domain retrieval: Office-Home and DomainNet. Office-Home dataset [37] comprises 65 classes and 4 domains (Art, Clipart, Product, Real). DomainNet dataset [29] consists of 6 domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. Following the criteria established by DD [13], we select 7 classes from the DomainNet dataset with over 200 images in each domain for training and testing.

Evaluation Metrics. Following DD [13], we assess retrieval performance on the Office-Home dataset using precision at top-k retrieved images, specifically at k = 1, 5, and 15 (i.e., P@1, P@5, and P@15). For the DomainNet dataset, we report precision at k = 50, 100, and 200 (i.e., P@50, P@100, and P@200).

Implementation Details. We use the pre-trained CLIP [30] architecture as the backbone framework. Following the consensus in this community, we adopt ResNet-50 [11] and ViT-B [7] as image encoder to investigate the effectiveness of TPSNet. During the DPG stage, the CLIP backbone is frozen, and only the learnable text tokens $[X]^1\,[X]^2 \ldots [X]^M$ are updated. In the TPDP stage, we unfreeze the last block of the image encoder, allowing fine-tuning. Each input image is augmented using random horizontal flipping, padding, cropping, and random erasing. The model is optimized using the Adam optimizer with a learning rate of 0.0001, following a cosine decay schedule. All the experiments are conducted on a single NVIDIA GeForce RTX 3090.

4.2. Comparison with State-of-the-Art Methods

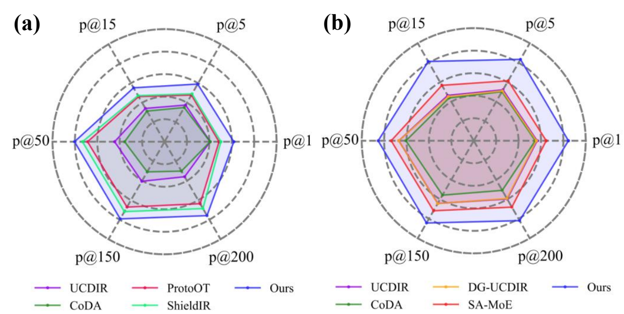

Figure 3. Average Accuracy (%) of UCDIR Methods using (a) ResNet-50 and (b) ViT-B as image encoders.

In our experimental validation, we evaluate the effectiveness of TPSNet by comparing it with the following methods: DD [13], CoDA [42], DG-UCDIR [14], ProtoOT [20], ShieldIR [34] and SA-MoE [38]. Table 1 and Table 2 show the detailed retrieval performance for Office-Home and DomainNet, respectively. TPSNet's impressive performance is evident from the summarized results in Figure 3, with detailed average accuracies provided in Table 3.

Through comparison, we observe the following: 1) TPSNet outperforms other methods on nearly all metrics across the two datasets, highlighting its effectiveness. 2) With either ResNet-50 or ViT-B as the image encoder, TPSNet outperforms other compared methods on nearly all metrics, demonstrating its robustness. 3) TPSNet exhibits more substantial performance gains on the Office-Home

Table 1. Detailed comparison with UCDIR methods on the Office-Home dataset across 12 individual cross-domain retrieval scenarios.

Method		$Art \rightarrow Real$			$Real \rightarrow Art$			$Art \rightarrow Product$			$Product \rightarrow Art$		
		$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$
ResNet-50	DD [13]	45.12	42.33	40.06	47.95	43.68	38.38	35.39	34.67	32.61	42.51	37.94	31.41
	CoDA [42]	44.77	40.99	36.64	44.88	37.54	38.12	34.52	33.96	31.06	40.98	32.24	30.54
	ProtoOT [20]	49.94	49.24	49.55	54.90	51.81	48.17	44.25	45.95	45.71	53.43	52.07	46.89
	ShieldIR [34]	51.46	50.20	49.27	57.33	53.50	49.26	44.75	46.16	45.33	56.31	52.71	47.39
	TPSNet	**69.06**	**67.22**	**65.02**	**74.00**	**69.72**	**62.73**	**59.41**	**57.21**	**53.50**	**63.82**	**59.98**	**52.31**
ViT-B	DD [13]	60.94	56.44	51.75	61.92	54.14	44.97	58.76	56.55	52.65	60.80	53.89	44.88
	CoDA [42]	59.65	53.68	46.84	58.34	52.82	40.84	57.32	53.86	50.80	57.53	51.54	40.11
	DG-UCDIR [14]	59.29	52.76	46.59	58.83	51.54	39.48	56.51	53.01	52.65	61.46	55.89	47.42
	SA-MoE [38]	71.12	68.93	66.10	73.86	68.85	60.37	64.69	62.39	58.63	66.57	62.82	55.12
	TPSNet	**89.53**	**88.61**	**87.77**	**91.74**	**90.71**	**86.81**	**81.17**	**80.60**	**79.43**	**88.80**	**86.14**	**81.01**

Method		$Clipart \rightarrow Real$			$Real \rightarrow Clipart$			$Product \rightarrow Real$			$Real \rightarrow Product$		
		$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$
ResNet-50	DD [13]	33.31	30.57	28.14	44.66	41.47	37.41	57.42	52.69	47.90	51.71	48.48	44.95
	CoDA [42]	30.12	27.10	24.02	43.65	35.21	29.06	57.37	50.98	42.82	55.23	49.18	44.36
	ProtoOT [20]	40.92	40.68	39.96	51.11	52.12	50.44	69.34	67.27	64.74	57.72	59.32	59.72
	ShieldIR [34]	43.36	42.58	41.55	52.05	52.38	50.81	70.37	68.33	66.08	61.46	62.08	61.85
	TPSNet	**51.68**	**49.91**	**47.56**	**63.53**	**61.36**	**57.40**	**78.85**	**75.87**	**72.48**	**74.39**	**73.00**	**70.94**
ViT-B	DD [13]	42.66	39.51	36.19	59.31	54.58	48.18	70.56	65.19	58.58	65.87	62.57	57.58
	CoDA [42]	41.06	38.46	34.83	57.42	51.47	45.92	70.43	63.62	56.84	63.94	60.13	55.16
	DG-UCDIR [14]	42.87	39.31	35.94	60.86	55.35	48.90	72.12	66.87	60.41	66.25	62.97	58.09
	SA-MoE [38]	52.99	50.00	47.33	70.41	65.78	60.06	78.91	74.96	70.40	76.52	74.54	71.18
	TPSNet	**73.36**	**72.73**	**71.85**	**89.03**	**88.01**	**84.86**	**93.26**	**92.82**	**91.78**	**92.10**	**92.05**	**91.78**

Method		$Product \rightarrow Clipart$			$Clipart \rightarrow Product$			$Art \rightarrow Clipart$			$Clipart \rightarrow Art$		
		$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$	$P@1$	$P@5$	$P@15$
ResNet-50	DD [13]	42.26	37.42	33.74	27.79	27.26	25.97	32.67	30.79	28.70	27.26	23.94	20.53
	CoDA [42]	47.21	35.43	28.33	27.10	24.77	24.00	36.69	32.19	26.37	25.64	21.17	21.37
	ProtoOT [20]	51.92	50.66	49.58	39.01	37.95	37.16	40.54	38.27	36.01	30.24	29.31	27.34
	ShieldIR [34]	52.10	52.43	50.23	40.84	39.87	39.45	41.24	39.05	37.44	32.21	30.13	28.39
	TPSNet	**60.71**	**57.49**	**53.26**	**46.05**	**43.77**	**42.22**	**56.98**	**54.07**	**50.29**	**43.14**	**39.66**	**35.58**
ViT-B	DD [13]	57.87	54.00	47.87	45.22	43.29	40.13	56.00	51.89	45.78	39.06	34.64	29.27
	CoDA [42]	54.44	51.97	44.68	44.84	42.62	38.91	55.76	50.41	44.63	38.42	33.02	27.41
	DG-UCDIR [14]	56.58	53.41	47.09	43.89	41.25	37.48	55.93	51.50	44.93	36.51	30.04	25.43
	SA-MoE [38]	65.28	61.09	56.01	47.47	45.81	43.36	61.80	57.82	53.23	46.90	44.39	39.97
	TPSNet	**88.31**	**86.74**	**84.49**	**72.67**	**72.30**	**71.79**	**84.55**	**81.94**	**79.36**	**69.87**	**68.67**	**64.55**

dataset, which contains a larger number of categories (65 in total). Specifically, when using ResNet-50 as the image encoder, TPSNet surpasses the SOTA by average margins of 11.51%, 9.99%, and 8.02% in $P@1$, $P@5$, and $P@15$, respectively. The improvements are even more pronounced with ViT-B as the image encoder, yielding gains of 19.82%, 22.00%, and 24.48% in the same metrics. These results highlight the effectiveness of TPSNet in handling more complex and challenging datasets. 4) We observe that TPSNet achieves substantial improvements on noisy domains with large visual variation. For instance, on the Infograph-real and Infograph-sketch scenarios of the DomainNet dataset, TPSNet outperforms SOTA by more than 20% in $P@50$ when ViT-B is used as the image encoder. These results demonstrate that TPSNet facilitates more effective semantic learning, thereby improving the model's robustness to the domain distribution gaps.

In addition, to comprehensively evaluate the effectiveness and robustness of TPSNet and ensure that the observed improvements are not solely attributed to CLIP, we replace the pre-trained CLIP image encoder with MoCov2 [6] and DINO [5], which are widely adopted in prior works [13, 20, 38]. All backbones are trained under identical settings for fair comparison. As shown in Table 3, TPSNet consistently outperforms state-of-the-art methods across both datasets, even when MoCov2 and DINO are used for initialization. These results strongly indicate that the performance gains are not merely dependent on CLIP,

Table 2. Detailed comparison with UCDIR methods on the DomainNet dataset across 12 individual cross-domain retrieval scenarios.

Method		*Clipart* → *Sketch*			*Sketch* → *Clipart*			*Infograph* → *Real*			*Real* → *Infograph*		
		$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$
ResNet-50	DD [13]	56.31	52.47	47.38	63.07	57.26	48.17	35.52	35.24	34.35	57.74	46.69	35.47
	CoDA [42]	44.56	35.86	35.14	49.00	38.61	38.49	27.12	27.21	26.43	36.98	30.44	30.02
	ProtoOT [20]	82.70	82.26	80.32	86.69	85.57	81.59	48.74	49.28	49.60	75.36	68.49	56.20
	ShieldIR [34]	87.16	87.12	86.74	89.95	89.54	88.56	54.41	54.93	55.34	82.26	75.35	64.15
	TPSNet	**95.43**	**95.31**	**94.63**	**95.50**	**95.31**	**94.26**	**70.04**	**69.55**	**68.74**	**86.07**	**81.91**	**71.99**
ViT-B	DD [13]	76.55	73.39	68.66	82.36	79.22	71.46	47.12	47.15	46.89	78.01	68.07	53.17
	CoDA [42]	73.62	70.22	65.51	80.21	76.00	67.68	46.11	45.98	45.13	73.69	61.67	45.62
	DG-UCDIR [14]	77.47	73.83	69.31	82.86	80.85	72.19	48.92	47.97	45.20	78.72	68.71	52.37
	SA-MoE [38]	83.97	82.08	78.40	88.11	86.07	81.21	57.29	57.55	57.67	87.37	80.08	66.24
	TPSNet	**97.92**	**97.84**	**97.73**	**98.27**	**98.20**	**98.06**	**78.72**	**78.55**	**78.31**	**93.01**	**93.04**	**85.71**

Method		*Infograph* → *Sketch*			*Sketch* → *Infograph*			*Painting* → *Clipart*			*Clipart* → *Painting*		
		$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$
ResNet-50	DD [13]	31.29	29.33	26.54	43.66	36.14	28.12	66.42	56.84	46.72	52.58	50.10	46.11
	CoDA [42]	24.94	22.48	22.42	27.65	23.85	23.48	57.30	44.25	44.15	46.51	42.38	41.22
	ProtoOT [20]	45.72	45.51	43.77	62.92	56.41	47.08	91.04	90.09	86.67	82.70	82.66	81.38
	ShieldIR [34]	47.39	47.10	46.34	72.29	67.68	55.00	93.96	**93.47**	**90.99**	85.32	85.34	84.30
	TPSNet	**67.59**	**67.00**	**65.96**	**90.20**	**86.27**	**74.90**	**94.07**	93.27	90.61	**94.56**	**93.99**	**91.08**
ViT-B	DD[13]	49.94	48.72	45.11	71.97	62.05	49.32	93.89	93.24	90.09	86.00	85.57	83.72
	CoDA [42]	45.85	44.52	40.35	67.51	55.69	43.87	85.93	82.34	80.29	82.56	80.18	79.65
	DG-UCDIR [14]	50.38	49.24	45.62	72.03	63.15	51.99	92.40	93.17	89.43	86.17	85.83	82.86
	SA-MoE [38]	53.76	53.07	50.37	79.56	72.29	58.62	95.00	94.59	93.23	90.60	90.42	89.25
	TPSNet	**75.87**	**75.61**	**75.03**	**95.21**	**92.04**	**84.78**	**98.49**	**98.45**	**98.26**	**97.84**	**97.73**	**96.64**

Method		*Painting* → *Quickdraw*			*Quickdraw* → *Painting*			*Quickdraw* → *Real*			*Real* → *Quickdraw*		
		$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$	$P@50$	$P@100$	$P@200$
ResNet-50	DD [13]	39.72	38.59	37.63	33.45	33.81	34.29	30.24	28.40	26.14	25.00	22.26	19.73
	CoDA [42]	28.73	26.69	24.95	21.05	20.52	20.81	23.98	22.59	21.50	39.87	36.28	32.91
	ProtoOT [20]	63.06	59.92	56.73	56.51	55.66	52.04	61.55	61.60	61.63	67.24	67.02	65.41
	ShieldIR [34]	65.33	61.25	57.98	58.07	57.42	**55.11**	67.19	67.36	67.53	**75.70**	73.60	70.63
	TPSNet	**67.21**	**65.38**	**61.88**	**59.27**	**57.93**	54.30	**69.34**	**68.88**	**68.17**	75.06	**74.70**	**72.96**
ViT-B	DD [13]	63.19	61.19	58.09	53.55	53.46	53.09	48.71	48.74	48.68	51.65	50.48	48.97
	CoDA [42]	50.24	48.58	44.68	43.49	42.65	41.94	37.67	29.63	26.21	39.85	31.84	29.49
	DG-UCDIR [14]	65.06	62.90	59.89	54.18	53.95	53.21	45.20	44.30	42.63	50.62	49.24	46.80
	SA-MoE [38]	72.54	71.16	69.22	61.32	61.21	61.03	55.78	55.59	55.49	59.02	57.55	55.81
	TPSNet	**81.67**	**77.04**	**71.34**	**65.12**	**64.51**	**63.31**	**62.50**	**62.21**	**61.56**	**79.34**	**76.89**	**72.14**

Table 3. Average accuracy (%) for evaluating the impact of image encoder initialization on the Office-Home and DomainNet datasets using ResNet-50 and ViT-B backbones.

Method	ResNet-50					
	Office-Home			Domainnet		
	$P@1$	$P@5$	$P@15$	$P@50$	$P@100$	$P@200$
SOTA [34]	50.29	49.12	47.25	73.25	71.68	68.56
TPSNet (MoCov2)	56.71	54.64	52.14	74.04	72.80	70.00
TPSNet	**61.80**	**59.11**	**55.27**	**80.36**	**79.13**	**75.79**
Method	ViT-B					
	Office-Home			Domainnet		
	$P@1$	$P@5$	$P@15$	$P@50$	$P@100$	$P@200$
SOTA [38]	64.71	61.45	56.81	73.69	71.81	68.05
TPSNet (DINO)	67.62	65.31	62.29	80.97	79.30	75.48
TPSNet	**84.53**	**83.44**	**81.29**	**85.33**	**84.34**	**81.91**

but instead reflect the intrinsic effectiveness and generalizability of TPSNet. Additional detailed comparisons are provided in Supplementary Material B.4.

4.3. Ablation Study

Quantitative Evaluation. To evaluate the contribution of each component in TPSNet, we perform an ablation study on two datasets using ViT-B as the image encoder, with results summarized in Table 4. The results indicate that the integration of domain prompts provides substantial performance gains by introducing additional semantic supervision, yielding an improvement of 11.91% in $P@1$ on the Office-Home dataset and 6.69% in $P@50$ on DomainNet. Furthermore, the inclusion of TPFE and PPFE modules leads to additional performance enhancements by strengthening semantic information while reducing domain discrepancies. When all components are combined, TPSNet achieves further performance gains, proving the efficacy of our method.

Table 4. Ablation study of TPSNet components with ViT-B as the image encoder. Results averaged over 12 scenarios.

DPG	TPFE	PPFE	Office-Home			DomainNet		
			$P@1$	$P@5$	$P@15$	$P@50$	$P@100$	$P@200$
✗	✗	✗	65.18	60.77	54.87	70.94	68.09	62.58
✓	✗	✗	77.09	75.10	71.43	77.63	74.72	69.18
✓	✗	✓	81.34	80.22	78.07	81.88	79.73	75.75
✓	✓	✗	81.63	80.13	77.72	82.05	79.94	75.97
✓	✓	✓	**84.53**	**83.44**	**81.29**	**85.33**	**84.34**	**81.91**

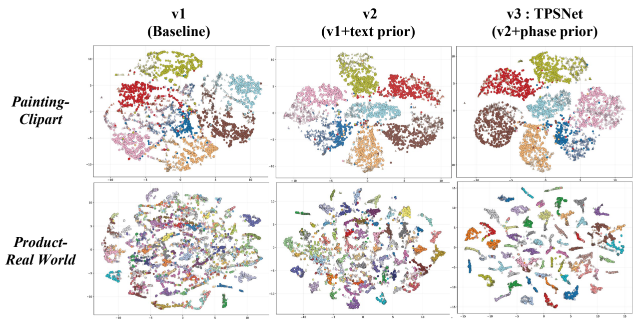

Figure 4. t-SNE visualizations of last-layer features for the baseline model (v1), baseline with text prior (v2), and TPSNet (v3) across two scenarios from two datasets.

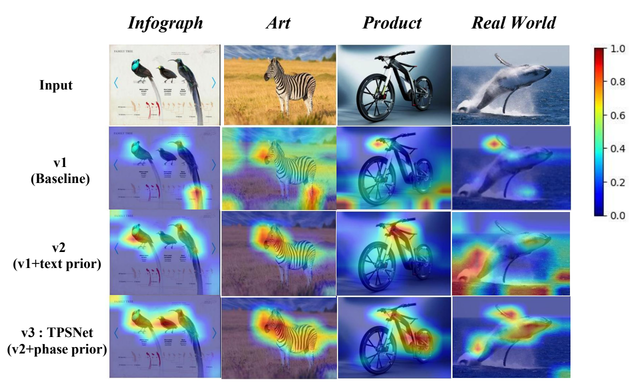

Figure 5. Grad-CAM visualizations of last-layer features for the baseline model (v1), baseline with text prior (v2), and TPSNet (v3) on randomly selected samples.

Qualitative Evaluation. To evaluate the explainability of the proposed text-phase dual priors in semantic feature extraction and domain distribution alignment, we perform two sets of qualitative assessments. Specifically, we compare three models: the baseline model (v1), the baseline model enhanced with the text prior (v2), and TPSNet (v3), which integrates all components.

First, we visualize the t-SNE [36] representations of features extracted from the final layer of the image encoder, as shown in Figure 4. The results show that incorporating the text prior facilitates more compact clustering of semantically similar samples, while the phase prior further aligns similar samples across domains into a shared feature space. This highlights the effectiveness of TPSNet in learning domain-invariant semantic representations. Subsequently, we present Grad-CAM [32] visualizations of the last-layer features from the backbone for models v1, v2, and v3 in Figure 5. The results indicate that the synergy of the text-phase dual priors enables TPSNet to more accurately focus on target objects, such as the bird in a complex background and the whale leaping across the sea surface. More visualizations can be found in Supplementary Material B.9.

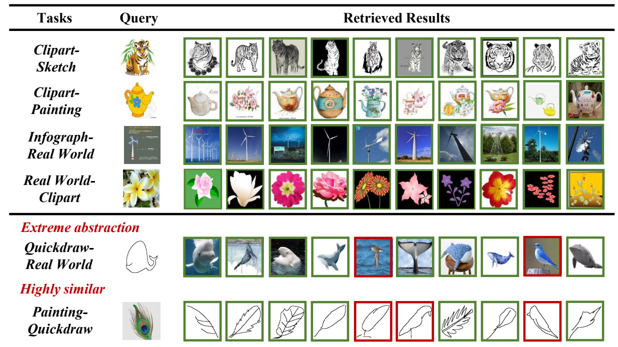

Figure 6. Top-10 retrieval results in various cross-domain retrieval scenarios from two datasets. The green and red boxes denote correct and incorrect retrievals, respectively.

Visualization of the Retrieval Results. We randomly select various cross-domain retrieval scenarios from two datasets for evaluation. As shown in Figure 6, TPSNet achieves 100% top-10 retrieval accuracy in most scenarios. We further analyze the failure cases and find that retrieval errors mainly arise under two conditions: 1) queries come from highly abstract domains such as Quickdraw, where fine-grained semantic information is severely limited, and 2) query and retrieved samples exhibit strong visual similarity, for example when bird and feather instances become nearly indistinguishable in the Quickdraw domain.

5. Conclusion

In this paper, we address the challenge of unsupervised cross-domain image retrieval (UCDIR) by proposing a novel method called TPSNet. TPSNet leverages domain prompts derived from image-text multi-modal alignment as a text prior to refine inaccurate pseudo-labels and enhance semantic representations. In addition, it integrates phase features as a phase prior to reduce domain discrepancies while preserving semantic information. Through the synergy of text-phase dual priors, TPSNet effectively extracts domain-invariant semantic representations for UCDIR. Extensive experiments on two datasets demonstrate its significant superiority over state-of-the-art methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62476056, T24B2005, and 62306070) and the Southeast University StartUp Grant for New Faculty under Grant 4009002309, and the Fundamental Research Funds for the Central Universities (2242025K30024). Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

References

[1] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Proceedings of the International Conference on Learning Representations*, 2019. 3

[2] Shuanghao Bai, Min Zhang, Wanqi Zhou, Siteng Huang, Zhirong Luan, Donglin Wang, and Badong Chen. Prompt-based distribution alignment for unsupervised domain adaptation. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 729–737, 2024. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 7

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 9912–9924, 2020. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9650–9660, 2021. 6, 3

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2020. 5

[8] Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, and Heng Tao Shen. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17292–17301, 2024. 1

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 3

[10] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1160–1170, 2023. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[12] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 440–450, 2022. 1

[13] Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 529–544. Springer, 2022. 2, 5, 6, 7, 3

[14] Conghui Hu, Can Zhang, and Gim Hee Lee. Unsupervised feature representation learning for domain-generalized cross-domain image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11016–11025, 2023. 2, 5, 6, 7

[15] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1062–1070, 2015. 1, 2

[16] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1654–1662, 2017. 2

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916, 2021. 2

[18] Sri Karnila, Suhendro Irianto, and Rio Kurniawan. Face recognition using content based image retrieval for intelligent security. *International Journal of Advanced Engineering Research and Science*, 6(1):91–98, 2019. 1

[19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning*, pages 5583–5594, 2021. 2

[20] Bin Li, Ye Shi, Qian Yu, and Jingya Wang. Unsupervised cross-domain image retrieval via prototypical optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3009–3017, 2024. 2, 5, 6, 7, 3

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022. 2, 4, 5

[22] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning*, pages 6028–6039, 2020. 3

[23] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed:

Universal multimodal retrieval with multimodal llms. In *Proceedings of the International Conference on Learning Representations*, 2025. 5, 7

[24] Haoran Liu, Ying Ma, Ming Yan, Yingke Chen, Dezhong Peng, and Xu Wang. Dida: Disambiguated domain alignment for cross-domain retrieval with partial labels. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 3612–3620, 2024. 2

[25] Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 5, 7

[26] Mingsheng Long, ZHANGJIE CAO, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2018. 3

[27] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 2

[28] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[29] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 5

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5

[31] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2

[32] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[33] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Mobile product image search by automatic query object extraction. In *Proceedings of the European Conference on Computer Vision*, pages 114–127. Springer, 2012. 1

[34] Zixin Tang, Haihui Fan, Jinchao Zhang, Hui Ma, Xiaoyan Gu, Bo Li, and Weiping Wang. Shieldir: Privacy-preserving unsupervised cross-domain image retrieval via dual protection transformation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6383–6392, 2025. 2, 5, 6, 7

[35] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 3, 4, 5

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 8

[37] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 5

[38] Kai Wang, Jiayang Liu, Xing Xu, Jingkuan Song, Xin Liu, and Heng Tao Shen. Unsupervised cross-domain image retrieval with semantic-attended mixture-of-experts. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 197–207, 2024. 2, 5, 6, 7, 3

[39] Lixu Wang, Xinyu Du, and Qi Zhu. Semantic feature learning for universal unsupervised cross-domain retrieval. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024. 2

[40] Mengzhu Wang, Shanshan Wang, Ye Wang, Wei Wang, Tianyi Liang, Junyang Chen, and Zhigang Luo. Boosting unsupervised domain adaptation: A fourier approach. *Knowledge-Based Systems*, 264:110325, 2023. 2, 3

[41] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 5, 7

[42] Xu Wang, Dezhong Peng, Ming Yan, and Peng Hu. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10200–10208, 2023. 2, 5, 6, 7

[43] Ling Xiao and Toshihiko Yamasaki. Boosting fine-grained fashion retrieval with relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8229–8234, 2024. 1

[44] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 3

[45] Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025. 5, 7

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11975–11986, 2023. 2, 4, 5

[47] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7404–7413, 2019. 3