DUDE: DIFFUSION-BASED UNSUPERVISED CROSS-DOMAIN IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised cross-domain image retrieval (UCIR) aims to retrieve images of the same category across diverse domains without relying on annotations. Existing UCIR methods, which align cross-domain features for the entire image, often struggle with the domain gap, as the object features critical for retrieval are frequently entangled with domain-specific styles. To address this challenge, we propose DUDE, a novel UCIR method building upon feature disentanglement. In brief, DUDE leverages a text-to-image generative model to disentangle object features from domain-specific styles, thus facilitating semantical image retrieval. To further achieve reliable alignment of the disentangled object features, DUDE aligns mutual neighbors from within domains to across domains in a progressive manner. Extensive experiments demonstrate that DUDE achieves superior performance across three benchmark datasets over 13 domains. The code will be released.

1 INTRODUCTION

Cross-domain image retrieval (CIR) aims to retrieve relevant images from one domain given a query from another domain Zhou et al. (2022). CIR has widespread applications, such as matching user-uploaded photos with product listings on shopping platforms Liu et al. (2016); Huang et al. (2015b) and matching forensic sketches with suspect databases in law enforcement Koley et al. (2023); Ouyang et al. (2016). To effectively address the CIR problem, existing works Wang et al. (2023a); Ji et al. (2017); Zhou et al. (2022); Huang et al. (2015a) typically rely on annotated labels or paired data as supervision signals, which are costly and labor-intensive to obtain in practice. To reduce the labeling burden, recent efforts have pivoted toward unsupervised cross-domain image retrieval (UCIR) Wang et al. (2023c); Kim et al. (2021); Yue et al. (2021); Wu et al. (2018), liberated from manual annotations. To achieve UCIR, most existing methods resort to instance- and cluster-wise contrastive learning for cross-domain alignment Hu & Lee (2022); Li et al. (2024a); Wang et al. (2023b). A recent study proposes phase-enhanced contrastive learning Hu et al. (2023), which narrows the domain gap by converting RGB images into phase images.

Despite notable progress, existing UCIR methods still suffer from the domain gap inherent in cross-domain data. This challenge arises from the entanglement of domain-specific styles and domain-invariant objects in images. For instance, as illustrated in Fig. 1 (a), "an infograph of a bear" contains dense textual descriptions, whereas "a sketch of a bear" is rendered in simple lines. As existing methods align cross-domain features for the entire image, the prevalence of domain-specific styles can overwhelm the semantically important objects, leading to performance degradation. To enable effective UCIR, it is highly expected that the domain-invariant object features could be disentangled from domain-specific styles.

However, disentangling the semantic and style features is particularly challenging in the unsupervised setting Mathieu et al. (2016). As a classic solution, β-VAE Higgins et al. (2017) suggests that unsupervised feature disentanglement can be achieved based on assumptions of prior distributions. Instead of relying on carefully designed priors, we leverage a text-to-image model as a disentangler. The intuition nestles in the following: if a text-to-image model can synthesize images that faithfully reflect both the style and object specified in a prompt, it should likewise be capable of disentangling style and object from an image in the token space. The disentangled object features offer clean semantic alignment across domains. Compared with exhaustively mitigating the domain gap from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

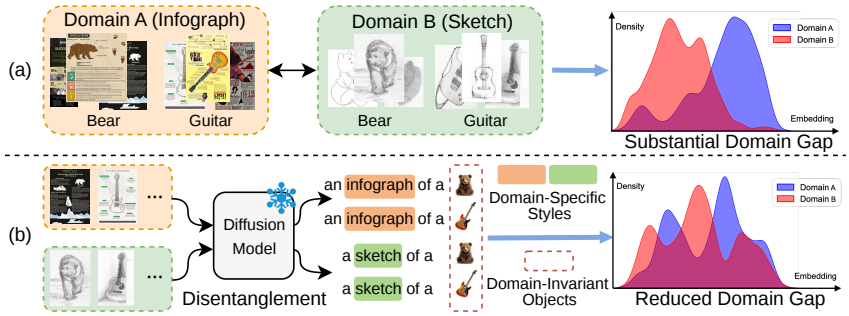


Figure 1: Our key idea. (a) Existing UCIR methods align cross-domain features for the entire image, in which domain-specific styles (*e.g.*, infograph texts and sketch lines) often overshadow the object features, leading to a substantial domain gap. (b) Our method mitigates this gap by disentangling domain-invariant object features from domain-specific styles. In the figure, the right part illustrates the feature density distributions across two domains, before and after disentanglement.

entangled image features, it would yield twice the effect with half the effort by disentangling object features from domain-specific styles before the cross-domain alignment.

Based on such a motivation, we propose DUDE, a **D**iffusion-based **U**nsupervised cross-**D**omain image r**E**trieval method. In brief, DUDE first employs an **Object Disentanglement Module** to extract domain-invariant object features, leveraging style-specific prompts to counteract domain styles. Specifically, DUDE formulates textual prompts using a handcrafted domain-style token (*e.g.*, painting, infograph, sketch) and a learnable object-semantic token. By deducing the textual prompt that reproduces the same image semantics, DUDE encourages the object semantic token to focus on domain-invariant object features present in the image, given that the image style is encapsulated by the domain-style token.

As illustrated in Fig. 1(b), the disentangled object features effectively narrow the domain gap. To further enhance cross-domain alignment of these object features, DUDE incorporates a **Progressive Alignment Module**. By progressively constructing positive contrastive pairs from single instances, in-domain mutual neighbors, to cross-domain mutual neighbors, DUDE achieves reliable and accurate UCIR. Thanks to the proposed disentangle-then-align paradigm, DUDE achieves superior performance across 13 domains on three widely used benchmarks.

The main contributions of this work could be summarized as follows:

- To tackle the core domain gap challenge in UCIR, we propose disentangling domain-invariant object features from domain-specific styles. To the best of our knowledge, this could be the first effort to investigate diffusion-guided feature disentanglement for solving UCIR.
- To further align the disentangled object features, we introduce a progressive alignment strategy that gradually expands the positive contrastive pairs, from single instances, within-domain mutual neighbors, to cross-domain mutual neighbors, leading to more accurate UCIR results.
- The proposed DUDE outperforms existing UCIR methods by a substantial margin on three standard benchmarks over 13 domains. For instance, DUDE achieves an average P@200 improvement of 18.78% compared with the best competitor on DomainNet.

2 Related Work

In this section, we briefly review two fields related to this work, namely, cross-domain image retrieval (CIR) and stylized image generation.

2.1 CROSS-DOMAIN IMAGE RETRIEVAL

CIR plays a pivotal role in computer vision across many real-world scenarios, where the query and database images originate from different domains. For example, users may submit freehand sketches as queries to retrieve photos from a database of real photographs with the same category Smeulders et al. (2000); Chaudhuri et al. (2023); Sain et al. (2023). Although CIR methods have demonstrated strong performance, they heavily rely on large-scale annotations that are labor-intensive to obtain Gajic & Baldrich (2018); Bao et al. (2023); Wang et al. (2023a); Ji et al. (2017); Kucer & Murray (2019). Such a limitation has aroused several recent efforts in unsupervised cross-domain image retrieval (UCIR) Wang et al. (2023c); Yue et al. (2021); Wang et al. (2023b); Hu & Lee (2022); Li et al. (2024a). For instance, DD combines intra-domain representation learning with cross-domain alignment to reduce the domain gap Hu & Lee (2022). ProtoOT leverages optimal transport to capture intra-domain consistency and cross-domain correspondence simultaneously Li et al. (2024a). Another branch of study adjusts spatial channel-wise or adjusts frequency components of images to narrow the domain gap Liu et al. (2023); Xu et al. (2021); Nam et al. (2021). For example, a recent work finds that the domain gap could be reduced in phase images, and thereby designs phase-enhanced contrastive learning to facilitate UCIR Hu et al. (2023). More recently, foundation-model-based priors, such as CLIP-based semantic representations, have also been explored for UCIR, further extending the conventional ImageNet-pretrained paradigm Yang et al. (2026).

Distinct from existing UCIR studies that focus on designing new feature extraction and alignment strategies for the entire image, this work attempts to solve UCIR from a novel disentanglement perspective. Namely, we propose to extract retrieval-relevant object features by disentangling them from domain-specific styles. By extracting clean object semantics before performing cross-domain alignment, our method significantly enhances the retrieval performance.

2.2 STYLIZED IMAGE GENERATION

Stylized image generation, a task that involves generating images of the given specific style Wang et al. (2023d), has been widely adopted across diverse applications, including creative design Karras et al. (2020); Li et al. (2024b), digital art Xu et al. (2023); Le & Carlsson (2022), and short video entertainment Zhang et al. (2023b); Wu et al. (2023). To control the style in image generation, existing methods resort to style-specific fine-tuning Keskar et al. (2019); Lu et al. (2023), attribute control via conditional variational autoencoders Kaneko et al. (2017); Shao et al. (2020), and prompt-based techniques Chai et al. (2022); Jeong et al. (2024). For text-to-image generation, style control is commonly achieved by fine-tuning the U-Net on target-style images, as in Stable Diffusion Rombach et al. (2022), or through lightweight fine-tuning methods such as low-rank adaptation Hu et al. (2022), which injects trainable rank-decomposition matrices for efficient model adaptation.

In addition to generation, diffusion generative models have also demonstrated potential in representation learning Yu et al. (2024); Dhariwal & Nichol (2021); Zhang et al. (2023a), as well as discriminative tasks such as classification Prabhudesai et al. (2023) and segmentation Amit et al. (2021); Rahman et al. (2023). Beyond existing applications, this work exhibits diffusion models' capability in cross-domain image retrieval, exploring and mining their feature disentanglement power. We hope this work could inspire further research toward the largely untapped potential of diffusion models for tasks beyond generation.

3 METHOD

In this section, we elaborate on the proposed method, DUDE, as illustrated in Fig. 2. In brief, Sec. 3.1 provides a brief overview of the preliminaries on unsupervised cross-domain image retrieval (UCIR) and diffusion models. Then, Sec. 3.2 introduces the Object Disentanglement Module built upon Stable Diffusion (SD) to disentangle object features from domain-specific styles. Next, Sec. 3.3 presents a Progressive Alignment Module that constructs positive contrastive pairs from instances, then in-domain neighbors, and gradually cross-domain neighbors, to achieve stable and accurate cross-domain alignment.

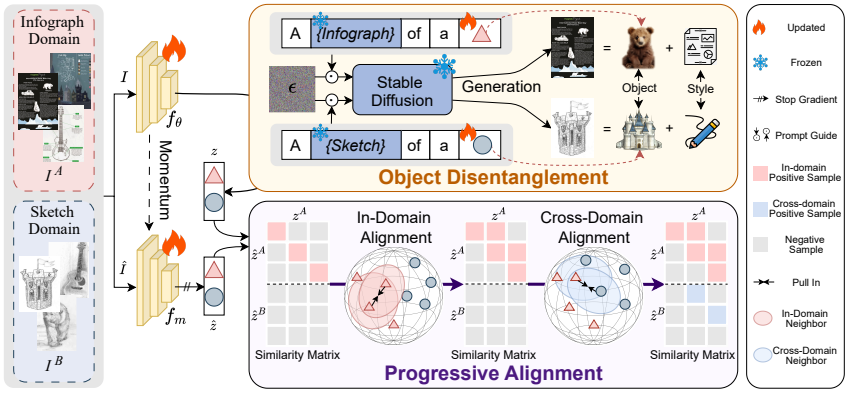


Figure 2: An overview of the proposed DUDE, which consists of two modules: (a) The **Object Disentanglement Module** leverages a frozen Stable Diffusion and varying prompts with different domain-specific tokens (*e.g.*, $\{infograph\}$ and $\{sketch\}$). It guides the feature extractor f_θ to extract object features by disentangling them from domain-specific styles. (b) To achieve reliable cross-domain alignment, the **Progressive Alignment Module** encourages the clean object features z to align with their augmentations, then with in-domain mutual neighbors, and progressively with cross-domain mutual neighbors.

3.1 Preliminary

We begin with the problem formulation for the UCIR task and a brief overview of SD.

Problem Formulation. In the UCIR task, we deal with unlabeled images from two domains $D^A = \{I_i^A\}_{i=1}^n, D^B = \{I_i^B\}_{i=1}^m$. Given a query image I_i^A belonging to category C, the goal of UCIR is to retrieve images of the same category C from domain B. To achieve this, it is required to train a valid feature extractor $f_\theta : I_i \to z_i$ which projects the input image to a common embedding space.

Stable Diffusion. SD incorporates a Variational Autoencoder (VAE) Doersch (2016) γ_θ to first encode an image $x \in \mathbb{R}^{512\times512}$ into a latent representation $x_0 \in \mathbb{R}^{64\times64}$, and then reconstruct an image from the latent representation. To enable high-quality reconstruction, SD employs a denoising process where Gaussian noise is progressively added to the latent representation during the forward process and then removed step by step across T timesteps during the reverse process.

Formally, at timestep $t \in \{1, 2, \ldots, T\}$, the noisy latent x_t is computed as:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{1}$$

where x_0 is the latent representation, $\bar{\alpha}_t$ is a pre-defined noise schedule parameter, and ϵ is Gaussian noise sampled from $\mathcal{N}(0, I)$.

To denoise x_t at this timestep, the model is trained to predict the added noise using the following objective:

$$\min_{\phi} \mathbb{E}_{\epsilon,x,c,t}\left[\|\epsilon - \epsilon_\phi(x_t, c, t)\|_2^2\right], \tag{2}$$

where c is the encoded text prompt, and ϵ_ϕ denotes the noise predictor parameterized by ϕ.

3.2 Object Disentanglement

For the UCIR problem, directly extracting features from the entire image often leads to entanglement of domain-invariant objects with domain-specific styles such as sketch lines, textual elements, or richly colored patterns, which degrades cross-domain retrieval performance. Therefore, a critical step is to disentangle domain-invariant objects from these domain-specific styles.

Recalling that SD can generate consistent objects under different stylistic prompts, this capacity suggests an implicit separation between objects and styles within the prompt token space. Inspired by this, we propose a diffusion-based object disentanglement strategy, leveraging the stylized image generation capabilities of SD to disentangle object features from domain styles.

Let $f_\theta : I_i \to z_i$ be a feature extractor that projects an image I_i into the token space. For each image I_i^d from domain D^d ($d \in \{A, B\}$), we construct a prompt: "a $\{domain\}$ of a $[z_i^d]$." Here, $\{domain\}$ is the domain-style token (*e.g.*, painting, infograph, sketch) shared among all samples in domain D^d, while the object-semantic token $[z_i^d]$ encodes the feature $z_i^d = f_\theta(I_i^d)$. This prompt is used to guide the pre-trained SD in reconstructing the image from a noisy latent version of I_i^d. Rather than optimizing the image generation capability of SD itself, we freeze all parameters of SD, including the noise predictor ϵ_ϕ and the VAE encoder γ_ϕ, and train the feature extractor f_θ. The training objective is to compress the domain-invariant object feature in I_i^d into the object-semantic token $[z_i^d]$.

Formally, given an image I_i^d, we first encode I_i^d into a latent representation using the frozen VAE γ_ϕ and calculate the noisy latent at a diffusion timestep t as:

$$x_{i,t}^d = \sqrt{\bar{\alpha}_t}\,\gamma_\phi(I_i^d) + \sqrt{1-\bar{\alpha}_t}\,\epsilon. \tag{3}$$

Then, we compute the disentanglement loss as:

$$\mathcal{L}_{\text{OD}} = \mathbb{E}_{\epsilon,t}\left[\left\|\epsilon - \epsilon_\phi(x_{i,t}^d, c_\theta, t)\right\|_2^2\right], \tag{4}$$

where c_θ is the prompt embedding constructed from the frozen text encoder using the feature z_i^d inserted into the prompt "a $\{domain\}$ of a $[z_i^d]$", and ϵ_ϕ denotes the frozen pre-trained SD noise predictor.

Such a design inverts the typical use of generative models. Rather than generating from text, SD is leveraged as a disentangler of domain-invariant object features $[z_i^d]$ and domain-specific styles $\{domain\}$, as gradients are backpropagated to solely optimize f_θ.

3.3 Progressive Alignment

Thanks to the Object Disentanglement Module, the feature extractor f_θ effectively captures object features while stripping away domain-specific styles. These clean object features, when properly aligned, are expected to form compact clusters for the same class, regardless of their domain of origin. To facilitate reliable and effective alignment, we propose a progressive alignment strategy that constructs positive contrastive samples in stages, from instance to in-domain mutual nearest neighbors, and then to cross-domain mutual nearest neighbors.

First, contrastive learning is conducted between each image and its augmentation, denoted as $\{I_i^d, \hat{I}_i^d\} \in D^d$, where $d \in \{A, B\}$ indicates the domain index. The clean object features are computed as $z_i^d = f_\theta(I_i^d)$ and $\hat{z}_i^d = f_m(\hat{I}_i^d)$, using the feature extractor f_θ and the momentum extractor f_m, respectively. The momentum extractor is updated following the MoCo strategy He et al. (2020), and its outputs are stored in domain-specific memory banks $\mathcal{Q}^d$.

In-domain Alignment. The most reliable alignment occurs when only augmented views of the same sample are treated as positive pairs, while all other samples in the dataset are considered negatives. The loss function is defined as:

$$\mathcal{L}_{\text{aug}} = \sum_{i \in \mathbf{I}} -\log \frac{\exp\left(z_i^\top \hat{z}_i/\tau\right)}{\sum_{a \in \mathbf{I}} \exp\left(z_i^\top \hat{z}_a/\tau\right)}, \tag{5}$$

where the index set $\mathbf{I}$ represents all the samples in the same domain, and τ is the temperature parameter.

Next, we extend the alignment progressively to in-domain neighbors. To capture feature relationships within a domain, we define a directed neighborhood graph for each sample z_i^d based on the top-k cosine similarity:

$$\mathcal{N}_k^d(i) = \text{Top-}k\left(\left\{j \mid \cos\left(z_i^d, z_j^d\right)\right\}\right). \tag{6}$$

To ensure mutual reliability between neighbors, we define a symmetric adjacency matrix $\mathbf{M}_{\text{in}}^d \in \{0,1\}^{n^d \times n^d}$, where $n^d = |D^d|$. Then, it is defined as:

$$[\mathbf{M}_{\text{in}}^d]_{ij} = \begin{cases} 1, & \text{if } i \in \mathcal{N}_k^d(j) \text{ and } j \in \mathcal{N}_k^d(i), \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

This adjacency matrix captures only those sample pairs that mutually recognize each other as top-k neighbors, filtering noisy or incidental neighbors, thereby promoting more reliable alignment.

The in-domain alignment loss is formulated using an InfoNCE Oord et al. (2018) objective, where only mutual neighbors in $\mathbf{M}_{\text{in}}^d$ contribute as positives:

$$\mathcal{L}_{\text{in}}^d = -\frac{1}{n^d}\sum_{i=1}^{n^d}\frac{\sum_{j=1}^{n^d}([\mathbf{M}_{\text{in}}^d]_{ij}\cdot\ell_{ij}^d)}{\sum_{j=1}^{n^d}[\mathbf{M}_{\text{in}}^d]_{ij}+\epsilon}, \tag{8}$$

$$\ell_{ij}^d = \frac{(z_i^d)^\top \hat{z}_j^d}{\tau} - \log\sum_{q=1}^{|\mathcal{Q}^d|}\exp\left(\frac{(z_i^d)^\top \hat{z}_q^d}{\tau}\right), \tag{9}$$

where $\hat{z}_j^d$ and $\hat{z}_q^d$ represent the features in memory bank, τ denotes the temperature parameter, and ϵ denotes a small constant to avoid division by zero. By combining the in-domain losses from both domains, the total in-domain loss is defined as:

$$\mathcal{L}_{\text{in}} = \mathcal{L}_{\text{in}}^A + \mathcal{L}_{\text{in}}^B. \tag{10}$$

The progressive loss in the first stage is defined as:

$$\mathcal{L}_{\text{PA1}} = \mathcal{L}_{\text{aug}} + \beta\mathcal{L}_{\text{in}}. \tag{11}$$

Cross-domain Alignment. Once strong in-domain alignment is established, we extend the alignment to cross-domain neighbors. Here, we only describe the alignment process from domain A to domain B, as the process from domain B to domain A is symmetric. For each feature z_i extracted from image $I_i^A \in D^A$, we identify its top-k nearest neighbors in domain B:

$$\mathcal{N}_k^{A\to B}(i) = \text{Top-}k\left(\left\{j \mid \cos\left(z_i^A, z_j^B\right)\right\}\right). \tag{12}$$

To reduce noisy neighbors, we further enforce a mutual cross-domain adjacency matrix $\mathbf{M}_{\text{cross}}^{A\to B} \in \{0,1\}^{n\times m}$:

$$[\mathbf{M}_{\text{cross}}^{A\to B}]_{ij} = \begin{cases} 1, & \text{if } i \in \mathcal{N}_k^{B\to A}(j) \\ & \text{and } j \in \mathcal{N}_k^{A\to B}(i), \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

This matrix captures only the mutually nearest neighbor pairs across domains, ensuring that alignment occurs only between features with high similarity. The cross-domain loss from domain A to domain B is:

$$\mathcal{L}_{\text{cross}}^{A\to B} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}[\mathbf{M}_{\text{cross}}^{A\to B}]_{ij}\cdot\ell_{ij}^{A\to B}}{\sum_{j=1}^{m}[\mathbf{M}_{\text{cross}}^{A\to B}]_{ij}+\epsilon}, \tag{14}$$

$$\ell_{ij}^{A\to B} = \frac{(z_i^A)^\top \hat{z}_j^B}{\tau} - \log\sum_{q=1}^{|\mathcal{Q}^B|}\exp\left(\frac{(z_i^A)^\top \hat{z}_q^B}{\tau}\right), \tag{15}$$

where $\hat{z}_j^B$ and $\hat{z}_q^B$ denote the features in memory bank $\mathcal{Q}^B$, and τ is the temperature parameter. The cross-domain loss from domain B to domain A $\mathcal{L}_{\text{cross}}^{B\to A}$ is computed symmetrically. Consequently, the total loss for cross-domain alignment is given by:

$$\mathcal{L}_{\text{cross}} = \mathcal{L}_{\text{cross}}^{A\to B} + \mathcal{L}_{\text{cross}}^{B\to A}. \tag{16}$$

The progressive loss in the second stage is defined as:

$$\mathcal{L}_{\text{PA2}} = \mathcal{L}_{\text{in}} + \lambda\mathcal{L}_{\text{cross}}. \tag{17}$$

After training, the outputs of f_θ are used for cross-domain retrieval.

4 EXPERIMENTS

In this section, we evaluate the performance of our proposed DUDE against existing baselines on three datasets. Following this, we conduct ablation studies, parameter analysis, and visualization to further validate the effectiveness of DUDE.

Table 1: Average retrieval performance on PACS, Office-Home, and DomainNet.

	Average	ID	ProtoNCE	CDS	PCS	DD	CoDA	ProtoOT	**DUDE**	Improvement
PACS	P50	17.72	38.91	33.79	36.42	45.87	41.04	50.26	**71.09**	**+20.83**
	P@100	17.15	28.75	31.50	34.03	43.01	38.25	48.78	**67.49**	**+18.71**
	P@200	16.85	25.56	28.54	31.12	38.54	34.63	46.38	**61.78**	**+15.40**
Office-Home	P@1	33.18	35.49	37.17	38.07	40.67	40.68	43.85	**52.36**	**+8.51**
	P@5	29.76	32.15	34.63	35.01	37.60	35.06	41.29	**50.84**	**+9.55**
	P@15	25.78	28.30	31.30	31.42	34.15	31.39	37.98	**48.04**	**+10.06**
DomainNet	P@50	43.96	43.67	45.98	47.73	50.82	39.25	66.63	**82.62**	**+15.99**
	P@100	38.80	37.05	38.37	41.47	45.54	33.13	62.98	**80.07**	**+17.09**
	P@200	32.37	30.48	31.05	34.38	39.10	32.67	56.12	**74.90**	**+18.78**

Table 2: Ablation Study on our loss components. The best performance results are in **bold**.

$\mathcal{L}_{\text{OD}}$	$\mathcal{L}_{\text{PA1}}$	$\mathcal{L}_{\text{PA2}}$	PACS			Office-Home			DomainNet			Average		
			P@50	P@100	P@200	P@1	P@5	P@15	P@50	P@100	P@200	P@50(1)	P@100(5)	P@200(15)
ResNet50			32.28	29.83	26.89	32.62	29.09	25.02	40.49	34.18	28.10	35.13	31.03	26.66
✓			46.17	43.20	39.39	39.17	34.56	30.33	65.90	62.59	56.08	50.41	46.78	41.93
	✓		41.19	37.55	32.06	43.28	39.99	36.23	52.78	45.82	37.40	45.75	41.12	35.23
		✓	47.67	44.08	39.02	35.88	32.50	28.76	53.49	48.21	40.92	45.68	41.60	36.23
✓	✓		70.97	67.71	60.19	49.86	48.53	45.73	76.56	73.89	67.96	65.80	63.38	57.96
✓		✓	50.23	47.85	44.32	44.83	42.13	38.62	79.16	77.06	71.45	58.07	55.68	51.46
	✓	✓	43.06	39.34	33.76	44.52	41.59	37.37	57.89	51.67	42.88	48.49	44.20	38.00
✓	✓	✓	**71.09**	**67.79**	**61.78**	**52.36**	**50.84**	**48.04**	**82.62**	**80.07**	**74.90**	**68.69**	**66.23**	**61.57**

4.1 Comparison with State-of-the-Art Methods

Tabs. 5-7 provide detailed retrieval performance on PACS, Office-Home and DomainNet, respectively. As summarized in Tab. 1, DUDE achieves substantial performance improvements across all three datasets. On Office-Home, DUDE outperforms the leading baseline ProtoOT Li et al. (2024a) with average gains of 8.50%, 9.55%, and 10.06% in P@1, P@5, and P@15, respectively. On DomainNet, DUDE achieves average gains of 15.99%, 17.08%, and 18.78% in P@50, P@100, and P@200, respectively. The improvements are even more pronounced on PACS, where DUDE achieves average gains of 20.83%, 18.71%, and 15.40% in P@50, P@100, and P@200, respectively. The results demonstrate that the proposed method significantly enhances the performance of unsupervised cross-domain image retrieval, thanks to the disentangle-then-align strategy that narrows the domain gaps.

4.2 Ablation Study

To understand the impact of each component in the proposed DUDE, we conduct ablation studies on loss components. The results in Tab. 2 demonstrate the effectiveness of different loss components in our DUDE. From the results, one could arrive at three conclusions: *i)* Comparing line 1 (a baseline using the pre-trained ResNet50 features) and line 2 shows that adopting $\mathcal{L}_{\text{OD}}$ alone leads to a substantial performance improvement, indicating that our disentanglement strategy helps the feature extractor learn clean, domain-invariant object features and significantly enhances feature discriminability; *ii)* Comparing Lines 3, 4, and 7, we first observe that the improvement from Line 3 to Line 7 underscores the benefit of our gradual transition from intra- to inter-domain alignment. However, compared to Line 7, Line 4 yields fluctuating results. For instance, a low P@1 of 35.88% on Office-Home versus a high P@50 of 47.67% on PACS. This disparity arises because the initial pre-trained ResNet50 features in Office-Home are not well aligned within domains, whereas PACS exhibits strong in-domain alignment. As a result, relying solely on $\mathcal{L}_{\text{PA2}}$ benefits PACS but leads to suboptimal optimization on Office-Home. These findings further underscore the necessity of our stable progressive alignment strategy; *iii)* Comparing lines 7 and 8, introducing $\mathcal{L}_{\text{OD}}$ leads to a notable performance boost, such as a 27% gain in P@50 on PACS. Similarly, comparing lines 2 and 8, jointly applying both $\mathcal{L}_{\text{PA1}}$ and $\mathcal{L}_{\text{PA2}}$ results in an improvement such as a 25% gain in P@50 on PACS. The optimal performance is observed when all three loss components are applied jointly, confirming that object disentanglement and progressive alignment are both indispensable.

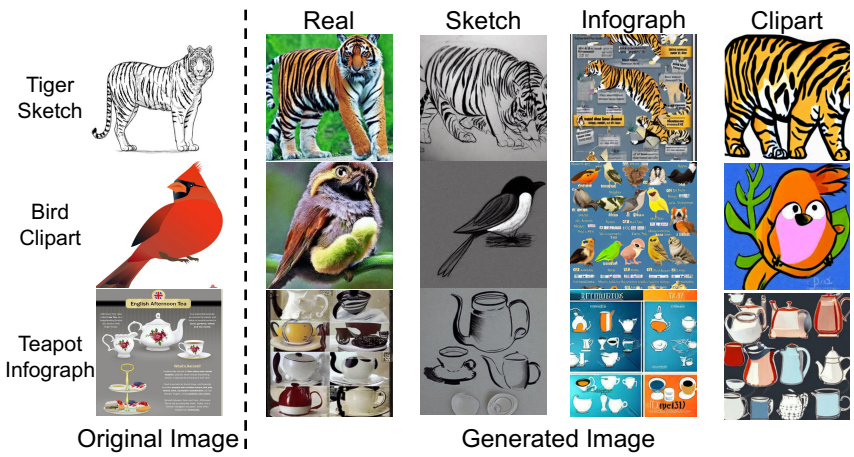
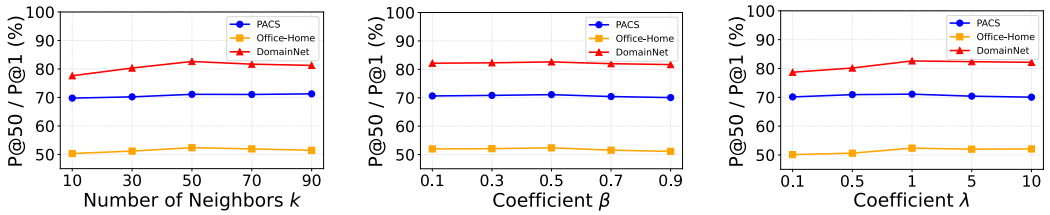


Figure 3: Cross-domain generation results using disentangled object features from one domain combined with other domain prompts. A Prompt-swapping evidence for object-style disentanglement. Each row uses the object token $[z]$ extracted from a source image, while each column changes only the domain prompt. The generated images preserve the source object semantics across columns but adopt the style specified by the target domain prompt, providing qualitative evidence that object semantics are primarily preserved in $[z]$, while domain styles are largely controlled by $\{domain\}$.



(a) Analysis on neighbors k. (b) Analysis on coefficient β. (c) Analysis on coefficient λ.

Figure 4: Parameter analysis on neighbors k, coefficient β, and coefficient λ.

4.3 Analysis of Computational Cost

To clarify the computational cost of our DUDE, we compare it with the competitive baseline ProtoOT. As shown in Tab. 3, DUDE adopts a two-stage training strategy, where the first stage takes around 236.1 s per epoch and the second stage takes around 20.7 s per epoch. Although DUDE uses additional memory, it requires less total training time than ProtoOT. More importantly, DUDE achieves substantially better retrieval performance, improving P@50 from 50.26% to 71.09%, with an absolute gain of 20.83%. This demonstrates that DUDE achieves a better trade-off between training cost and retrieval performance.

Table 3: Comparisons of training cost and retrieval performance.

Method	Stage	Epoch Time	Epochs	Total Time	Memory	P@50
ProtoOT	–	109.6 s	200	21,920 s	14,800 MB	50.26%
DUDE	Stage 1	236.1 s	50	12,841 s	19,106 MB	71.09%
	Stage 2	20.7 s	50			

4.4 Parameter Analysis

To evaluate the robustness of DUDE, we conduct experiments by varying three key parameters: the number of neighbors k, the loss coefficient β in $\mathcal{L}_{\text{PA1}}$ (Eq. 11), and the coefficient λ in $\mathcal{L}_{\text{PA2}}$ (Eq. 17). The results are shown in Fig. 4.

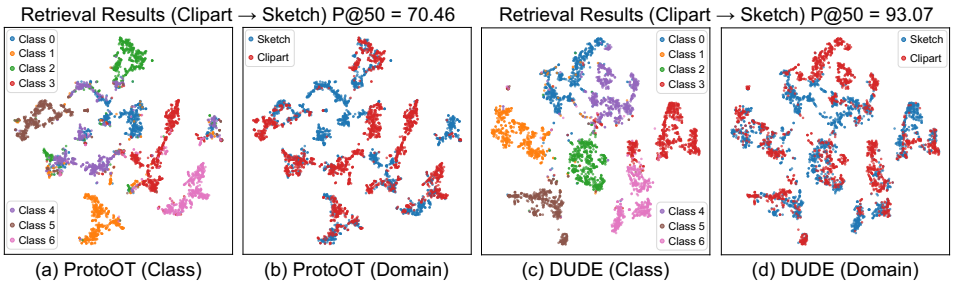


Figure 5: T-SNE visualization of learned features on Clipart-Sketch in DomainNet. (a) and (b) show the class- and domain-colored results of ProtoOT; (c) and (d) show the corresponding results of our DUDE.

Recall that DUDE aligns only those sample pairs that mutually identify each other as top-k neighbors. In Fig. 4a, we evaluate the effect of varying k. As k increases, the retrieval performance initially improves and then declines slightly. This can be attributed to two competing factors: on the one hand, larger k introduces more potential true positive neighbors; on the other hand, overly large k brings in noisy neighbors. Nonetheless, our mutual top-k neighbor strategy remains robust. To evaluate the influence of β, we vary its value in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. As shown in Fig. 4b, DUDE demonstrates stable performance across all the values. For λ, we evaluate values in $\{0.1, 0.5, 1, 5, 10\}$, as shown in Fig. 4c. DUDE performs well when λ falls across this range. Based on these observations, we set $k = 50$, $\beta = 0.5$, and $\lambda = 1$ for all experiments.

4.5 Visualization

To provide a more intuitive comparison between our DUDE and the most promising state-of-the-art method, ProtoOT Li et al. (2024a), we present t-SNE visualizations on the Clipart-Sketch pair from DomainNet.

To further illustrate the discriminative power of the learned embeddings, we visualize t-SNE projections for the Clipart-Sketch domains in DomainNet, colored by class and by domain. As shown in Fig. 5, (a) and (b) represent the class-wise and domain-wise t-SNE of ProtoOT, while (c) and (d) correspond to those of DUDE. *i)* Comparing (a) and (c), DUDE achieves more compact intra-class distributions and clearer inter-class separations, indicating improved class discriminability. *ii)* Comparing (b) and (d), in DUDE t-SNE, features exhibit better aggregation between different domains. This suggests that DUDE suppresses domain-specific style in the learned representation.

5 Conclusion

In this work, we present DUDE, a novel unsupervised cross-domain image retrieval (UCIR) method that explicitly addresses the domain gap through feature disentanglement. Unlike prior UCIR methods that extract and align features for the entire image, DUDE disentangles the domain-invariant object features from the domain-specific styles, by mining the stylized image generation capability of diffusion models in reverse with the help of style-varying prompts. Further empowered by the proposed progressive alignment strategy, DUDE achieves reliable and accurate cross-domain retrieval results, significantly outperforming existing UCIR baselines. Looking ahead, DUDE not only offers a new perspective for tackling the UCIR task through feature disentanglement, but also opens up new application avenues for diffusion models beyond the image generation task.

6 AI use statement

Generative AI tools were used to assist with data analysis, language polishing, translation, and improving the clarity and readability of the manuscript. All AI-assisted analyses and text were independently reviewed and verified by the authors. The authors take full responsibility for the methodology, results, claims, and final content of this paper.

REFERENCES

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Chen Bao, Xudong Zhang, Jiazhou Chen, and Yongwei Miao. Mmfl-net: multi-scale and multi-granularity feature learning for cross-domain fashion retrieval. *Multimedia Tools and Applications*, 82(24):37905–37937, 2023.

Junyi Chai, Reid Pryzant, Victor Ye Dong, Konstantin Golobokov, Chenguang Zhu, and Yi Liu. Fast: Improving controllability for text generation with feedback aware self-training. *arXiv preprint arXiv:2210.03167*, 2022.

Abhra Chaudhuri, Ayan Kumar Bhunia, Yi-Zhe Song, and Anjan Dutta. Data-free sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 12084–12093, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1869–1871, 2018.

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*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *European Conference on Computer Vision*, pp. 529–544. Springer, 2022.

Conghui Hu, Can Zhang, and Gim Hee Lee. Unsupervised feature representation learning for domain-generalized cross-domain image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11016–11025, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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*, pp. 1062–1070, 2015a.

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*, pp. 1062–1070, 2015b.

Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention, 2024. URL `https://arxiv.org/abs/2402.12974`.

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*, pp. 1654–1662, 2017.

Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6089–6098, 2017.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9123–9132, 2021.

Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6850–6861, June 2023.

Michal Kucer and Naila Murray. A detect-then-retrieve model for multi-domain fashion item retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Minh-Ha Le and Niklas Carlsson. Styleid: Identity disentanglement for anonymizing faces. *arXiv preprint arXiv:2212.13791*, 2022.

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*, volume 38, pp. 3009–3017, 2024a.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Xiang Li, Lei Meng, Lei Wu, Manyi Li, and Xiangxu Meng. Dreamfont3d: personalized text-to-3d artistic font generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Shaolei Liu, Siqi Yin, Linhao Qu, and Manning Wang. Reducing domain gap in frequency and spatial domain for cross-modality domain adaptation on medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1719–1727, 2023.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Specialist diffusion: Plug-and-play sample-efficient fine-tuning of text-to-image diffusion models to learn any unseen style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14267–14276, 2023.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in neural information processing systems*, 29, 2016.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8690–8699, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Shuxin Ouyang, Timothy M Hospedales, Yi-Zhe Song, and Xueming Li. Forgetmenot: Memory-aware forensic facial sketch matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5571–5579, 2016.

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*, pp. 1406–1415, 2019.

Mihir Prabhudesai, Tsung-Wei Ke, Alexander Cong Li, Deepak Pathak, and Katerina Fragkiadaki. Test-time adaptation of discriminative models via diffusion generative feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=gUTVpByfVX.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11536–11546, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2765–2775, 2023.

Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. Controlvae: Controllable variational autoencoder. In *International conference on machine learning*, pp. 8655–8664. PMLR, 2020.

Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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*, pp. 5018–5027, 2017.

Haixin Wang, Hao Wu, Jinan Sun, Shikun Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Idea: An invariant perspective for efficient domain adaptive image retrieval. *Advances in Neural Information Processing Systems*, 36:57256–57275, 2023a.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16041–16050, 2022.

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*, volume 37, pp. 10200–10208, 2023b.

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*, volume 37, pp. 10200–10208, 2023c.

Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model. *arXiv preprint arXiv:2309.01770*, 2023d.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633, 2023.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Jiaolong Xu, Liang Xiao, and Antonio M López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14383–14392, 2021.

Wenju Xu, Chengjiang Long, and Yongwei Nie. Learning dynamic style kernels for artistic style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10083–10092, 2023.

Jing Yang, Hui Xue, Shipeng Zhu, and Pengfei Fang. Text-phase synergy network with dual priors for unsupervised cross-domain image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23891–23900, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13834–13844, 2021.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023a.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023b.

Xiaoping Zhou, Xiangyu Han, Haoran Li, Jia Wang, and Xun Liang. Cross-domain image retrieval: methods and applications. *International Journal of Multimedia Information Retrieval*, 11(3): 199–218, 2022.

APPENDIX OF DUDE: DIFFUSION-BASED UNSUPERVISED CROSS-DOMAIN IMAGE RETRIEVAL

In this appendix, we provide additional details and analyses to support the main paper. Specifically, Appendix A describes all disentanglement prompts. Appendix B provides a theoretical interpretation of the proposed diffusion-guided object disentanglement objective. Appendix C presents the datasets, implementation details, evaluation metrics, baseline methods, and comparisons between DUDE and existing baselines on three benchmark datasets. Appendix D reports an ablation study of the disentanglement prompts. Finally, Appendix E presents qualitative retrieval results on all three datasets, further demonstrating the effectiveness of DUDE for cross-domain image retrieval.

A DISENTANGLEMENT PROMPT SET

Our DUDE is built upon a text-to-image model, Stable Diffusion (SD). SD requires an image and a text condition as input. For each image in different domains, we construct a prompt with an object-semantic token as a text condition. The object-semantic token $[z]$ represents the output of feature extractor f_θ. Tab. 4 summarizes the prompt set for three datasets.

Table 4: Disentanglement prompt set used in three datasets.

Dataset	Domain	Prompt
PACS	Art Painting	a painting of a $[z]$
	Cartoon	a cartoon of a $[z]$
	Photo	a photo of a $[z]$
	Sketch	a sketch of a $[z]$
Office-Home	Art	a painting of a $[z]$
	Clipart	a clipart of a $[z]$
	Real	a photo of a $[z]$
	Product	a product photo of a $[z]$
DomainNet	Clipart	a clipart of a $[z]$
	Sketch	a sketch of a $[z]$
	Painting	a painting of a $[z]$
	Infograph	an infograph of a $[z]$
	Real	a photo of a $[z]$

B THEORETICAL INTERPRETATION OF DIFFUSION-GUIDED OBJECT DISENTANGLEMENT

We provide a theoretical interpretation of the proposed object disentanglement objective $\mathcal{L}_{OD}$. Rather than claiming that $\mathcal{L}_{OD}$ universally guarantees disentanglement, we show that, under conditional object identifiability and a minimum-sufficiency principle, its non-redundant minimizer preserves object information while discarding domain-style information already specified by the prompt.

Problem definition Let the image latent x_0 be generated from two factors:

$$x_0 \sim p(x_0 \mid O, S), \tag{18}$$

where O denotes retrieval-relevant object information and S denotes domain-specific style. In DUDE, the style factor is represented by a domain-style descriptor

$$D = g(S), \tag{19}$$

such as "photo", "sketch", or "infograph". The learnable representation $Z = f_\theta(x_0)$ is inserted into the prompt as the object token, and the frozen diffusion model is conditioned on $C = (D, Z)$.

Given Gaussian noise ϵ and timestep t, the noisy latent is

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \tag{20}$$

The object disentanglement loss is

$$\mathcal{L}_{OD} = \mathbb{E}_{\epsilon,t}\left[\|\epsilon - \epsilon_\phi(x_t, D, Z, t)\|_2^2\right], \tag{21}$$

where ϵ_ϕ is frozen. Thus, optimizing $\mathcal{L}_{OD}$ only updates f_θ.

For squared-error denoising, the Bayes-optimal predictor under condition C is

$$\epsilon^\star(x_t, C, t) = \mathbb{E}[\epsilon \mid x_t, C, t], \tag{22}$$

and the corresponding Bayes denoising risk is

$$R(C) = \mathbb{E}\left[\mathrm{Var}(\epsilon \mid x_t, C, t)\right]. \tag{23}$$

Since $C = (D, Z)$ in DUDE, we have

$$R(D, Z) = \mathbb{E}\left[\mathrm{Var}(\epsilon \mid x_t, D, Z, t)\right]. \tag{24}$$

Therefore, minimizing $\mathcal{L}_{OD}$ reduces the conditional uncertainty of the denoising target ϵ given (D, Z). Since D explicitly specifies the domain style, the remaining uncertainty reducible by Z should correspond to information not already provided by D, ideally the object information.

Assumption 1. Conditional object identifiability Given a domain-style descriptor $D = d$, distinct object factors induce distinct optimal denoising predictors. Specifically, for any $O_1 \neq O_2$, there exists a set of (x_t, t) with non-zero probability such that

$$\mathbb{E}[\epsilon \mid x_t, O_1, d, t] \neq \mathbb{E}[\epsilon \mid x_t, O_2, d, t]. \tag{25}$$

Assumption 2. Minimum-sufficiency principle Following the principle of minimal sufficient representation Tishby et al. (2000); Wang et al. (2022), we focus on the non-redundant representative among the minimizers of the diffusion reconstruction objective. Let

$$\mathcal{Z}_{\min} = \arg\min_Z R(D, Z) \tag{26}$$

be the set of global minimizers. Since this set may contain redundant solutions that encode both object and style information, we define the minimum-sufficient solution as

$$Z^\star \in \arg\min_{Z \in \mathcal{Z}_{\min}} I(Z; X), \tag{27}$$

where X denotes the input image. Thus, $Z^\star$ retains only the information necessary to attain the minimum denoising risk.

Object sufficiency We first show that $Z^\star$ preserves object-discriminative information. Suppose, for contradiction, that $Z^\star$ does not preserve the object factor. Then there exist $O_1 \neq O_2$ and a domain descriptor d such that

$$p(O_1 \mid Z^\star = z^\star, D = d) > 0, \quad p(O_2 \mid Z^\star = z^\star, D = d) > 0. \tag{28}$$

Under the same condition $(D = d, Z^\star = z^\star)$, the denoiser must use the same Bayes predictor for both object factors:

$$\mathbb{E}[\epsilon \mid x_t, D = d, Z^\star = z^\star, t]. \tag{29}$$

By the law of total variance,

$$\begin{aligned}\mathrm{Var}(\epsilon \mid x_t, D, Z^\star, t) = \mathbb{E}\left[\mathrm{Var}(\epsilon \mid x_t, D, Z^\star, O, t) \mid x_t, D, Z^\star, t\right] \\ + \mathrm{Var}\left(\mathbb{E}[\epsilon \mid x_t, D, Z^\star, O, t] \mid x_t, D, Z^\star, t\right).\end{aligned} \tag{30}$$

The second term is strictly positive on a set of non-zero probability by Assumption 1. Hence, merging distinct object factors increases the denoising risk:

$$R(D, Z^\star) > R(D, O), \tag{31}$$

contradicting the optimality of $Z^\star$. Therefore,

$$H(O \mid Z^\star, D) = 0. \tag{32}$$

That is, given $Z^\star$ and the domain-style descriptor D, the object factor is identifiable.

Style redundancy We next show that $Z^\star$ removes redundant domain-style information. Suppose that $Z^\star$ contains an additional style component:

$$Z^\star = (Z_O^\star, Z_S^\star), \tag{33}$$

where $Z_O^\star$ is object-relevant and $Z_S^\star$ is domain-style information. Since D already specifies the style condition used by the diffusion model, adding $Z_S^\star$ does not further reduce the conditional denoising uncertainty:

$$\mathrm{Var}(\epsilon \mid x_t, D, Z_O^\star, Z_S^\star, t) = \mathrm{Var}(\epsilon \mid x_t, D, Z_O^\star, t). \tag{34}$$

Thus,

$$R(D, Z_O^\star, Z_S^\star) = R(D, Z_O^\star). \tag{35}$$

The style component $Z_S^\star$ can therefore be removed without increasing the denoising risk, contradicting the minimum-sufficiency of $Z^\star$. Hence,

$$I(Z^\star; D \mid O) = 0. \tag{36}$$

Conclusion The analysis establishes two properties:

$$H(O \mid Z^\star, D) = 0, \qquad I(Z^\star; D \mid O) = 0. \tag{37}$$

The first states that $Z^\star$ is sufficient for identifying the object factor, while the second states that it contains no additional domain-style information after conditioning on the object. Therefore, under conditional object identifiability and minimum sufficiency, $\mathcal{L}_{OD}$ encourages an object–style disentangled representation: the prompt accounts for domain style, while the learned token captures the residual object-specific information required for diffusion reconstruction.

C Performance on Three Benchmarks

C.1 Datasets

To evaluate the effectiveness of the proposed DUDE method, we conduct experiments on three widely used datasets in cross-domain retrieval, including PACS Xu et al. (2019), Office-Home Venkateswara et al. (2017), and DomainNet Peng et al. (2019). Specifically, PACS consists of four different domains (Photo, Art Painting, Cartoon, Sketch) with seven categories. Office-Home comprises four domains (Art, Clipart, Product, Real) encompassing 65 categories. For DomainNet Peng et al. (2019), we utilize five domains (Clipart, Infograph, Painting, Real, and Sketch). Consistent with prior studies Hu & Lee (2022); Wang et al. (2023b); Li et al. (2024a), we utilize seven categories that contain over 200 images in DomainNet.

C.2 Implementation Details

Following previous UCIR studies, all compared methods use the same ImageNet-pretrained ResNet-50 backbone He et al. (2016) to ensure fair comparisons. We utilize the parameters of f_θ to update f_m following the MoCo strategy He et al. (2020). We train our model for 100 epochs using the Adam optimizer with a learning rate of 2.5×10^{-4} and a batch size of 64. The temperature parameter τ is set to 0.2 across all datasets. Without loss of generality, we apply Stable Diffusion v1.5 Rombach et al. (2022) in the object disentanglement module, which offers a good balance between generation speed and quality. In the first 50 epochs, we warm up the feature extractor f_θ using only the object disentanglement loss $\mathcal{L}_{\text{OD}}$. Afterward, we initialize f_m with the parameters of the warmed-up f_θ, and continue training f_θ using $\mathcal{L}_{\text{PA1}}$ for the next 30 epochs. Finally, we perform cross-domain alignment using $\mathcal{L}_{\text{PA2}}$ for the remaining 20 epochs.

C.3 Evaluation Metrics

Following previous studies, we evaluate the retrieval performance on Office-Home using precision at the top 1, 5, and 15 retrieved images (denoted as P@1/5/15), and report precision at the top 50, 100, and 200 retrieved images for both PACS and DomainNet (denoted as P@50/100/200). Consistent with prior works, our task focuses on category-level cross-domain retrieval, where a retrieved image is considered correct if it belongs to the same class as the query.

Table 5: Comparison of retrieval performance (P@50, P@100, P@200) across various domains on PACS. The best results are in **bold**.

Method	Art→Cartoon			Cartoon→Art			Art→Photo			Photo→Art		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID	26.32	25.45	24.66	27.13	25.93	24.43	19.17	18.78	18.48	19.65	19.62	18.81
ProtoNCE	28.45	25.38	22.63	27.63	24.69	21.96	28.14	25.35	24.33	30.50	26.90	23.53
CDS	45.84	40.96	35.93	38.01	36.09	32.00	41.01	37.30	33.27	84.32	79.60	65.91
PCS	48.98	44.38	39.20	42.00	39.55	35.13	45.43	41.82	37.56	84.70	79.95	66.75
DD	59.85	54.63	47.62	48.49	46.60	41.78	65.38	62.46	56.36	84.93	82.02	68.25
CoDA	53.73	48.99	42.81	45.49	42.98	38.07	58.23	55.04	50.18	84.93	79.90	67.53
ProtoOT	69.97	65.94	61.37	55.95	55.91	55.08	84.23	83.99	**79.88**	86.08	85.50	**83.89**
DUDE	**88.21**	**84.03**	**77.61**	**84.76**	**82.51**	**77.06**	**87.13**	**85.44**	76.93	**93.30**	**89.83**	81.73
	Art→Sketch			Sketch→Art			Cartoon→Photo			Photo→Cartoon		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID	14.68	14.92	15.22	13.54	13.53	13.46	12.79	12.85	13.56	15.16	15.18	15.37
ProtoNCE	33.73	29.72	25.34	25.94	23.40	24.35	36.23	30.75	27.05	40.62	34.31	28.04
CDS	19.06	18.73	19.99	18.69	17.44	16.09	24.66	24.52	23.63	45.89	41.39	34.24
PCS	20.60	20.15	21.02	18.95	17.79	16.25	25.68	25.60	24.76	47.21	42.52	36.91
DD	26.54	25.72	26.25	28.68	26.04	22.78	40.91	39.74	36.57	55.19	48.83	40.27
CoDA	23.26	22.68	23.14	25.87	22.95	20.01	36.27	35.74	33.63	55.57	49.08	40.39
ProtoOT	23.51	24.06	26.65	30.41	29.86	27.16	42.04	42.71	43.18	64.74	60.25	54.58
DUDE	**69.07**	**62.88**	**58.54**	**54.56**	**51.70**	**46.21**	**59.24**	**57.36**	**52.19**	**82.13**	**76.52**	**68.40**
	Cartoon→Sketch			Sketch→Cartoon			Photo→Sketch			Sketch→Photo		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
ID	21.65	19.54	18.67	17.10	16.42	16.55	12.48	12.39	12.42	12.96	11.21	10.57
ProtoNCE	45.61	39.77	32.44	34.45	30.07	27.50	38.23	32.77	27.08	25.34	21.96	22.58
CDS	34.09	31.51	30.27	22.56	21.08	20.54	18.84	18.77	20.77	12.54	10.57	9.79
PCS	42.58	39.87	38.79	25.37	23.87	23.21	22.34	21.78	23.45	13.15	11.11	10.47
DD	44.63	42.03	40.76	44.83	40.52	35.66	25.31	24.66	26.29	25.82	22.90	19.87
CoDA	40.16	37.82	37.01	32.04	28.91	27.81	22.67	21.36	22.16	14.29	13.61	12.84
ProtoOT	50.02	47.10	46.07	53.54	50.90	44.71	27.74	24.31	19.63	14.87	14.79	14.35
DUDE	**72.06**	**69.17**	**65.78**	**61.99**	**58.06**	**54.41**	**53.56**	**50.41**	**48.63**	**47.11**	**41.98**	**33.94**

C.4 Comparison with Baselines

To evaluate the effectiveness of our DUDE, we compare it against the following baselines: *i)* ID Wu et al. (2018), which learns instance-level representations by distinguishing individual instances; *ii)* ProtoNCE Li et al. (2020), which enhances clustering performance by incorporating prototypes to capture semantic structure; *iii)* CDS Kim et al. (2021), which considers both intra-domain and cross-domain discriminative relations among instances to address domain adaptation; *iv)* PCS Yue et al. (2021), which uses K-means to generate prototypes and aligns domains using similarity vectors in a self-supervised manner; *v)* DD Hu & Lee (2022), which introduces a distance-of-distance loss to quantify and minimize domain discrepancy; *vi)* CoDA Wang et al. (2023b), which employs in-domain self-matching supervision and cross-domain classifier alignment to learn domain-invariant and discriminative features; and *vii)* ProtoOT Li et al. (2024a), which unifies intra-domain and cross-domain representation learning through a prototype-based optimal transport framework to improve retrieval performance. For all baselines, we adopt the settings reported in their respective papers. Tables 5–7 report the complete retrieval results on PACS, Office-Home, and DomainNet. Consistent with the findings reported in the main paper, DUDE achieves the best overall performance across the three datasets and different evaluation metrics, demonstrating the effectiveness of the proposed disentangle-then-align framework.

D Ablation Study on Disentanglement Prompt

Recall that in our strategy to extract object features by disentangling them from domain-specific styles, we construct prompts in the format: "a $\{domain\}$ of a $[z]$", where $\{domain\}$ specifies the style in domain (*e.g.*, sketch, photo), and $[z]$ represents the extracted object features.

To investigate whether the disentanglement indeed arises from the design of our varying $\{domain\}$ term, we compare retrieval performance based on $[z]$ learned under two prompt variants: *i)* varying prompts in the form "a $\{domain\}$ of a $[z]$", and *ii)* a fixed prompt "a photo of a $[z]$" applied across all domains. The results for DomainNet are shown in Tab. 8. We observe that varying prompts consistently outperform the fixed prompt, particularly in domains involving highly different styles. For instance, when giving varying $\{domain\}$ terms, retrieval from Clipart to Sketch improves by

Table 6: Comparison of retrieval performance (P@1, P@5, P@15) across various domains on Office-Home. The best results are in **bold**.

Method	Art→Real			Real→Art			Art→Product			Product→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
ID	35.89	33.13	29.60	39.89	34.42	27.65	25.88	24.91	22.49	32.17	25.94	20.23
ProtoNCE	40.50	36.39	34.00	44.53	39.26	32.99	29.54	27.89	25.75	35.73	30.61	24.55
CDS	45.08	41.15	38.73	44.71	40.75	35.53	32.76	31.47	28.90	35.75	32.48	26.82
PCS	41.70	38.51	36.22	44.96	39.88	33.99	33.29	31.50	29.53	39.24	34.77	28.77
DD	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	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	47.38	45.49	43.52	50.61	46.64	41.51	38.11	36.50	35.10	46.47	41.63	34.47
DUDE	**59.41**	**57.59**	**55.45**	**63.27**	**58.44**	**52.20**	**46.39**	**46.55**	**45.98**	**57.55**	**53.93**	**47.31**
	Clipart→Real			Real→Clipart			Product→Real			Real→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
ID	29.48	26.48	23.25	35.51	32.17	27.96	50.73	45.03	39.05	45.12	41.46	38.01
ProtoNCE	25.25	22.66	20.83	41.15	37.66	31.95	53.84	48.25	42.21	47.74	44.85	41.21
CDS	32.51	30.30	27.80	38.88	36.48	33.16	54.00	50.07	45.60	49.39	47.27	43.98
PCS	29.07	26.06	24.00	40.60	38.11	34.06	56.45	50.78	45.37	49.90	47.11	43.73
DD	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	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	36.75	33.58	31.33	48.93	45.93	41.59	64.01	59.22	54.50	54.85	53.49	51.37
DUDE	**43.11**	**41.59**	**41.03**	**57.10**	**55.57**	**52.57**	**74.97**	**72.77**	**68.51**	**68.60**	**67.19**	**64.83**
	Product→Clipart			Clipart→Product			Art→Clipart			Clipart→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
ID	31.52	28.55	24.15	24.01	22.42	20.60	26.78	24.79	21.64	21.17	17.86	14.71
ProtoNCE	36.13	33.99	28.24	21.17	20.63	20.47	28.97	26.15	22.98	21.33	17.40	14.46
CDS	37.69	34.99	30.42	27.24	26.46	24.86	25.59	23.77	22.41	22.41	20.34	17.34
PCS	39.51	37.51	32.81	26.39	25.86	24.92	31.23	28.74	26.11	24.51	21.27	17.54
DD	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	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	44.76	42.64	38.83	29.92	30.15	29.29	35.39	34.57	32.05	28.96	25.58	22.21
DUDE	**51.81**	**51.79**	**48.47**	**34.61**	**34.77**	**34.27**	**40.79**	**40.56**	**38.93**	**30.81**	**29.42**	**26.98**

Table 7: Comparison of retrieval performance (P@50, P@100, P@200) across various domains on DomainNet. The best results are in **bold**.

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
ID	49.46	46.09	40.44	54.38	47.12	37.73	28.27	27.44	26.33	39.98	31.77	24.84
ProtoNCE	46.85	42.67	36.35	54.52	45.04	35.06	28.41	28.53	28.50	57.01	41.84	30.33
CDS	45.84	42.37	37.16	59.13	48.83	37.40	28.51	27.92	27.48	56.69	39.76	26.38
PCS	51.01	46.87	40.19	59.70	50.67	39.38	30.56	30.27	29.68	55.42	42.13	30.76
DD	56.31	52.74	47.38	63.07	57.26	48.17	35.52	35.24	34.35	57.74	46.69	35.47
CoDA	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	70.46	69.41	67.41	82.79	78.68	71.31	40.65	40.35	40.05	77.02	67.33	49.41
DUDE	**93.07**	**92.25**	**90.30**	**95.21**	**94.56**	**91.89**	**55.48**	**54.87**	**53.28**	**78.49**	**70.60**	**57.86**
	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
ID	30.35	29.04	26.55	42.20	34.94	27.52	64.67	54.41	40.07	42.37	39.61	35.56
ProtoNCE	28.24	26.79	24.23	39.83	31.99	24.77	55.44	43.74	32.59	39.13	35.87	32.07
CDS	30.55	29.51	27.00	46.27	36.11	27.33	63.15	47.30	32.93	37.75	35.18	32.76
PCS	30.27	28.36	25.35	42.58	34.09	25.91	63.47	53.21	41.68	48.83	46.21	42.10
DD	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	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	37.16	36.30	34.42	63.59	53.30	38.75	90.21	87.38	77.14	71.13	71.15	70.50
DUDE	**70.28**	**68.48**	**66.40**	**88.20**	**85.24**	**76.01**	**93.23**	**89.99**	**83.10**	**87.01**	**84.59**	**80.36**

9.34% in P@50. This indicates that the image styles could be encapsulated by the varying domain-style token, enabling the model to learn cleaner and more disentangled object features compared to using a fixed prompt across domains.

To further examine the disentanglement behavior of DUDE, we analyze three questions: *i)* whether $[z]$ preserves object semantics, *ii)* whether $[z]$ avoids carrying domain-specific style information, and *iii)* whether the target style is controlled by the domain prompt rather than hidden in $[z]$. Given an image from one domain in DomainNet, we extract its object token $[z]$ using f_θ, keep $[z]$ fixed,

Table 8: Comparison of retrieval performance on the disentangled object features using different prompts or the same prompt on DomainNet.

Prompt	Clipart→Sketch			Sketch→Clipart			Infograph→Real			Real→Infograph			Average		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
varying	75.76	71.86	65.37	82.71	80.34	74.08	36.17	36.40	35.66	75.21	65.17	48.54	67.46	63.44	55.91
fixed	66.42	63.45	58.44	76.70	74.35	68.34	34.87	34.46	34.51	71.36	62.18	46.94	62.33	58.61	52.06
Prompt	Infograph→Sketch			Sketch→Infograph			Painting→Clipart			Clipart→Painting			Average		
	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200	P@50	P@100	P@200
varying	42.78	40.48	37.29	64.52	56.82	43.91	91.01	89.00	82.11	73.27	70.90	65.71	67.89	64.30	57.25
fixed	38.19	36.53	34.13	57.48	49.94	39.47	90.65	87.99	81.42	70.91	69.88	67.89	64.31	61.08	55.73

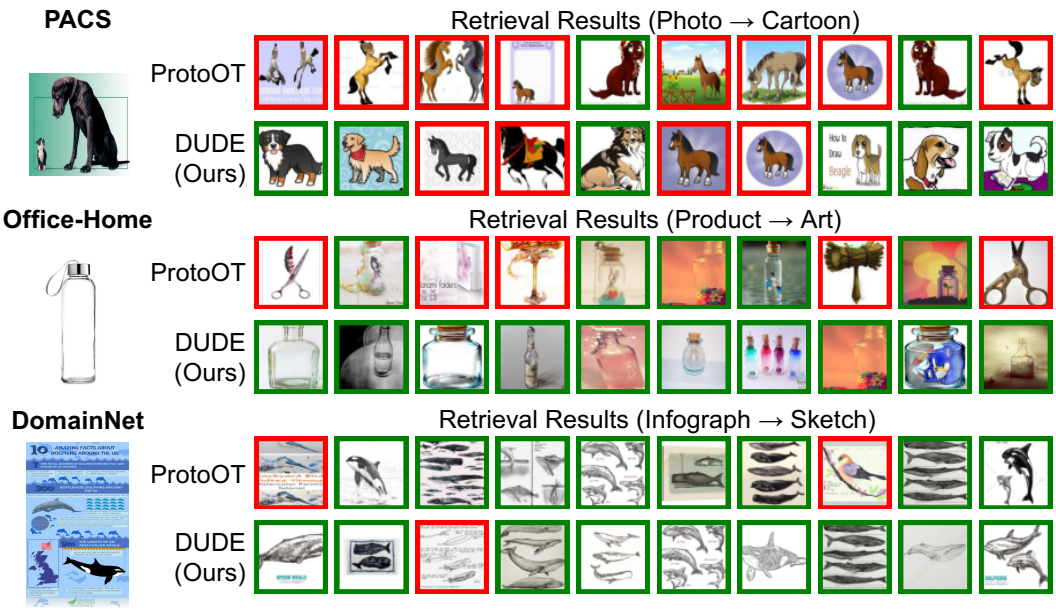


Figure 6: Top 10 unsupervised cross-domain image retrieval results using SOTA method ProtoOT and our DUDE. The retrievals are conducted on examples in three datasets. The red borders indicate falsely retrieved results while the green borders indicate correctly retrieved results.

and combine it with different domain terms, e.g., "photo", "sketch", "infograph", and "clipart". The resulting prompts are fed into the frozen Stable Diffusion model to generate cross-domain images. Since only the domain term changes while $[z]$ remains fixed, the generated images reveal whether object semantics and domain styles are separated. As shown in Fig. 3, the generated images preserve the source object semantics across different target domains, indicating that $[z]$ retains object-discriminative information. Meanwhile, their visual styles follow the target domain prompts rather than the source domain, e.g., sketch-specific strokes do not persist when the tiger token is combined with "photo", "infograph", or "clipart". These results suggest that $[z]$ contains limited source style and that domain-specific style is mainly controlled by the $\{domain\}$ prompt.

E Retrieval Results Visualization

To compare the retrieval performance of our DUDE with the most competitive state-of-the-art method ProtoOT Li et al. (2024a), we present retrieval visualizations on three datasets. We showcase the top-10 cross-domain retrieval results of our DUDE and ProtoOT. As illustrated in Fig. 6, the results are arranged top to bottom by dataset: Photo-Cartoon retrieval in PACS, Product-Art retrieval in Office-Home, and Infograph-Sketch retrieval in DomainNet. These examples demonstrate that DUDE consistently outperforms ProtoOT in most cases. For instance, in the query with the "whale" infograph, ProtoOT fails to retrieve the correct category in the top-1 result, while DUDE retrieves correct instances within both top-1 and top-2. Even in more challenging cases, such as the third query of a "dog" photo in PACS, DUDE retrieves 6 out of 10 correct matches, whereas ProtoOT only retrieves 2 out of 10.