SOURCE RETRIEVAL FOR GENERATIVE STYLIZED IMAGES VIA DESTYLIZED FEATURE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image editing enables users to easily create highly customized and stylized derivatives from reference images, posing new challenges to copyright protection and visual source tracing, particularly in the portrait domain. To address this problem, we formalize the task of *Source Retrieval for Stylized Images*, which aims to trace a stylized query image back to its exact source image. The main challenge is that existing pretrained visual representations can be sensitive to style variations and generation-induced appearance changes, weakening source discrimination under large style shifts. To enable systematic evaluation of this problem, we build MSIR-Bench, comprising 56K images derived from 5,500 source images and covering 25 artistic styles and five generative editing models. Building on this benchmark, we further present Destylized Feature Learning (DFL), a joint learning framework applicable to various visual backbones that encourages source-discriminative representations with reduced reliance on style-predictive cues by jointly employing a Source Discriminative Loss, a Destylized Triplet Loss, and Style Confusion Regularization. Across seven visual backbones, DFL consistently improves source retrieval over zero-shot and task-adapted baselines in the primary evaluation, while maintaining gains under unseen-source and unseen-style settings and achieving positive average improvements across datasets and generators. Our benchmark and associated resources will be made publicly available at `https://anonymous.4open.science/r/Source-Image-Retrieval-8172/`.

1 INTRODUCTION

The rapid evolution of generative AI has substantially reshaped visual content creation. Recent advances in text-guided image editing allow users to apply diverse stylistic transformations to existing images through simple text instructions. While these tools lower the barrier to visual content creation, they also introduce new risks of misuse, including intellectual property infringement, fraudulent advertising, manipulative content, and the generation of non-consensual intimate imagery. Such uses may harm individual rights and copyright owners' commercial interests, further increasing the need for effective visual source tracing.

Generative stylization further complicates visual asset protection and source tracing. Early neural style transfer models typically operated over a limited set of fixed styles and could substantially alter source structures. Later text-to-image (T2I) models primarily generated images from scratch, usually without a clearly defined source image. In contrast, modern instruction-driven image editing models condition on both a reference image and a text prompt, enabling substantial appearance changes while preserving considerable source content, layout, and structure, thereby introducing risks such as copyright infringement, identity misuse, and source concealment. To study this setting, we formalize the task of **Source Retrieval for Stylized Images**. Unlike face recognition, which targets identity-level matching, source retrieval is defined at the image level: the objective is to recover the specific reference image from which a stylized query was generated.

The task requires a representation space that withstands large stylistic shifts while preserving fine-grained source details. Overreliance on style cues hurts generalization to unseen styles, but removing appearance information too aggressively reduces separability among similar source images. The model must learn this trade-off from limited training styles and still generalize to unseen sources,

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.

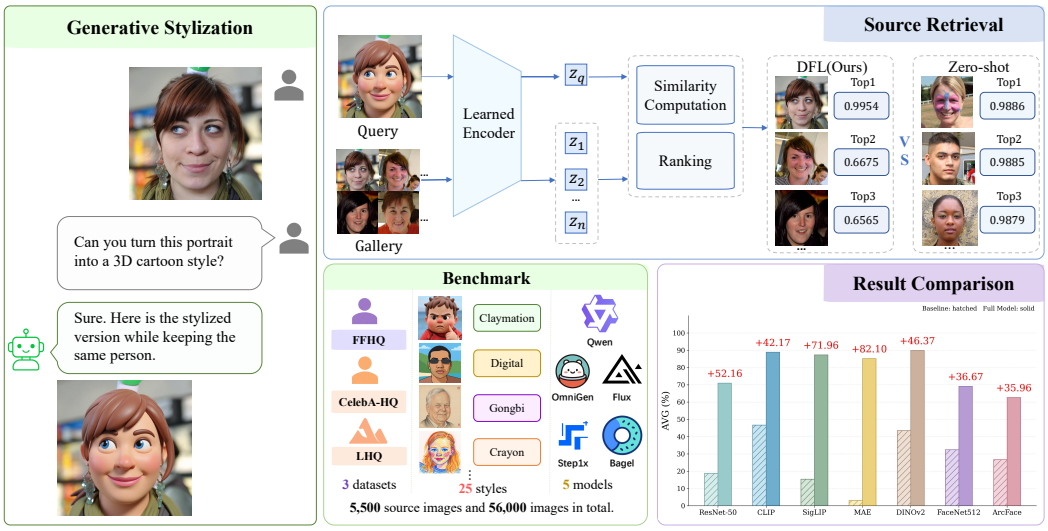

Figure 1: Overview of Source Retrieval for Stylized Images. Given a stylized query, the task aims to retrieve its exact source image from a reference gallery under large generative style shifts. DFL framework improves retrieval performance across different visual backbones.

new text-driven styles, new datasets, and different generators. In deployment, retrieval must remain reliable under post-processing and degradation such as compression, cropping, occlusion, watermarking, and screenshot. While our primary benchmark focuses on portraits, we also evaluate transferability to other portrait datasets and a landscape dataset to test out-of-domain generalization. This work is an initial study of source retrieval for stylized images, aimed at motivating broader research on this task.

To address these challenges, we propose Destylized Feature Learning (DFL), a joint learning framework applicable to different visual backbones. DFL is designed to reduce reliance on style-related cues while preserving fine-grained source discrimination. Its three components play complementary roles: a Source Discriminative Loss preserves separability among different source images; a Destylized Triplet Loss aligns cross-style variants of the same source while separating different sources; and Style Confusion Regularization uses a frozen auxiliary style classifier to reduce style predictability in the learned representation. We additionally employ curriculum-based negative sampling that gradually shifts from random negatives to harder same-style negatives from different sources, facilitating the optimization of the Destylized Triplet Loss.

Our main contributions are:

- We formally define the underexplored task of source retrieval for stylized images, which aims to trace a stylized query back to its exact source image and can support copyright protection and visual source tracing under modern generative image editing.
- We build MSIR-Bench, a comprehensive benchmark with 56K images covering 5,500 source images, 25 artistic styles, and five generative editing models.
- We propose Destylized Feature Learning (DFL), a joint learning framework that employs a Source Discriminative Loss, a Destylized Triplet Loss, and Style Confusion Regularization to reduce reliance on style-related cues while preserving fine-grained source discrimination.
- Experiments on seven visual backbones (five general foundation models and two face-recognition models) show that DFL consistently outperforms zero-shot and task-adapted baselines in the primary evaluation, improves under unseen-source and unseen-style settings, yields positive average gains across generators and datasets, and remains robust to common post-processing.

2 RELATED WORK

Generative image editing. Instruction-driven image editing has advanced quickly. Models such as InstructPix2Pix Brooks et al. (2023), IP-Adapter Ye et al. (2023), InstantStyle Wang et al. (2024), and FLUX.2 Klein Black Forest Labs (2026) can apply substantial style and appearance transformations while retaining much of the original content. Yet this form of generative stylization creates new difficulties for copyright enforcement and visual source tracing, because the resulting visual gap can make conventional low-level similaritybased matching methods unreliable Swaminathan et al. (2006).

Image matching and copy detection. Image copy detection and perceptual similarity approaches offer related ways to match images under visual transformations. SSCD Pizzi et al. (2022) learns descriptors that remain stable under typical copy edits, while LPIPS Zhang et al. (2018) and DreamSim Fu et al. (2023) quantify perceptual similarity across varying semantic scales. Nevertheless, these methods are largely designed for standard image transformations or perceptual resemblance, whereas generative restylization may dramatically change texture and overall appearance while still retaining content linked to the original source, thereby making precise source discrimination more difficult.

Generated image forensics. Current research in generated image forensics mainly targets identifying AI generated content, attributing images to particular generators, or flagging possible copyright violations. Approaches such as FSD Wu et al. (2025c), C2P-CLIP Tan et al. (2025), SAFE Li et al. (2025), DDA Chen et al. (2025), and CO-SPY Cheng et al. (2025) mainly leverage generation-specific artifacts, whereas PRISM Ricco et al. (2025) and OmniDFA Wu et al. (2025b) aim to determine which model produced the image. Likewise, copyright-focused methods like MCID Huang et al. (2025) and StyleSentinel Chen et al. (2026) emphasize infringement detection instead of pinpointing the exact source image. While related, these problems differ from our scenario, which demands fine-grained retrieval of the exact source image despite substantial generative style changes.

Visual representations for source retrieval. Pretrained visual encoders, including ResNet-50 He et al. (2016), MAE He et al. (2022), CLIP Radford et al. (2021), SigLIP Zhai et al. (2023), and DINOv2 Oquab et al. (2024), offer powerful representations for downstream retrieval tasks, whereas ArcFace Deng et al. (2019) and FaceNet512 Schroff et al. (2015) are tailored to face recognition. Yet, the substantial appearance changes caused by generative restylization can make naive reuse of these pretrained features for exact-source retrieval less effective, reducing discrimination at the source level and highlighting the need for task-specific adaptation.

3 BENCHMARK: MSIR-BENCH

We present MSIR-Bench, a benchmark tailored to Source Retrieval for Stylized Images. It contains 56K images in total, including 5,500 source images and 50,500 stylized images, covering 25 artistic styles and five generative editing models. In addition to offering a large-scale set of sourcestylized image pairs, MSIR-Bench specifies a standardized exact-source retrieval protocol and factorized evaluation configurations along four dimensions: source, style, generator, and dataset domain. Instead of bundling all distribution shifts into a single test split, these configurations independently assess generalization to unseen sources and styles, transfer across generators, and transfer across datasets, enabling controlled analysis of different causes of performance drops.

The training set includes 500 source images, each paired with 19 stylized variants. This extensive style coverage per source is intended to present identical underlying content with a wide range of visual transformations, supplying supervision to learn source-discriminative representations that stay stable across styles instead of depending on style-specific cues.

The evaluation set is split into a primary retrieval benchmark and a controlled generalization suite. The primary benchmark is built on 3,000 previously unseen FFHQ source images and assesses queries generated by Qwen-Image-Edit-2509 in both seen-style and unseen-style conditions. In the unseen-style condition, two styles are randomly chosen from a separate collection of six artistic styles that were not included in training. Because the source gallery remains unseen in both conditions and the only change is whether the query style was encountered during training, this setup

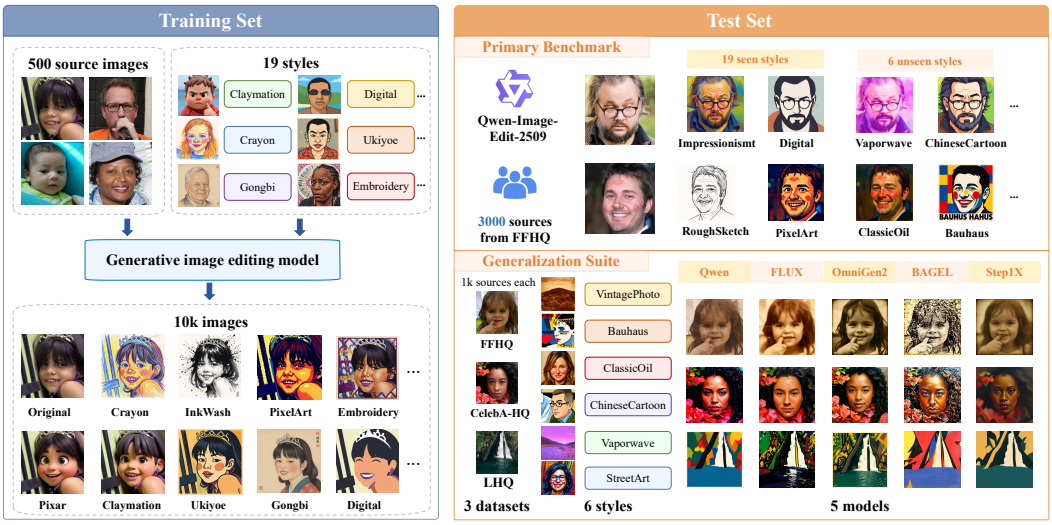

Figure 2: MSIR-Bench overview. The primary benchmark tests retrieval for both seen and unseen styles on 3,000 unseen FFHQ source images using Qwen-Image-Edit-2509. The generalization suite spans three datasets and evaluates five generative editing models: Qwen-Image-Edit-2509, FLUX.2 [klein] 4B Base, Step1X-Edit v1.2, OmniGen2, and BAGEL-7B-MoT.

disentangles generalization to new sources from the added impact of a style shift. Using a 3,000-image source gallery also makes exact-source retrieval notably challenging.

We also build a balanced generalization suite spanning three datasets and five generative editing models. For each dataset, we sample 1,000 source images unseen during training and use the same fixed set of sources for all five generators. Each source is then stylized two times by randomly selecting two of the six unseen styles, yielding 30,000 query images overall. This balanced design controls for source-content variation across generators and allows for straightforward comparisons across datasets and models under identical evaluation settings.

4 Method: Destylized Feature Learning Framework

Given the task and benchmark introduced above, we propose **Destylized Feature Learning** (DFL), a joint learning framework applicable to different visual backbones for source retrieval under generative stylization. DFL aims to reduce reliance on style-related cues while preserving the fine-grained information required for exact-source discrimination. The overall objective is primarily composed of three complementary terms: the Source Discriminative Loss enforces separability among sources; the Destylized Triplet Loss models feature space structure across different stylized variants of the same source image; and Style Confusion Regularization uses a frozen auxiliary style classifier to limit the adverse effect of style on source retrieval. In combination, these components balance cross-style invariance with source-level discrimination.

Figure 3 shows the proposed two-stage DFL pipeline and the test-time source retrieval procedure. Training has two phases. Phase I freezes the vision backbone and trains an auxiliary style classifier to extract style cues from pretrained features. Phase II freezes the classifier and optimizes the retrieval encoder with the full DFL objective. We use curriculum-based negative sampling, gradually shifting from random negatives to harder same-style negatives from different sources, so later training focuses on separating sources with similar styles.

4.1 Problem Formulation

Let the training set be

$$\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N}, \tag{1}$$

where $x_i \in \mathcal{X}$ is an input image, $y_i \in \mathcal{Y}$ its source-image label, and $s_i \in \mathcal{S}$ its style label. Here, $\mathcal{X}$ is the image domain, $\mathcal{Y}$ the set of source-image labels, and $\mathcal{S}$ the set of K training styles used by the

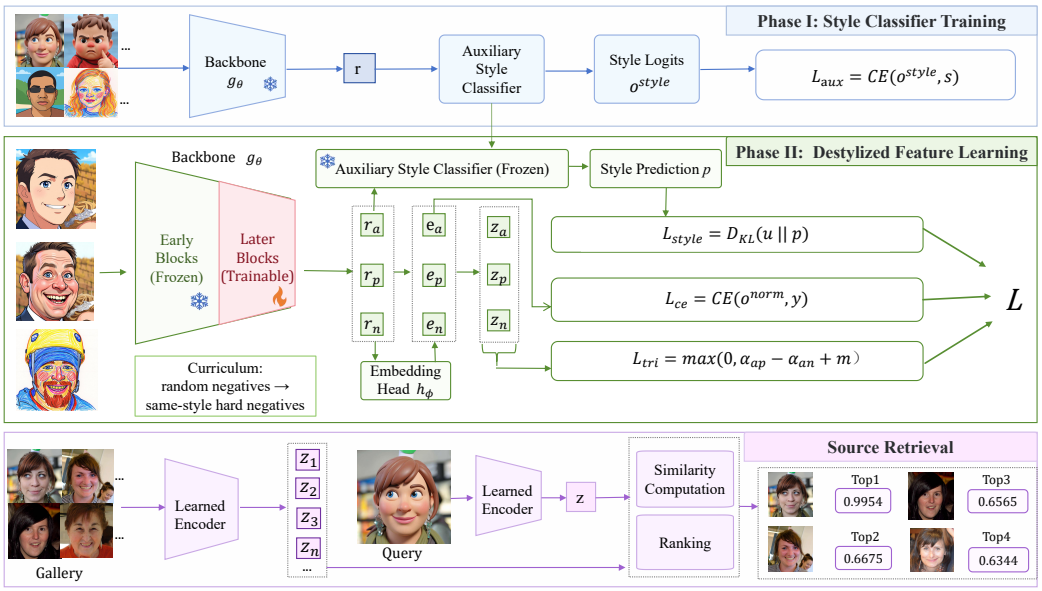

Figure 3: Overview of the Destylized Feature Learning framework. Phase I trains an auxiliary style classifier while freezing the backbone. Phase II freezes the classifier and optimizes the retrieval encoder with Source Discriminative Loss, Destylized Triplet Loss, Style Confusion Regularization. At test time, the encoder retrieves source images by ranking cosine similarity.

auxiliary style classifier. Our goal is to learn an embedding function

$$f_\theta : \mathcal{X} \to \mathbb{R}^d \tag{2}$$

that maps images to a d-dimensional space so that images from the same source are close regardless of style, while images from different sources are well separated.

4.2 Framework and Training Pipeline

The proposed framework is compatible with a variety of visual backbones. For an input image x, a backbone encoder g_θ extracts a visual representation $\mathbf{r}$. Next, a learnable embedding head h_ψ maps $\mathbf{r}$ into a retrieval embedding $\mathbf{e}$, which is then ℓ_2-normalized to yield the final retrieval feature $\mathbf{z}$:

$$\mathbf{r} = g_\theta(x), \quad \mathbf{e} = h_\psi(\mathbf{r}), \quad \mathbf{z} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}. \tag{3}$$

The backbone g_θ may be implemented using different visual encoders, including CLIP, SigLIP, and DINOv2 Radford et al. (2021); Zhai et al. (2023); Oquab et al. (2024).

Our goal is to learn an embedding $\mathbf{e}$ that effectively distinguishes sources while being less influenced by style-specific signals. To achieve this, we fine-tune only the backbones last layers and keep the earlier layers frozen. We enforce source-discriminative learning by attaching a linear classifier to $\mathbf{e}$ to predict the source. This classifier is used only during training and removed at inference time, whereas the embedding head serves as a learnable projection for retrieval. We further encourage reduced dependence on style cues through a Destylized Triplet Loss and Style Confusion Regularization.

4.3 Destylized Feature Learning Objective

The destylized feature learning objective jointly optimizes three complementary terms: a source discriminative loss for source-image separability, a destylized triplet loss for cross-style compactness and inter-source separation, and style confusion regularization to reduce reliance on style-related cues. These objectives jointly shape the retrieval space so that style dependence is reduced without sacrificing source-image discrimination.

Source Discriminative Loss. To encourage separability among source images, we use the normalized softmax cross-entropy loss (NormCE). Let $\mathbf{w}_c$ denote the classifier weight vector for source class c. We normalize the weights and compute cosine logits scaled by τ:

$$\hat{\mathbf{w}}_c = \frac{\mathbf{w}_c}{\|\mathbf{w}_c\|_2}, \qquad o_c^{\text{norm}} = \tau \cdot \hat{\mathbf{w}}_c^\top \mathbf{z}. \tag{4}$$

Let $\mathbf{o}^{\text{norm}}$ be the logit vector over all source classes. The source discriminative loss is then formulated as:

$$\mathcal{L}_{\text{ce}} = \text{CE}(\mathbf{o}^{\text{norm}}, y) = -\log \frac{\exp(o_y^{\text{norm}})}{\sum_j \exp(o_j^{\text{norm}})}, \tag{5}$$

where y is the ground-truth source class index. This formulation applies source-image supervision in angular space, aligning with the cosine-similarity retrieval metric Wang et al. (2017); Ranjan et al. (2017). Because the source classifier is discarded at inference, $\mathcal{L}_{\text{ce}}$ only optimizes the retrieval representation and its angular embedding distribution.

Destylized Triplet Loss. While the Source Discriminative Loss enforces global source separability, it does not capture feature structure across style variations. We therefore build triplets based on sourcestyle relationships and call the objective the Destylized Triplet Loss. The positive shares the anchors source image but uses a different style, while the negative comes from a different source. This encourages style-invariant source consistency while separating different sources. The loss is computed in angular feature space, matching the cosine similarity used at test time. Let the ℓ_2-normalized embeddings be $\mathbf{z}_a$, $\mathbf{z}_p$, and $\mathbf{z}_n$. The Destylized Triplet Loss is defined as:

$$\mathcal{L}_{\text{tri}} = \max\left(0,\ \alpha_{ap} - \alpha_{an} + m\right), \tag{6}$$

where $\alpha_{ap} = \arccos(\text{clip}(\mathbf{z}_a^\top \mathbf{z}_p, -1+\epsilon, 1-\epsilon))$ and $\alpha_{an} = \arccos(\text{clip}(\mathbf{z}_a^\top \mathbf{z}_n, -1+\epsilon, 1-\epsilon))$ are embedding angles, m is the angular margin, and $\epsilon = 10^{-6}$ improves numerical stability.

Combined with curriculum-based negative sampling, training shifts from random negatives to harder same-style negatives from other sources, emphasizing source discrimination when styles are similar.

Style Confusion Regularization. To reduce reliance on style cues, we use a frozen auxiliary style classifier as a fixed reference. In Phase I, the backbone is frozen while the auxiliary style classifier is trained on $\mathbf{r}$.

$$\mathcal{L}_{\text{aux}} = \text{CE}(\mathbf{o}^{\text{style}}, s), \tag{7}$$

where s is the training style label. This loss is used only to train the auxiliary style classifier and is excluded from the Phase II retrieval objective.

In Phase II, the auxiliary style classifier is frozen and no longer updated with the retrieval encoder, serving as a fixed reference to measure style predictability in the current feature $\mathbf{r}$. Let its prediction be

$$\mathbf{p} = \text{softmax}(W_{\text{style}}\mathbf{r} + b_{\text{style}}). \tag{8}$$

We define a uniform target distribution $\mathbf{u}$ over K training style labels with $u_k = 1/K$, and define

$$\mathcal{L}_{\text{style}} = D_{\text{KL}}(\mathbf{u} \,\|\, \mathbf{p}). \tag{9}$$

The auxiliary classifier is trained on pretrained features to predict style, then frozen during retrieval training. Minimizing $\mathcal{L}_{\text{style}}$ makes the current feature $\mathbf{r}$ less style-predictive under this fixed reference. Combined with source discrimination and Destylized Triplet Loss, it suppresses style cues while preserving source discrimination.

Overall loss. The Phase II objective is

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}}. \tag{10}$$

λ_{ce} and λ_{tri} are fixed weights for source discrimination and the destylized triplet loss. λ_{style} sets the style regularization strength. We apply style regularization only in Phase II and warm it up so early training focuses on source discrimination and cross-style alignment.

4.4 SOURCE RETRIEVAL

At retrieval time, we remove both the source classifier and the auxiliary style classifier, keeping only the backbone encoder together with the embedding head. These components constitute the retrieval encoder $f_{\theta,\psi}$. For a query image x_q and a source gallery $\mathcal{R} = \{x_r\}_{r=1}^{M}$, we rank the gallery samples by their cosine similarity to x_q and return the closest source image:

$$\hat{x}_q = \arg\max_{x_r \in \mathcal{R}} \frac{f_{\theta,\psi}(x_q)^\top f_{\theta,\psi}(x_r)}{\|f_{\theta,\psi}(x_q)\|_2 \|f_{\theta,\psi}(x_r)\|_2}. \tag{11}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Evaluation protocol. We train and evaluate our framework on MSIR-Bench (see Section 3). The training set contains 500 source images and 9,500 stylized variants (19 per source), yielding 10,000 training images in total.

For the primary evaluation, we use a gallery of 3,000 unseen source images and construct two query subsets: "Seen Style" and "Unseen Style". In "Seen Style", each source image is stylized using two styles randomly selected from the 19 styles used during training, resulting in 6,000 queries. In "Unseen Style", each source image is stylized using two styles randomly sampled from a separate pool of six unseen styles, again yielding 6,000 queries. We adopt the gallery-query retrieval protocol described in Section 4.4: for each stylized query, we rank the 3,000 source images using cosine similarity in the learned embedding space. Unless noted otherwise, we report Top-1 retrieval accuracy.

Models and variants. We evaluate seven representative visual backbones: five general-purpose models including ResNet-50 He et al. (2016), CLIP ViT-B/32 Radford et al. (2021), SigLIP ViT-B/16 Zhai et al. (2023), MAE ViT-B/16 He et al. (2022), and DINOv2 ViT-B/14 Oquab et al. (2024), and two widely used face recognition models including FaceNet512 Schroff et al. (2015) and ArcFace Deng et al. (2019).

For each backbone, we evaluate three settings: **Zero-shot**, which directly uses the pretrained representation; **Source Disc.**, which fine-tunes the backbone with the Source Discriminative Loss; and **DFL**, which uses the full proposed objective with curriculum-based negative sampling. All task-adapted variants use the same training data and comparable optimization budgets.

Metric. We use Top-1 accuracy as the primary metric and also report Top-5 accuracy and Source Accuracy (Source Acc.) in selected experiments. Source Acc. measures retrieval consistency across stylized queries from the same source image: a source image is correct if more than half of its stylized queries retrieve the correct source at Top-1. Source Acc. is the fraction of source images meeting this criterion.

5.2 EFFECTIVENESS EVALUATION AND COMPARISON

Table 1 presents a comparison of Zero-shot, Source Disc., and the full DFL framework on seven visual backbones, covering five general-purpose vision models and two face-recognition-specific models, ArcFace Deng et al. (2019) and FaceNet512 Schroff et al. (2015). This side-by-side evaluation assesses both how well various pretrained representations transfer directly to source retrieval under generative stylization and how effective DFL is across different categories of visual backbones.

Table 1 shows that zero-shot retrieval performs poorly for most backbones, underscoring the importance of task-specific adaptation when generative stylization is involved. Relative to Source Disc., the full DFL framework yields consistent gains in Top-1 accuracy for both Seen Style and Unseen Style across all seven backbones, spanning general-purpose vision encoders and dedicated face recognition models. Notably, DINOv2 ViT-B/14 delivers the best overall results and is thus selected as the default backbone in the remaining experiments.

We also compare DFL with task-adjacent image-matching methods under the same unseen-style retrieval protocol (Appendix C.3).

Table 1: Performance comparison across seven visual backbones and three training settings. "Seen" and "Unseen" evaluate unseen source images under seen and unseen styles, respectively.

Setting	Metric	ResNet-50	CLIP ViT-B/32	SigLIP ViT-B/16	MAE ViT-B/16	DINOv2 ViT-B/14	FaceNet512	ArcFace
Zero-shot	Seen	12.07	43.22	12.68	4.65	37.25	26.08	17.97
	Unseen	25.62	50.13	18.05	1.48	49.83	38.83	35.48
	AVG	18.84	46.67	15.37	3.07	43.54	32.45	26.72
Source Disc.	Seen	72.57	88.12	84.07	89.30	93.22	71.42	61.75
	Unseen	61.75	75.60	73.22	75.47	80.23	61.62	59.45
	AVG	67.16	81.86	78.64	82.38	86.72	66.52	60.60
DFL	Seen	**77.93**	**95.37**	**94.33**	**91.32**	**95.72**	**74.70**	**64.90**
	Unseen	**64.08**	**82.30**	**80.33**	**79.03**	**84.10**	**63.53**	**60.47**
	AVG	**71.00**	**88.84**	**87.33**	**85.17**	**89.91**	**69.12**	**62.68**

Table 2: Ablation of the DFL objectives on DINOv2 ViT-B/14 and CLIP ViT-B/32.

Configuration	DINOv2			CLIP		
	Seen ↑	Unseen ↑	Avg. ↑	Seen ↑	Unseen ↑	Avg. ↑
Source Disc. only	93.22	80.23	86.72	88.12	75.60	81.86
Des. Triplet only	93.10	80.45	86.78	84.00	72.83	78.42
Source Disc. + Sty. Con.	93.00	83.62	88.31	90.82	77.10	83.96
Source Disc. + Des. Triplet	95.37	82.92	89.14	94.33	80.90	87.62
Full DFL	**95.72**	**84.10**	**89.91**	**95.37**	**82.30**	**88.84**

5.3 Ablation Study

To assess how each DFL objective contributes, we perform objective-level ablations on both DINOv2 ViT-B/14 and CLIP ViT-B/32. Because the main results already show DFLs effectiveness across seven backbones, we concentrate here on isolating the impact of the individual training objectives and checking whether these effects are consistent across different pretrained representations.

Table 2 indicates that Source Discriminative Loss and Destylized Triplet Loss each offer strong supervision when applied on their own. Incorporating Style Confusion Regularization into Source Disc. primarily boosts retrieval for unseen styles, whereas pairing Source Disc. with Destylized Triplet Loss yields gains on both seen and unseen styles. We observe the same patterns for both DINOv2 and CLIP. Using the complete DFL objective delivers the best overall results across both backbones, suggesting that source discrimination, cross-style alignment, and style confusion regularization contribute complementary advantages.

To confirm that the gains do not simply stem from adding more metric-learning supervision, we evaluate DFL against standard Triplet Loss and SupCon under comparable training conditions. Neither general-purpose objective matches DFLs performance, underscoring the task-specific nature of our framework. Additional results are reported in Appendix C.2.

5.4 Generalization Evaluation

We additionally assess DFL on three datasets and five generators under the challenging setting of unseen source images and unseen styles, including four generators that were not used in training. As reported in Table 3, DFL delivers steady improvements for most dataset–generator pairs, raising the overall average performance compared with Source Disc., and yielding positive average gains across both generators and dataset domains.

5.5 Robustness Evaluation

Table 4 reports our evaluation of DFL under typical image degradations and post-processing operations. DFL shows consistently strong retrieval performance across these perturbations, with Top-1 accuracy staying above 70% in all tested settings. Results are especially steady under JPEG compression and remain fairly robust to watermarking and screenshots.

Table 3: Cross-dataset and cross-generator generalization of DFL. Each cell shows Top-1 retrieval accuracy on unseen source images with unseen styles. "Average" columns average across datasets per generator; the bottom row averages across generators per dataset. "Gain" is DFLs average improvement over Source Disc. across datasets.

Generator	CelebA-HQ		FFHQ		LHQ		Average		Gain
	Src.	DFL	Src.	DFL	Src.	DFL	Src.	DFL	
BAGEL	69.95	**76.70**	71.85	**78.45**	60.15	**62.95**	67.32	**72.70**	**+5.38**
FLUX.2	74.85	**82.65**	79.85	**83.85**	81.95	**86.10**	78.88	**84.20**	**+5.32**
OmniGen2	**81.90**	79.90	79.40	**81.00**	63.05	**69.35**	74.78	**76.75**	**+1.97**
Qwen	79.10	**82.90**	82.20	**83.90**	84.40	**88.15**	81.90	**84.98**	**+3.08**
Step1X	61.45	**73.20**	73.60	**82.90**	65.45	**67.30**	66.83	**74.47**	**+7.64**
Average	73.45	**79.07**	77.38	**82.02**	71.00	**74.77**	**73.94**	**78.62**	**+4.68**

Table 4: Robustness of DFL under common image degradations and post-processing operations, evaluated on the unseen-style test set.

Condition	Setting	Top-1 ↑	Top-5 ↑	Source Acc. ↑
Clean	None	**84.10**	**89.23**	**71.57**
JPEG	Q=70	84.07	89.23	71.50
JPEG	Q=30	84.08	89.20	71.57
Crop	70% retained	83.13	88.58	69.67
Crop	50% retained	77.72	85.75	61.83
Occlusion	Random 25%	72.70	80.93	55.37
Mixed corruption	Random combination	81.20	87.20	67.13
Watermark	Height 14%, opacity 0.55	83.07	88.63	69.83
Watermark	Height 22%, opacity 0.70	80.80	87.02	66.13
Screenshot	Mobile-post screenshot	79.18	86.28	64.83
Screenshot + JPEG	Q=65	79.23	86.20	64.90

6 Conclusion

We introduced **Source Retrieval for Stylized Images**, a task that aims to retrieve the exact source image of a generatively stylized query under large appearance and style variations. To enable systematic evaluation of this problem, we constructed **MSIR-Bench**, a benchmark containing 56K images derived from 5,500 source images, covering 25 artistic styles and five generative editing models, with evaluation settings designed to examine generalization across sources, styles, generators, and dataset domains. Building on this benchmark, we proposed **Destylized Feature Learning (DFL)**, which reduces reliance on style-related cues while preserving fine-grained source discrimination. Extensive experiments across multiple visual backbones show that DFL consistently improves exact-source retrieval over zero-shot and task-adapted baselines, while maintaining strong generalization and robustness under diverse evaluation conditions. We hope that the task, benchmark, and baseline framework established in this work can provide a foundation for future research on source tracing under generative image editing.

AI USE STATEMENT

Generative AI tools were used in this work for two purposes. First, generative image-editing models were used to create the stylized images in MSIR-Bench, as detailed in the dataset creation process and the experimental configuration. Second, LLM-based tools supported writing and editorial refinement of the manuscript. All AI-assisted text was examined and validated by the authors. The authors take full responsibility for the final content of the paper and for the artifacts generated with the assistance of generative AI.

REFERENCES

Black Forest Labs. FLUX.2 [klein]. https://bfl.ai/models/flux-2-klein, 2026. Accessed: 2026-04-30.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Lingxiao Chen, Liqin Wang, and Wei Lu. StyleSentinel: Reliable artistic copyright verification via stylistic fingerprints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 35348–35356, 2026.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, 2025.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13455–13465, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Google. Nano Banana 2. https://gemini.google/overview/image-generation/, 2026. Accessed: 2026-05-07.

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 (CVPR)*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Chuanwei Huang, Zexi Jia, Hongyan Fei, Yeshuang Zhu, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Xiaoyue Duan, Jinchao Zhang, and Jie Zhou. Mcid: Multi-aspect copyright infringement detection for generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16154–16164, October 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

OpenAI. GPT Image 2. `https://developers.openai.com/api/docs/models/gpt-image-2`, 2026. Accessed: 2026-05-07.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

Qwen Team. Qwen-Image-Edit-2509. `https://huggingface.co/Qwen/Qwen-Image-Edit-2509`, 2025. Accessed: 2026-05-01.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Rajeev Ranjan, Carlos D. Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

Emanuele Ricco, Elia Onofri, Lorenzo Cima, Stefano Cresci, and Roberto Di Pietro. PRISM: Phase-enhanced radial-based image signature mapping framework for fingerprinting ai-generated images. *arXiv preprint arXiv:2509.15270*, 2025.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14144–14153, 2021.

Ashwin Swaminathan, Yinian Mao, and Min Wu. Robust and secure image hashing. *IEEE Transactions on Information Forensics and Security*, 1(2):215–230, 2006.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. NormFace: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1041–1049, 2017.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025a.

Shiyu Wu, Shuyan Li, Jing Li, Jing Liu, and Yequan Wang. OmniDFA: A unified framework for open set synthesis image detection and few-shot attribution. *arXiv preprint arXiv:2509.25682*, 2025b.

Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025c.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A MSIR-BENCH DATASET DETAILS

A.1 DATASET OVERVIEW

MSIR-Bench is constructed for **Source Retrieval for Stylized Images**: given a generatively stylized query, the goal is to retrieve its exact source image from a gallery of source images.

MSIR-Bench includes 56,000 images from 5,500 sources across 25 styles and 5 generative editing models. It provides an FFHQ-based primary benchmark and a generalization suite over 3 datasets and 5 generators to test generalizability to changes in sources, styles, generators, and domains.

A.2 SOURCE IMAGE COLLECTION

We build the training set and primary benchmark using FFHQ Karras et al. (2019). From this dataset, we choose 500 source images for training and a separate, non-overlapping set of 3,000 source images to form the primary evaluation gallery.

To test cross-dataset generalization, we additionally consider CelebA-HQ Karras et al. (2018) and LHQ Skorokhodov et al. (2021). The generalization suite includes 1,000 source images per dataset. The 1,000 FFHQ sources are drawn from the 3,000-image primary evaluation gallery, so they do not add new unique sources, whereas CelebA-HQ and LHQ each add 1,000 new source images. Altogether, this yields 5,500 unique source images.

In these datasets, different source images represent different identities for portraits and different scenes for landscapes. FFHQ and CelebA-HQ are used for portrait-domain evaluation, while LHQ measures transfer to the landscape domain. We do not use demographic attributes for supervision; both training and evaluation use only source-image and style labels.

A.3 STYLIZED IMAGE GENERATION

The training set and primary benchmark are generated using Qwen-Image-Edit-2509 Qwen Team (2025). For training, each source image is edited using all 19 seen styles, producing 9,500 stylized images from 500 source images.

For the primary benchmark, we use 3,000 FFHQ source images not seen during training as the retrieval gallery. From each source image, we create two query types: in the *Seen Style* setting, two styles are randomly selected from the 19 training styles; in the *Unseen Style* setting, two styles are randomly selected from a separate set of six unseen styles. This yields 6,000 query images per setting. In addition, we create 1,000 *Seen Source + Seen Style* queries as an auxiliary reference split included in MSIR-Bench, where both the source images and the styles were observed during training.

To assess cross-generator generalization, we also consider five generative editing models: Qwen-Image-Edit-2509, FLUX.2 [klein] 4B Base Black Forest Labs (2026), Step1X-Edit v1.2 Liu et al. (2025), OmniGen2 Wu et al. (2025a), and BAGEL-7B-MoT Deng et al. (2025). For each of the three generalization datasets, we keep the same 1,000 source images across all five generators. For every source image and generator, two styles are randomly sampled from the six unseen styles to generate the query images, ensuring a consistent protocol for evaluation across datasets and generators.

A.4 STYLE TAXONOMY

MSIR-Bench includes 25 artistic styles: 19 seen styles used for training and six unseen styles held out for evaluation. The training set spans a wide range of visual transformations, including line-based rendering, ink painting, material changes, cartoonization, digital rendering, and geometric abstraction, exposing the model to a broad range of appearance diversity during training.

The 19 seen styles are: *Impressionist*, *LineArt*, *InkWash*, *Gongbi*, *Ukiyoe*, *PencilSketch*, *RoughSketch*, *Crayon*, *PopArt*, *Claymation*, *Embroidery*, *Pixar*, *Digital*, *LowPoly*, *PixelArt*, *Glitch*, *Anime*, *Comic*, and *Caricature*.

Table 5: Unique-image composition of MSIR-Bench. Images reused across protocols are counted once.

Component	# Source Images	# Stylized Images	# Unique Images
Training set	500	9,500	10,000
Primary FFHQ gallery	3,000	–	3,000
Primary: Seen Style	–	6,000	6,000
Primary: Unseen Style	–	6,000	6,000
Seen Source + Seen Style	–	1,000	1,000
Additional generalization queries	–	28,000	28,000
CelebA-HQ gallery	1,000	–	1,000
LHQ gallery	1,000	–	1,000
Total	**5,500**	**50,500**	**56,000**

The 6 unseen styles are: *Bauhaus*, *ChineseCartoon*, *ClassicOil*, *StreetArt*, *Vaporwave*, and *VintagePhoto*. These styles are excluded from training and are reserved for the primary Unseen Style evaluation as well as the generalization suite.

A.5 Dataset Statistics

Table 5 summarizes MSIR-Benchs unique-image composition. The training set includes 500 source images and 9,500 stylized images (10,000 total). The primary benchmark adds a 3,000-image FFHQ gallery, plus 6,000 Seen Style and 6,000 Unseen Style queries; we also keep 1,000 Seen Source + Seen Style queries as a reference.

The generalization suite provides 30,000 queries across three datasets and five generators. Since 2,000 FFHQ–Qwen queries overlap with the primary Unseen Style set, they are counted once, yielding 28,000 additional unique queries. CelebA-HQ and LHQ each add 1,000 source images, while the 1,000 FFHQ generalization sources are already in the primary gallery. In total, MSIR-Bench contains 5,500 unique source images and 50,500 unique stylized images (training stylizations + evaluation queries), for 56,000 unique images.

B Implementation and Training Details

B.1 Training Procedure

DFL trains in two phases. In Phase I, the visual backbone is frozen and an auxiliary style classifier is trained on its features to classify the 19 seen styles with an additional original-image category. In Phase II, the style classifier is frozen and used for Style Confusion Regularization, while the retrieval encoder is fine-tuned with Source Discriminative Loss, Destylized Triplet Loss, and Style Confusion Regularization. At inference, retrieval uses only the image encoder and embedding, not the auxiliary style or source classifiers.

Destylized Triplet Loss uses curriculum-based negative sampling: training starts with random negatives from different source images, then shifts to harder negatives, preferring same-style samples from different sources when available. Style Confusion Regularization is warmed up during Phase II to avoid disrupting early source-discriminative features.

B.2 Optimization Settings

Unless otherwise specified, the retrieval model in Phase II is trained for 20 epochs with a batch size of 16. Input images are resized to 224×224 and normalized using the pretrained statistics of the corresponding backbone. We apply standard data augmentations, including random resized crop, horizontal flip, color jitter, random grayscale, Gaussian blur, and random erasing. Mixed-precision training is used, and gradients are clipped with a maximum norm of 5.0. The retrieval embedding dimension is set to 512, followed by ℓ_2 normalization before retrieval.

In Phase I, the auxiliary style classifier is implemented as a single linear layer and trained for 3 epochs on frozen backbone features using AdamW with a learning rate of 5×10^{-4} and weight

Table 6: Random-seed stability of DFL with the DINOv2 ViT-B/14 backbone. Results report Top-1 retrieval accuracy (%) over five random seeds.

Seed	Seen Style Top-1 ↑	Unseen Style Top-1 ↑
7	95.83	83.40
42	95.72	84.10
123	96.03	83.62
2025	95.95	83.77
3407	96.05	83.58
Mean ± Std.	95.92 ± 0.12	83.69 ± 0.23

decay of 10^{-4}. Phase II uses the AdamW optimizer with weight decay 10^{-4} and parameter-group-specific learning rates of 1×10^{-5} for the trainable backbone layers and 1×10^{-4} for the embedding head and source classifier. A MultiStepLR scheduler multiplies the learning rates by 0.1 at epochs 12 and 16. Phase II is trained for 20 epochs and adopts curriculum-based negative sampling: random negatives are used during the first 10 epochs, followed by harder negatives during the last 10 epochs.

For the Source Discriminative Loss, we use label smoothing of 0.05 and a cosine-logit scale of $\tau = 30$. The angular margin in the Destylized Triplet Loss is set to $m = 0.2$ radians. The loss weights are set to $\lambda_{\text{ce}} = 0.3$ and $\lambda_{\text{tri}} = 4.0$, while the coefficient λ_{style} for Style Confusion Regularization is gradually warmed up to 0.05 during Phase II.

B.3 Fine-Tuning Scope

To adapt pretrained representations to source retrieval while preserving their general visual features, we apply partial fine-tuning to all backbones. Specifically, only the last four backbone blocks, the 512-dimensional embedding head, and the source classifier used during training are updated, while the remaining backbone layers stay frozen. For Vision Transformers, a block refers to a transformer block; for CNN-based backbones, it refers to the corresponding final backbone blocks in the implementation.

After training, both the source classifier and the auxiliary style classifier are discarded, and retrieval uses only the visual encoder and embedding head. All images are mapped to 512-dimensional ℓ_2-normalized embeddings, and source images in the gallery are ranked by cosine similarity at test time.

C Supplementary Experiments

C.1 Random-Seed Stability

To evaluate the stability of DFL with respect to random initialization, we repeat the default DINOv2 ViT-B/14 experiment using five random seeds. As shown in Table 6, DFL exhibits small performance variations across different random seeds under both seen- and unseen-style settings.

C.2 Comparison with Generic Metric-Learning Objectives

To further examine the role of the Destylized Triplet Loss, we compare DFL with generic metric-learning objectives, including Standard Triplet and SupCon, under matched training settings. We additionally compare Source Disc. + Standard Triplet with Source Disc. + Destylized Triplet to directly compare the proposed Destylized Triplet Loss with a standard triplet objective under the same Source Disc. supervision.

As shown in Table 7, Standard Triplet and SupCon do not achieve the same performance as the full DFL framework. Under the same Source Disc. supervision, Destylized Triplet further improves retrieval performance over Standard Triplet, particularly under unseen styles. These results provide additional evidence that incorporating source–style relationships into triplet learning is beneficial for source retrieval under generative stylization.

Table 7: Comparison with generic metric-learning objectives under the same DINOv2 backbone, training data, and optimization setting. AVG is the arithmetic mean of Seen and Unseen Top-1 accuracy.

Method	Seen ↑	Unseen ↑	AVG ↑
Standard Triplet	90.68	77.15	83.92
SupCon	86.32	73.57	79.95
Source Disc.	93.22	80.23	86.72
Source Disc. + Standard Triplet	94.88	82.07	88.48
Source Disc. + Destylized Triplet	95.37	82.92	89.14
Full DFL	**95.72**	**84.10**	**89.91**

Table 8: Direct transfer of existing image-matching methods under the unseen-style retrieval setting. SSCD, DreamSim, and LPIPS use their publicly available pretrained models without task-specific adaptation. All values report retrieval accuracy (%).

Method	Top-1 ↑	Top-5 ↑
SSCD	62.82	67.25
DreamSim	61.97	70.07
LPIPS-Alex	62.40	67.58
DINOv2 + DFL	**84.10**	**89.23**

C.3 Transferability of Existing Image-Matching Methods

We further examine how existing image-matching methods transfer to our task, including SSCD, DreamSim, and LPIPS. These methods represent different image-matching paradigms related to source retrieval. SSCD is designed for image copy detection and learns retrieval descriptors that are robust to common image transformations, while DreamSim and LPIPS primarily measure perceptual similarity between images.

For these methods, we directly use their publicly available pretrained models without introducing a unified source-classification adaptation objective. They are originally trained under different learning paradigms and matching formulations; in particular, LPIPS operates as a pairwise perceptual distance rather than a standard global retrieval representation. Applying the same source-classification objective to all of them would therefore modify their original training formulations and would not provide a natural method-preserving comparison. We instead use this experiment to examine how directly existing image-matching representations transfer to **Source Retrieval for Stylized Images**.

All methods are evaluated under the same unseen-style retrieval setting, with a gallery of 3,000 unseen source images and 6,000 queries generated using unseen styles. We report Top-1 and Top-5 retrieval accuracy.

As shown in Table 8, SSCD, DreamSim, and LPIPS retain useful source-matching capability under generative style shifts, but their direct transfer performance remains substantially below the task-trained DINOv2 + DFL model. DINOv2 + DFL is included here as a task-trained reference point rather than a training-matched baseline. These results suggest that representations developed for copy detection or perceptual similarity preserve some source-related information, but do not directly address exact-source retrieval under large generative appearance changes.

C.4 Analysis of Curriculum-Based Negative Sampling

To analyze the effect of the switching point in curriculum-based negative sampling, we vary the epoch at which training transitions from random negatives to harder negatives. The switching epoch denotes the number of initial epochs using random negatives before switching to hard negatives. Therefore, an epoch of 0 corresponds to using hard negatives throughout training, while an epoch of 20 corresponds to using random negatives throughout training.

Table 9 shows that transitioning from random to hard negatives around the middle of training provides the best overall performance. The relatively small variation across switching points further suggests that the strategy is not overly sensitive to the exact transition epoch.

Table 9: Effect of the switching epoch in curriculum-based negative sampling. The switching epoch indicates the number of initial epochs using random negatives before switching to hard negatives.

Switching Epoch	Negative-Sample Schedule	Seen ↑	Unseen ↑	AVG ↑
0	Hard negatives throughout	93.93	83.93	88.93
5	5 random + 15 hard	95.13	82.78	88.96
10	10 random + 10 hard	**95.72**	**84.10**	**89.91**
15	15 random + 5 hard	95.45	82.68	89.07
20	Random negatives throughout	95.28	83.37	89.33

Table 10: Source retrieval performance under different gallery sizes. All values report retrieval accuracy (%). Gain denotes the improvement of DFL over Source Disc. in percentage points. Gain is computed from the unrounded accuracies before values are rounded to two decimal places.

Gallery	Queries	Top-1 ↑			Top-3 ↑			Top-5 ↑		
		Src.	DFL	Gain	Src.	DFL	Gain	Src.	DFL	Gain
500	1,000	84.70	**88.30**	+3.60	88.50	**91.90**	+3.40	90.10	**93.40**	+3.30
1,000	2,000	83.30	**87.55**	+4.25	87.35	**91.00**	+3.65	89.00	**91.95**	+2.95
2,000	4,000	81.73	**85.15**	+3.43	86.15	**88.80**	+2.65	88.08	**90.50**	+2.43
3,000	6,000	80.23	**84.10**	+3.87	84.77	**87.60**	+2.83	86.55	**89.23**	+2.68

C.5 Gallery-Size Scaling

To examine how retrieval performance changes as the number of candidate source images increases, we further compare Source Disc. and DFL under gallery sizes of 500, 1,000, 2,000, and 3,000. These settings contain 1,000, 2,000, 4,000, and 6,000 unseen-style queries, respectively.

As shown in Table 10, retrieval becomes more challenging as the gallery size increases, with both Source Disc. and DFL showing lower Top-K accuracy at larger gallery sizes. Nevertheless, DFL consistently outperforms Source Disc. across all evaluated gallery sizes and across Top-1, Top-3, and Top-5 metrics. For Top-1 retrieval, DFL improves over Source Disc. by 3.60, 4.25, 3.43, and 3.87 percentage points at gallery sizes of 500, 1,000, 2,000, and 3,000, respectively. Similar improvements are observed for Top-3 and Top-5 retrieval, indicating that the advantage of DFL is maintained across different gallery scales and retrieval cutoffs.

C.6 Per-Style Analysis on Unseen Styles

To further analyze the effect of different unseen styles on source retrieval, Table 11 reports Top-1 accuracy separately for the six unseen styles and compares DINOv2 zero-shot, SSCD, DreamSim, LPIPS, and the full DFL. All methods are evaluated using the same unseen-style test set and source gallery. The Overall row reports accuracy over the complete unseen-style query set and is therefore not computed as the unweighted mean of the six per-style accuracies.

The results reveal substantial differences in retrieval difficulty across unseen styles. ClassicOil and VintagePhoto are relatively easy for most methods, whereas Bauhaus and StreetArt are more challenging. Compared with DINOv2 zero-shot, DFL consistently improves retrieval performance across different unseen styles, with particularly clear gains on styles where the original DINOv2 representation performs poorly. This indicates that DFL effectively adapts the pretrained representation to source retrieval under style variations. DFL also achieves the best or tied-best performance across all six unseen styles, showing more consistent generalization to different unseen styles.

C.7 Exploratory Open-Set Source Retrieval

Our primary setting focuses on closed-set source retrieval, where the true source image is assumed to be present in the candidate gallery. To further examine the applicability of DFL in a more open scenario, we conduct an exploratory open-set evaluation in which the true source images of some queries are absent from the gallery. The goal is to examine whether out-of-gallery queries can be rejected while preserving source retrieval performance for in-gallery queries.

Table 11: Top-1 retrieval accuracy (%) for individual unseen styles. Overall reports accuracy over the complete unseen-style query set rather than the unweighted mean of the six style-specific accuracies.

Style	DINOv2 Zero-shot	SSCD	DreamSim	LPIPS-Alex	Full DFL
Bauhaus	11.76	18.25	17.75	29.91	**84.45**
ChineseCartoon	38.47	34.39	33.00	55.86	**68.39**
ClassicOil	70.25	91.76	92.46	91.26	**100.00**
StreetArt	5.00	41.73	41.94	21.02	**57.96**
Vaporwave	88.14	90.32	86.36	75.59	**93.28**
VintagePhoto	98.80	**100.00**	99.90	99.90	**100.00**
Overall	49.83	62.82	61.97	62.40	**84.10**

Table 12: Exploratory open-set source retrieval at a target FAR of 1%. Results are reported as mean $\pm$ standard deviation over three source-disjoint folds. All metric values are reported in percentage form.

Method	AUROC ↑	Empirical FAR ↓	DIR ↑	Accepted Top-1 Precision ↑
Similarity Threshold	87.99 ± 0.27	1.17 ± 0.49	65.85 ± 2.00	$\mathbf{99.89 \pm 0.02}$
Binary Classifier	$\mathbf{90.65 \pm 0.12}$	$\mathbf{1.07 \pm 0.82}$	$\mathbf{68.91 \pm 1.90}$	99.84 ± 0.10

We adopt a source-disjoint three-fold evaluation protocol. The evaluation set contains 3,000 source images and 6,000 corresponding unseen-style queries. In each fold, 2,000 source images form the retrieval gallery, while the remaining 1,000 sources are treated as out-of-gallery sources. Source images used for classifier fitting, threshold calibration, and final testing are disjoint. We set the target false acceptance rate (FAR) to 1% and report the mean and standard deviation over the three folds.

We consider two simple extensions for open-set retrieval. **Similarity Threshold** directly uses the maximum cosine similarity between a query and the gallery, with the rejection threshold selected on a separate calibration subset. **Binary Classifier** keeps the DFL embeddings and cosine retrieval ranking fixed, and trains a lightweight logistic classifier using summary statistics of the query-to-gallery similarity distribution, including top-ranked similarities, similarity margins, and distribution statistics. The classifier is used only for accept/reject decisions and does not rerank the retrieved source candidates. We report AUROC, empirical FAR, detection-and-identification rate (DIR), and Top-1 precision among accepted queries.

As shown in Table 12, maximum cosine similarity already provides a useful signal for rejecting out-of-gallery queries, while the lightweight binary classifier further improves known/unknown separation and the detection-and-identification rate under the same 1% target FAR. At the same time, the Top-1 precision among accepted queries remains high.

We emphasize that open-set source retrieval is not the primary focus of this work. Our main study investigates exact-source retrieval under generative style variations when the true source is available in the gallery. The purpose of this experiment is therefore not to provide a complete solution to open-set retrieval, but to examine whether the learned DFL representation can be extended to a more open setting without retraining the retrieval model. The results show that simple post-hoc modeling of gallery-similarity statistics already provides meaningful rejection capability, suggesting a promising direction for extending the current framework to open-set source retrieval.

C.8 Small-Scale External-Generator Evaluation

We further conduct a small-scale test using two external generators, GPT Image 2 OpenAI (2026) and Nano Banana 2 Google (2026). The evaluation contains 10 portrait and 10 landscape source images, with 40 unseen-style queries generated by each model. We directly evaluate the trained DINOv2 ViT-B/14 DFL model without additional fine-tuning.

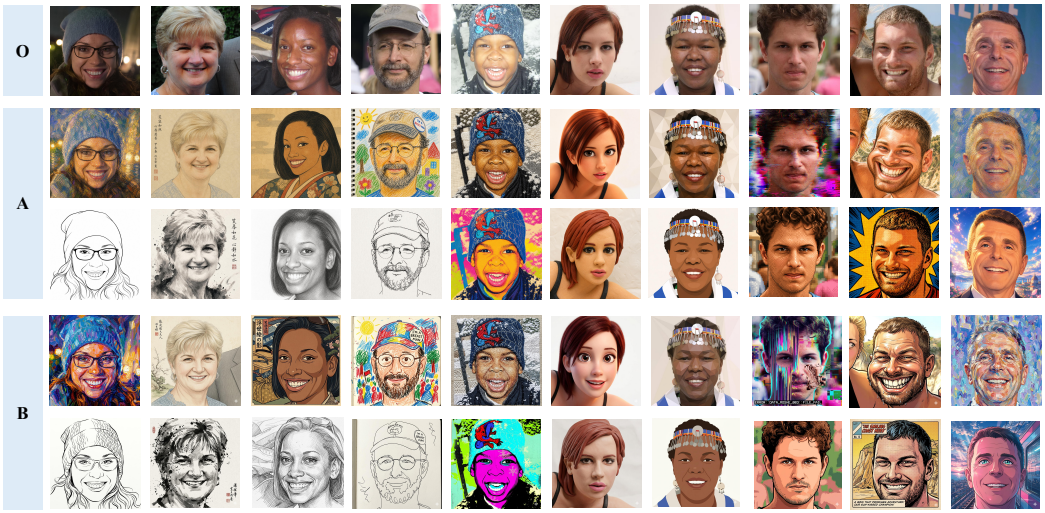

Figure 4: External-generator stress test on portrait images. The first row shows the source images (O), while rows A and B show unseen-style queries generated by GPT Image 2 and Nano Banana 2, respectively. All queries are correctly retrieved to their corresponding source images.

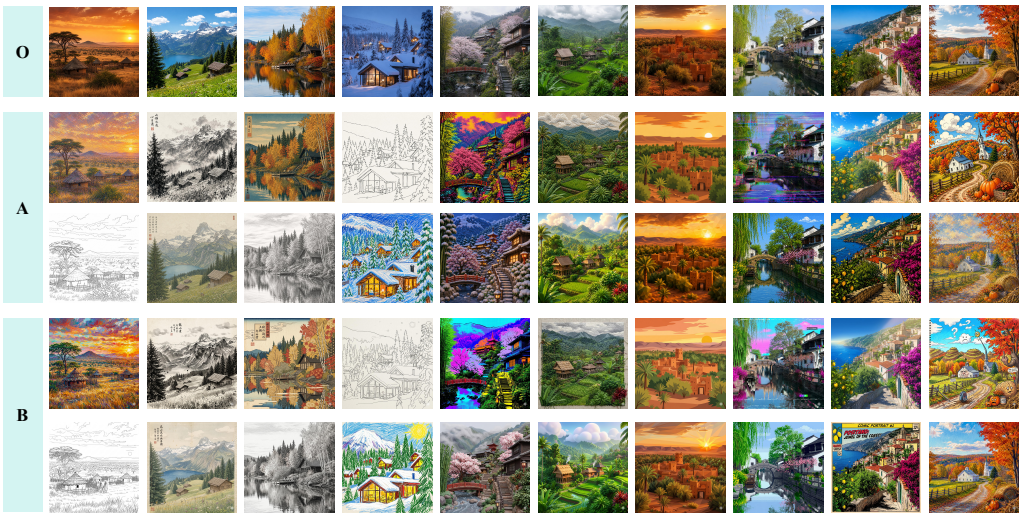

Figure 5: External-generator stress test on landscape images. The first row shows the source images (O), while rows A and B show unseen-style queries generated by GPT Image 2 and Nano Banana 2, respectively. All queries are correctly retrieved to their corresponding source images.

All 80 queries generated by the two external models are successfully retrieved to their corresponding source images. Figures 4 and 5 show examples on portrait and landscape images, respectively. Given its limited scale, we use this experiment only as a supplementary test to the generalization evaluation in the main paper.

C.9 Failure Case Analysis

To further analyze the performance difference between Source Disc. and DFL, we first compare their predictions on all 6,000 queries in the unseen-style test set. Table 13 summarizes the query-level outcomes.

Table 13: Query-level comparison between Source Disc. and DFL on the unseen-style test set. The 6,000 queries are grouped according to whether each method retrieves the correct source image at Top-1.

Outcome	Number of Queries	Percentage (%)
Both correct	4,708	78.47
DFL correct, Source Disc. incorrect	338	5.63
Source Disc. correct, DFL incorrect	106	1.77
Both incorrect	848	14.13

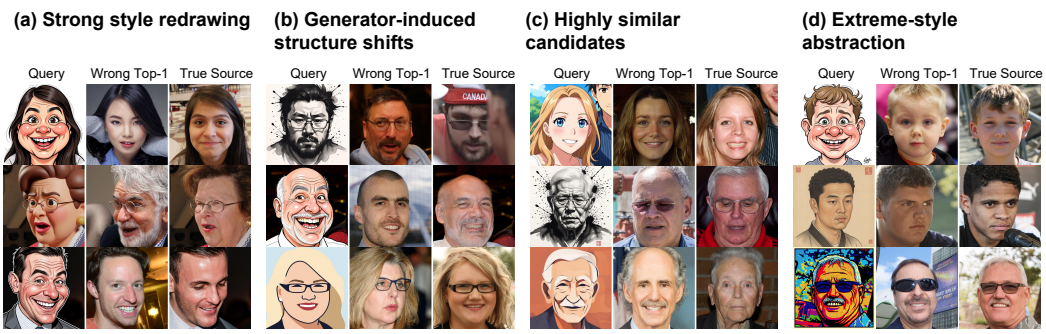

Figure 6: Representative Top-1 retrieval failures under unseen styles, including (a) strong style redrawing, (b) generator-induced structural changes, (c) highly similar source candidates, and (d) extreme style abstraction. Each example shows the stylized query, the incorrect Top-1 retrieved source, and the ground-truth source image.

DFL corrects 338 queries that are missed by Source Disc., while introducing 106 regressions on queries that Source Disc. retrieves correctly. This corresponds to a net gain of 232 correctly retrieved queries, consistent with the Top-1 improvement from 80.23% to 84.10% (+3.87 percentage points). The query-level comparison therefore shows that the overall gain of DFL is associated with substantially more corrected errors than newly introduced errors.

We further analyze representative Top-1 retrieval failures from the unseen-style test set. As shown in Figure 6, representative failures include four common patterns: strong style redrawing, generator-induced structural changes, highly similar source candidates, and extreme style abstraction. Each example shows a stylized query, the incorrect Top-1 retrieved source, and the ground-truth source image.

These cases illustrate that exact-source retrieval remains challenging when generative stylization substantially changes the appearance or structure of the source image. Retrieval can also be confused by visually similar candidate sources in the gallery. These examples highlight remaining challenges under strong generative transformations and fine-grained source ambiguity.

D Ethical Considerations and Limitations

D.1 Ethical Considerations

The source retrieval techniques studied in this work are intended to help identify the potential source image of a generatively stylized image, with possible applications in content provenance, copyright analysis, and generated-content moderation. However, retrieval predictions should not be treated as definitive evidence of copyright ownership, content ownership, or the underlying generation process. In practical applications, retrieval results should therefore be considered together with other evidence, such as content credentials, metadata, and human verification.

Source retrieval may also introduce potential misuse risks. For example, such techniques could be used for unauthorized image tracking, while incorrect retrieval results could lead to false source attribution. We therefore view the proposed method as an assistive tool for source analysis rather

than an automated system for determining copyright ownership or responsibility. Human review should be retained in consequential applications, and conclusions should not be based solely on a single retrieval prediction.

MSIR-Bench is intended solely for source retrieval and related research purposes. We do not support the use of the dataset, or models developed using it, for unauthorized identity tracking, surveillance, harassment, or other applications that may infringe upon individual privacy or rights. Users of the dataset are expected to comply with the applicable terms of the original data sources and generative models, and to use the resource responsibly.

D.2 LIMITATIONS

This work focuses on exact-source retrieval under generative stylization and formulates the problem as closed-set retrieval, where the true source image is present in the candidate gallery. This setting allows us to focus on the effect of style variations on source-image correspondence and to systematically evaluate generalization across unseen styles, generators, and image domains. In more open real-world scenarios, the true source may not always be available in the gallery. Appendix C.7 provides an initial exploration of this setting, while a comprehensive treatment of open-set source retrieval is beyond the scope of this work.

In addition, generative image-editing techniques continue to evolve rapidly, with new generators and styles emerging over time. It is therefore difficult for any fixed benchmark to exhaustively cover all possible generative variations. We aim to provide representative coverage through multiple generators, 25 styles, and different image datasets, with particular emphasis on generalization to unseen styles, unseen generators, and different image domains. As generative editing techniques continue to evolve, the benchmark can be further extended with new generators and data distributions to evaluate source retrieval under emerging generative variations.