

Enlightening Photographic Style Transfer with a Self-Supervised Photographic Embedding

Chengxuan Zhu¹, Jiacong Fang^{1,2}, Shuchen Weng^{1*}, Youwei Lyu²,
Jiajun Tang², Qingnan Fan², Chao Xu¹, and Boxin Shi^{1*}

¹ Peking University ² vivo BlueImage Lab
{peterzhu,shuchenweng,shiboxin}@pku.edu.cn, jiacong_fang@stu.pku.edu.cn

Abstract. Photographic style transfer aims to render a content image with the retouching style of a reference photo while preserving scene structure and texture. Unlike artistic style, photographic style is dominated by subtle, continuous tone/color changes and smooth spatial effects, which are difficult to represent with semantics-oriented or text-supervised image embeddings. We propose **PETAL**, a **P**hotographic **E**mbedding for **T**ransfer with an **A**daptive **L**UT. PETAL learns a dedicated photographic embedding through self-supervised photographic style augmentation and uses it to condition a lightweight adaptive 5D neural LUT. The learned embedding outperforms prior representations in photographic style retrieval, and the full framework enables reference-based style transfer without test-time optimization, achieving stronger quantitative and perceptual performance than state-of-the-art baselines. Project page: <https://petal-pku.pages.dev/>.

Keywords: Photography · Image Retouch · Self-supervised Learning

1 Introduction

Style transfer is a long-standing challenge [25] in computer vision, as it requires correctly understanding the “style” of reference images and faithfully rendering it onto the content images. However, existing methods transfer either conspicuous artistic brushstrokes or only global lightness and tint, which is far coarser than what photographers consider when processing photos. This poses the even more challenging task of *photographic style transfer*: to understand and transfer the nuanced *photographic style*, capturing photographer-introduced global adjustments (*e.g.*, lightness, color and tone), alongside smooth spatial effects like vignetting, while preserving content and texture.

Equal contribution. * Corresponding authors.

¹ Chengxuan Zhu and Chao Xu are with National Key Lab of General AI, School of Intelligence Science and Technology, Peking University. Jiacong Fang, Shuchen Weng, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

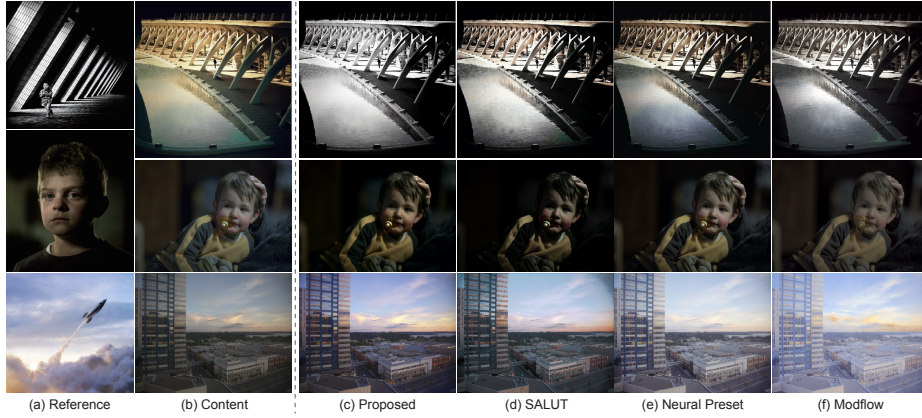


Fig. 1: The task is to transfer the photographic style of (a) reference images to (b) content images while preserving the content and texture of (b). With our photographic embedding, (c) PETAL robustly transfers diverse photographic styles, including high-contrast monochrome, low-key portraits, and saturated purple skies with orange sunlight, all with proper exposure and saturation. Baselines either produce unrealistic results, such as (d) SALUT [28] and (f) Modflow [47], or fail to capture complex styles, as in (e) Neural Preset [44].

Differing in scope from our work, style transfer methods often operate on much coarser features, ranging from the color [18, 53, 65, 66, 71, 102] to the deep features [37, 40, 57] of the reference image, transferring them to the content image, but they tend to distort the texture. Methods based on the lookup table (LUT) [28, 44, 55, 99, 103–105] avoid such distortion through their spatially independent design, but tend to produce unrealistic or altered appearances, as shown in Fig. 1(d, e, f), where none of the baselines can render the first image in black and white, and all yield incorrect contrast levels in the other two examples. We attribute these unsatisfactory results to misaligned features, and argue for the necessity of a fine-grained photographic style embedding to enlighten photographic style transfer.

However, previous image embeddings focus either on semantics [50, 51, 70] or on apparent artistic features [49, 80], leaving the subtle photographic style less explored. Moreover, current training paradigms for image representation are not sensitive to subtle differences among input images, since most previous methods rely on discrete supervision, such as text prompts [70] and classification categories [49, 80]. The continuous nature of photographic style is ill-suited to training with such discrete modalities, as they cannot properly reflect the small but perceptible changes in photographic style.

We therefore introduce **PETAL**, a Photographic Embedding for reference-based photographic style Transfer with an Adaptive LUT. As shown in Fig. 1(c), PETAL is able to transfer the monochrome style and high contrast in the first example, the dark shadow and high contrast in the second example, and the

futuristic pink tint in the third example. Since the LUT has proven effective for style adjustment in previous works [28, 55, 99, 103–105], the key challenge lies in the photographic embedding.

To extract an embedding that truly reflects the photographic style, we face the challenges of representation collapse and unstable optimization. Previous works partially address these problems with an expectation–maximization algorithm, implemented via a moving average [30] or a stop-gradient [16], without requiring negative samples. For photographic style, arbitrary distant patches are unreliable positives because local illumination or object-specific retouching may differ within one image. We therefore use restricted local views as positives and use controlled photographic augmentations to generate style-contrasting negatives.

We design a vision transformer (ViT) paired with histogram encoders in CIE-Lab color space to extract the photographic style. The network training incorporates the stop-gradient operation [16] and hard negative examples [43, 73, 92] derived from photographic style augmentation, for more stable self-supervised learning. On the tasks of retrieval and clustering, the photographic embedding is verified to focus on style rather than on semantics. Bridged by the photographic embedding extracted from any reference image, we then demonstrate that the photographic style of a given image can be transferred with a LUT conditioned on pixel color and location. Without any test-time optimization, PETAL surpasses previous methods in style consistency and visual quality.

In brief, the contributions of this paper are as follows:

- a photographic embedding network that extracts photographic style features from images, enabling similar-style retrieval and the measurement of style difference;
- a self-supervised learning scheme with photographic style augmentation that overcomes data scarcity and learns the underlying photographic style;
- a lightweight LUT that transfers the photographic style from any given reference image to a content image without additional test-time optimization.

2 Related Works

2.1 Image Embedding

Extracting a compact representation from images is a fundamental task in computer vision research. From hand-crafted features [5, 20, 56] to neural embeddings [46, 70, 77], methods increasingly aim to capture semantic content. Early approaches focused on category alignment [35, 46, 77, 84], enabling broad downstream use. Later work aligned images with text [7, 23, 48, 59], with large-scale cross-modal pretraining showing strong results [52, 70, 106]. Notably, CLIP [70] has been widely used in language [1, 8, 85–87], text-to-image [14, 68, 74], and text-to-video models [93, 94, 100], due to its effective cross-modal contrastive learning. Yet prior embeddings still fail to fully capture the nuanced photographic style, which, in addition to the content, is a vital part of visual semantics.

2.2 Image-Only Contrastive Learning

Without text labels or other modalities, contrastive learning can also learn an embedding space for images, by pulling positive samples together and pushing negative samples apart during optimization. The main problems are model collapse and unstable training, caused by under-constrained optimization and a non-convex landscape, respectively. Although various loss functions have been proposed [19, 31, 60, 76, 79], they require an enormous batch size [15, 16] or a specially designed memory bank to provide older representations [34], both of which are computationally expensive. Recent works attempt to remove the need for negative samples. VICReg [10] introduces a regularizer that explicitly supervises the variance and covariance among samples. BYOL [30] and SimSiam [16] adopt the expectation-maximization approach, using a moving average and a stop-gradient operation, respectively, to avoid model collapse. For the task of this paper, however, a vanilla network design may fail to distinguish photographic styles, such as different exposure levels. We borrow the stop-gradient from SimSiam [16] and incorporate negative samples into the self-supervised learning scheme.

2.3 Style Transfer

Prior methods align global lightness using physical models [26, 72, 82, 101], or model color mapping with 3D LUTs [55, 99, 103–105] and parameter sets [36, 44]. Although these methods can seemingly align the overall color and lightness, the naively extracted photographic style differs from human perception. White-box sequential edits [38, 62, 89, 95] offer finer control but struggle with the interdependencies among image-editing operations. Diffusion-based approaches [22, 91] add semantic awareness, but can distort structure, drift in color, and are computationally heavy. We propose a lightweight neural 5D LUT that transfers the extracted photographic style, conditioned on pixel color in CIE-Lab color space and spatial coordinates. Enlightened by the trained photographic embedding, the proposed photographic style transfer achieves a finer understanding of the actual photographic style.

3 Methods

In Sec. 3.1, we first introduce the photographic embedding network. The ViT backbone merges the pixel values and the global distribution from the histogram feature with a learnable [CLS] token. Through a self-supervised contrastive learning scheme with photographic style augmentation for negative samples, as discussed in Sec. 3.2, the extracted photographic embedding is able to reflect the photographic style, as further validated by style retrieval and clustering. Finally, we present the simple yet effective photographic style transfer design that builds upon the photographic embedding, detailed in Sec. 3.3.

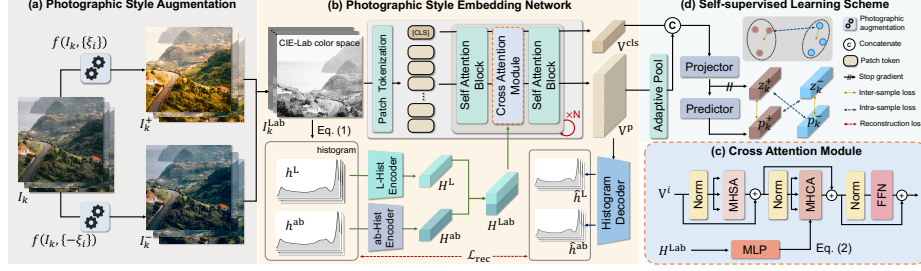


Fig. 2: The pipeline of the photographic embedding network and the training process. (a) An anchor crop is jittered into K overlapping local views, and each I_k is then applied with the photographic style augmentation controlled by $\{\xi_i\}$ and $\{-\xi_i\}$ to derive two photographically contrastive batches. (b) Each batch is converted to the CIE-Lab color space, and processed by the ViT-based network to extract photographic embeddings. (c) The luminance and chromaticity histograms (h^L, h^{ab}) are encoded and injected via cross attention module to guide the representation learning. (d) The photographic embeddings z_k^+ and z_k^- are projected to p_k^+ and p_k^- for contrastive learning.

3.1 Photographic Embedding

As a complex combination of color, lightness, and other visual cues, photographic style is challenging to describe in words. Its continuous nature is also quite different from natural language: for example, a change in exposure may be far more subtle than the change from “dim” to “slightly dim”, yet text prompts cannot reflect this, as they are discrete by nature. Nevertheless, professional photographers can easily recognize such styles, as well as slight changes in photographic style. Inspired by this observation, we design a ViT-based embedding network $\mathcal{M}$ to extract a photographic embedding by integrating global semantics, aggregated through the [CLS] token, with tonal and color statistical features guided by luminance–chromaticity histograms in the CIE-Lab color space, as illustrated in Fig. 2(b). The [CLS] token serves as a compact global descriptor that complements pooled patch-token statistics; as shown in Tab. 3, using the [CLS] token alone is insufficient, motivating its concatenation with aggregated patch features.

Given an input image I , we first convert it to CIE-Lab color space (I^{Lab}) to decouple luminance and chromaticity components [39, 42]. I^{Lab} is then divided into patches and tokenized, along with the [CLS] token, to form the input sequence for the main transformer blocks. Meanwhile, to extract tonal and color statistical features, we compute the luminance histogram $h^L \in \mathbb{R}^H$ from L channel (I^L) and the chromaticity joint histogram $h^{ab} \in \mathbb{R}^{H \times H}$ from the ab channels (I^{ab}). Following previous methods [3, 4], the histogram computation is implemented as a differentiable process using the inverse-quadratic kernel [4] $\mathcal{K}(\cdot)$:

$$h^L(i) = \sum_{p=1}^N \mathcal{K}(I^L(p), b_i), \text{ and } h^{ab}(i, j) = \sum_{p=1}^N \mathcal{K}(I^a(p), b_i) \cdot \mathcal{K}(I^b(p), b_j), \quad (1)$$

where N is the number of pixels, b_i and b_j are the bin centers, and p denotes a pixel location. These two histograms are processed by their respective encoders (L-Hist Encoder and ab-Hist Encoder), and the resulting features are then stacked to form the histogram feature H^{Lab} . The histogram features explicitly contain global tonal and chromatic statistics, which are difficult to infer reliably from pixel values. These features are injected via cross-attention, allowing global distribution cues to guide spatial representations without disrupting their structural properties. As illustrated in Fig. 2(d), each cross-attention module contains an MLP that projects the histogram feature H^{Lab} into $\tilde{H}^{\text{Lab}}$.

Finally, the photographic style embedding is obtained by integrating global and local representations as

$$z = h([V^{\text{cls}}; \mathcal{P}(V^{\text{P}})]). \quad (2)$$

where V^{cls} and V^{P} denote the feature representations of the [CLS] token and the patch tokens obtained from the ViT network, respectively. $\mathcal{P}(\cdot)$ denotes the learnable adaptive pooling over the patch tokens, and $h(\cdot)$ is a projection MLP following the asymmetric design of SimSiam [16].

3.2 Self-Supervised Embedding Learning

With the module designs providing the network with the information necessary to extract the photographic style, the remaining problem is the absence of high-quality labels. Although many previous methods manage to learn a semantic representation with text supervision [70, 88], the discrete text representation is unfit to supervise the subtle differences in photographic style. We therefore propose a self-supervised learning paradigm that optimizes the photographic embedding space by relying on images alone.

To optimize the encoded latent space specifically for the photographic style, we design a self-supervised embedding learning scheme. As illustrated in Fig. 2(a), we first sample an anchor crop from image I and apply small spatial jitters to obtain K overlapping local views. Each patch is augmented with two sets of parameters, which aim to perturb the input image toward contrary photographic styles. Denoted $f(I, \{\xi_i\})$, the transform takes a set of parameters $\xi_i \in [-1, 1]$, with $f(I, \mathbf{0}) \equiv I$ being the identity mapping. Each parameter ξ_i controls the adjustment of one of the M predefined aspects of photographic style (*e.g.*, exposure, saturation, tint, highlight, and shadow). In this way, we create two photographically augmented versions of the sample, which serve as negative examples to each other.

Photographic Style Augmentation (PSA). For reproducibility, $f(\cdot, \{\xi_i\})$ is designed as a random composition of differentiable photo-editing operators implemented in PyTorch. Each operator is parameterized by a scalar $\xi \sim \mathcal{U}([-1, 1])$ and reaches its maximum strength at $|\xi| = 1$; we keep the ranges intentionally conservative to avoid introducing out-of-distribution artifacts while still producing hard negatives in photographic style. The specific choices of style augmentations are detailed in the supplementary material.

Following previous methods [12] and our objective, the overlapping local views under the same PSA parameter in the same image are considered positives. They share most content, local illumination context, and smooth tone/color rendering, which mitigates false positives from distant patches or strongly localized edits. As for the negative samples, we choose the pair of images from the same patch that undergo contrary augmentations, namely $f(I_k, \{\xi_i\})$ and $f(I_k, \{-\xi_i\})$. For notational convenience, we denote $z_k^+ = \mathcal{M}(f(I_k, \{\xi_i\}))$ and $z_k^- = \mathcal{M}(f(I_k, \{-\xi_i\}))$. The predictor outputs are abbreviated as $p_k^+ = h(z_k^+)$ and $p_k^- = h(z_k^-)$.

As shown in Fig. 2(d), owing to the symmetric design of the proposed pipeline, we define an ‘‘intra-sample loss’’ $\mathcal{L}_-$ that attracts positive samples cropped from the same image, and an ‘‘inter-sample loss’’ $\mathcal{L}_+$ that repels negative samples from different augmented versions of the same patch, given by

$$\mathcal{L}_- = -\mathbb{E}\left[\left\langle p_k^+, \frac{\text{sg}(\sum_{i=1}^K z_i^+)}{K} \right\rangle + \left\langle p_k^-, \frac{\text{sg}(\sum_{i=1}^K z_i^-)}{K} \right\rangle\right], \quad (3)$$

$$\mathcal{L}_+ = \mathbb{E}\left[\text{CE}\left(\left\langle p_k^+, \text{sg}(z_k^+) \right\rangle, \left\langle p_k^-, \text{sg}(z_k^+) \right\rangle\right) + \text{CE}\left(\left\langle p_k^+, \text{sg}(z_k^-) \right\rangle, \left\langle p_k^-, \text{sg}(z_k^-) \right\rangle\right)\right]. \quad (4)$$

where $\text{sg}(\cdot)$ is the stop-gradient operation, and $\langle \cdot, \cdot \rangle$ and $\text{CE}(\cdot, \cdot)$ are notations for cosine similarity and binary cross entropy respectively, defined by $\text{CE}(s_{\text{pos}}, s_{\text{neg}}) = -\log \frac{\exp(s_{\text{pos}})}{\exp(s_{\text{pos}}) + \exp(s_{\text{neg}})}$. Eqs. (3) and (4) are designed to encourage matching the prediction to its own-view embedding while suppressing the opposite augmentation. The intra-sample loss matches each prediction to the (stop-gradient) center of its own local-view batch, encouraging invariance to crop and spatial jitter. In contrast, the inter-sample loss separates two contrary PSA views of the *same* patch. Because the content and layout are fixed while tone/color are deliberately changed in opposite directions, these samples prevent content-matching shortcuts and form hard negatives for photographic style.

In addition to the aforementioned contrastive objective, we introduce the histogram reconstruction loss $\mathcal{L}_{\text{rec}}$, which operates on the patch token features V^{p} , inspired by recent representation learning practices that regularize intermediate features with auxiliary reconstruction objectives (*e.g.*, SigLIP-2 [88] and related self-supervised vision encoders). We design a lightweight CNN to reconstruct both the luminance histogram (h^{L}) and the chromaticity joint histogram (h^{ab}) from V^{p} , yielding the reconstructions $\hat{h}^{\text{L}}$ and $\hat{h}^{\text{ab}}$. The loss $\mathcal{L}_{\text{rec}}$ is defined by

$$\mathcal{L}_{\text{rec}} = \mathcal{H}(h^{\text{L}}, \hat{h}^{\text{L}}) + \mathcal{H}(h^{\text{ab}}, \hat{h}^{\text{ab}}), \quad (5)$$

where $\mathcal{H}(\cdot)$ denotes the Hellinger distance [4]. Consequently, the final overall optimization objective is given by $\mathcal{L} = \mathcal{L}_- + 0.5 * \mathcal{L}_{\text{rec}} + 0.1 * \mathcal{L}_+$, where these coefficients are set to normalize the losses to comparable magnitudes.

3.3 Photographic Style Transfer

PETAL pairs the photographic style embedding from $\mathcal{M}$ with a lightweight adaptive neural LUT to enable reference-based photographic style transfer, as

illustrated in Fig. 3. The neural LUT is composed of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, both of which are 1×1 CNNs, so that the detailed texture can be preserved.

Let I_c denote the content image and I_s denote the style reference image. We first compute per-pixel features $x_c = \mathcal{E}(I_c, P)$, where $P \in [0, 1]^2$ is the normalized pixel coordinate map. Denote by $\mu(\cdot)$ and $\sigma(\cdot)$ the channel-wise mean and standard deviation of x_c , and let $g_\mu(\cdot)$ and $g_\sigma(\cdot)$ be two lightweight MLPs that map a photographic embedding to an affine shift in mean/variance. We then apply an embedding-conditioned AdaIN-like re-normalization:

$$\hat{I} = \mathcal{D}(\tilde{\mu} + \text{Norm}(x_c) \odot \tilde{\sigma}), \text{ where } \text{Norm}(x_c) = \frac{x_c - \mu(x_c)}{\sigma(x_c) + \epsilon}. \quad (6)$$

In addition, $\tilde{\mu} = \mu(x_c) + g_\mu(\mathcal{M}(I_s)) - g_\mu(\mathcal{M}(I_c))$, and $\tilde{\sigma} = \sigma(x_c) + g_\sigma(\mathcal{M}(I_s)) - g_\sigma(\mathcal{M}(I_c))$ are the standard deviation and mean for re-normalization. $\mathcal{E}$ and $\mathcal{D}$ are the encoder and decoder, implemented with 1×1 CNNs. This choice ensures computational efficiency, avoids spatial mixing to preserve local textures, and admits a natural 5D (Lab+xy) look-up-table interpretation, facilitating responsive deployment.

This injects the style delta between I_s and I_c through the embedding space. Since x_c and $\hat{I}$ are determined by pixel color (in CIE-Lab) and spatial coordinates, the module is a 5D neural LUT by design. Eq. (6) reduces to approximate identity mapping when $I_s \approx I_c$, since $\mathcal{M}(I_s) \approx \mathcal{M}(I_c)$.

To train the style-transfer network, the learned embedding network $\mathcal{M}$ is frozen. We sample a content patch I_c and synthesize a paired style target $I_s = f(I_c, \{\xi_i\})$ using the aforementioned PSA; I_s therefore serves as both the style reference and the supervised ground truth. The LUT is trained to map $I_c \mapsto I_s$ by conditioning on $(\mathcal{M}(I_c), \mathcal{M}(I_s))$. We supervise the training of this part with a weighted sum of: (i) pixel-wise mean squared error, (ii) perceptual loss [107], (iii) an embedding-based style loss $1 - \langle \mathcal{M}(\hat{I}), \mathcal{M}(I_s) \rangle$, and (iv) an identity loss $\|\mathcal{D}(\mathcal{E}(I_c)) - I_c\|_2$ to stabilize the near-identity regime.

Note that we adopt two-stage training mainly for stability and efficiency. Contrastive learning of $\mathcal{M}$ is memory-intensive, and joint training would require backpropagating through $\mathcal{M}$ while optimizing the dense per-pixel LUT, at least doubling the memory cost and slowing optimization. In practice, freezing $\mathcal{M}$ yields a strong and stable conditioning signal for LUT training.

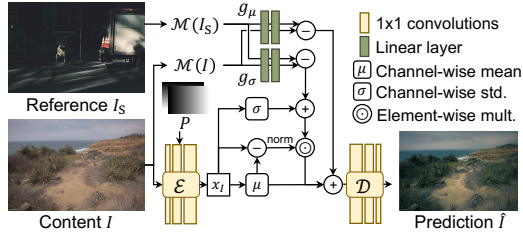


Fig. 3: The PETAL transfer network design, with $\mathcal{M}$ being the trained photographic style embedding network.

4 Experiments

Datasets. For training, we use high-quality images from Unsplash [17] to avoid reliance on discrete text representations.

To evaluate style retrieval, we adopt three datasets. **PPR10K subset** [54] contains 1,681 portrait scenes, each captured in three distinct photographic styles. Due to the high resolution (>20M pixels), we use the first 245 scenes for benchmarking. **FiveK-Concept** [11] is constructed from FiveK [11] to assess sensitivity to fundamental photographic concepts. Images are grouped by genre and subject, then clustered into 623 groups (7 images per group on average) using K-means on Qwen3-VL [9] annotations encoded by EmbeddingGemma [90]. A professional photographer edits each group’s RAW files in Adobe Lightroom [2] by varying exactly one concept (*e.g.*, contrast, temperature, or tint) to produce a single consistent style per group. **INSRANK** comprises 32 Instagram-based groups curated from photographers’ works. Each group includes one query and five candidates, which are independently ranked by two professional photographers according to style similarity; their rankings serve as pseudo ground truth.

For the style transfer evaluation, we conduct quantitative comparisons on PST50 [28] and PPR10K subset [54]. In PPR10K, the first image of each expert-defined style serves as the reference, and the remaining images in the same scene are used for evaluation. Importantly, both PPR10K and PST50 provide ground-truth retouched targets under consistent style definitions; thus, the reference-based metrics in Tab. 2 compare $\hat{I}$ against these ground truths, reflecting both content preservation and style fidelity. Although PPR10K is portrait-focused, its benchmark labels mainly reflect whole-image global and smooth regional retouching; face-specific local retouching remains outside our primary scope and is discussed in the limitations.

For qualitative comparison, we evaluate on the DPST test set [57], and additionally use images from the Hasselblad Masters dataset [32] and INSRANK as style references due to their strong individual photographic characteristics.

Baselines. For a scope similar to that of this paper, we compare against the photo-style embeddings of CSD [80], DiffSim [81], DreamSim [24], FreeStyleRet (FSR) [49], and ALADIN [75], as well as classic Gram-matrix features from Gatys *et al.* [25]. We also include the color discriminator (ColorD) from Neural Preset [44] and the intermediate representation of Modflow [47], both of which can serve as style-aware representations. For fairness of comparison, we also report DINOv3* [78] and CLIP* [70] fine-tuned with the same training data and objectives as our method. Pretrained image embeddings are also considered, including models pretrained on text-image pairs, such as CLIP [70], BLIP [51], and BLIP-2 [50], and those trained with self-supervision, such as DINOv2 [61], ImageBind [27], SigLIP-2 [88] and DINOv3 [78], which are designed for general image embedding tasks.

For photographic style transfer, we compare with methods that operate without per-reference test-time optimization, aligning with our method for fair comparison. Representative methods in photo-realistic style transfer are compared, including PhotoWCT² [18], DSTN [37], StyTr² [21], CAP-VSTNet [97], and SALUT [28]. We also compare with image retouching methods of Deep Preset [36], Neural Preset [44] (closed source, application access only), and Modflow [47]. We also consider the commercial products of Pixelmator Pro [67] where

Table 1: Quantitative photographic style retrieval performance for embeddings. All the values are shown as percentages, and higher numbers indicate better performance. Best values are in **bold**, and second best are underlined. * indicates fine-tuning on our training data with the same objectives.

Method	#Para.	PPR10K [54] subset				FiveK-Concept [11]				INSRANK		
		R@1	R@5	mAP	F1@5	R@1	R@5	mAP	F1@5	R@1	ρ [83]	τ [45]
General image embeddings												
CLIP [70]	428M	0.08	95.33	31.62	23.82	0.89	3.07	9.72	0.72	34.38	15.32	13.17
BLIP [51]	225M	8.81	89.60	35.78	26.76	1.52	2.40	10.21	0.72	46.88	15.62	11.87
BLIP-2 [50]	3.745B	1.54	91.07	31.46	22.90	1.28	2.55	10.11	0.69	37.50	29.06	23.75
DINOv2 [61]	304M	1.48	90.32	31.23	22.54	0.05	0.14	8.10	0.03	37.50	27.81	21.25
DINOv3 [78]	303M	1.48	97.36	30.81	22.15	0.02	0.10	8.42	0.02	43.75	25.31	21.25
ImageBind [27]	1.201B	0.62	97.09	32.02	23.84	0.09	24.53	17.71	4.89	37.50	16.87	14.37
SigLIP-2 [88]	1.137B	1.48	90.61	31.22	22.64	0.70	1.53	9.63	0.35	37.50	32.19	23.75
Image style embeddings												
CSD [80]	305M	6.77	98.00	38.44	32.42	8.35	23.14	17.85	6.70	37.50	<u>40.94</u>	<u>33.12</u>
FSR [49]	940M	0.26	94.16	30.25	21.62	0.88	4.27	9.21	0.94	25.00	4.06	3.75
DiffSim [81]	1.06B	7.03	94.99	33.51	25.82	8.67	10.47	11.31	1.38	<u>50.00</u>	19.37	15.00
DreamSim [24]	93.2M	8.80	93.23	34.17	23.69	1.04	7.83	8.57	1.65	40.62	24.06	19.37
ALADIN [75]	86.3M	32.44	98.87	52.69	44.35	27.67	55.88	32.46	21.26	46.88	33.12	29.37
Gram [25]	0.113M	26.24	96.48	42.39	31.82	25.49	51.77	25.90	17.46	31.25	32.19	26.87
Modflow [47]	59.63M	28.18	93.04	43.47	33.33	19.49	44.27	22.59	14.50	34.38	14.69	13.12
ColorD [44]	42.56M	<u>65.80</u>	97.86	<u>64.15</u>	<u>50.25</u>	<u>45.45</u>	79.76	<u>41.25</u>	<u>31.87</u>	46.88	13.44	8.12
Fine-tuned baseline representations												
DINOv3* [78]	303M	59.30	94.38	61.78	46.59	34.72	66.49	36.61	24.88	34.38	15.31	13.13
CLIP* [70]	428M	54.42	93.87	58.99	44.01	34.66	65.85	37.90	26.81	<u>50.00</u>	19.37	17.50
Proposed emb.	362M	74.27	<u>98.40</u>	71.99	56.60	52.94	<u>78.56</u>	51.84	39.01	62.50	52.19	46.88

we use the style matching tool, and also Nano Banana Pro [29] where we specify style transfer prompts.

Evaluation Metrics. To evaluate the style embedding, we follow previous works and report the recall among top- k predictions ($R@k$) [6, 27, 41, 49, 50, 60, 88], mean Average Precision (mAP) [13, 61, 63, 64, 69], and F1 score among top- k predictions ($F1@k$). Higher $R@k$ values, especially $R@1$, indicate a stronger ability of the embedding to distinguish photographic style from image content. mAP measures the proportion of false predictions, and $F1@k$ provides a balanced view of precision and recall. The higher these metrics are, the better the performance. Note that $R@5$ is defined differently in this field than in classification tasks. For INSRANK, we evaluate the alignment of predictions with human preference using Spearman’s ρ [83] and Kendall’s τ [45]. The mathematical definitions of the above metrics are provided in the supplementary material.

As for the photographic style transfer task, we report Peak Signal to Noise Ratio (PSNR), structural similarity index (SSIM) [96], LPIPS [107] and histogram correlation (H-Corr) [28] to measure the similarity between the prediction and the ground truth. Together, the metrics measure per-pixel color accuracy, structural preservation, and histogram similarity. To further measure perceptual style fidelity, we further report LIQE [108], Q-Align [98], MS-SWD [33], and Gemini-based VLM judging [85] in the supplementary material, which supports the same conclusion.

Table 2: Quantitative comparisons for photographic style transfer on the PPR10K [54] subset and PST50 [28] dataset, the user preference scores, and average inference time. $\uparrow$ ($\downarrow$) denote higher (lower) is better.

Method	Time (s)	PPR10K [54] subset				PST50 [28]				User Study		
		LPIPS↓	PSNR↑	SSIM↑	H-Corr↑	LPIPS↓	PSNR↑	SSIM↑	H-Corr↑	(i)↑	(ii)↑	(iii)↑
Open-source methods												
SALUT [28]	0.3117	0.2464	16.47	0.6307	0.2464	0.1557	21.93	0.8733	0.4036	4.26	3.38	3.62
Modflow [47]	0.3680	0.2160	19.64	0.7914	0.3815	0.1707	20.01	0.8599	0.3373	4.15	3.68	3.85
CAP-VSTNet [97]	0.4633	0.2172	20.05	0.7858	0.4199	0.1698	20.12	0.8400	0.3669	3.65	4.03	3.21
Deep Preset [36]	0.0329	0.1717	<u>20.88</u>	0.8137	0.4410	0.1365	<u>23.40</u>	0.8895	0.4194	4.65	3.12	3.97
StyTr ² [21]	0.6565	0.2779	18.62	0.6644	<u>0.4667</u>	0.4424	19.54	0.7224	0.4054	2.09	3.29	3.03
PhotoWCT ² [18]	0.3408	<u>0.1538</u>	18.95	<u>0.8366</u>	0.3989	0.1762	18.43	0.8365	0.3450	2.73	3.62	3.41
DSTN [37]	0.9809	0.2982	16.84	0.5726	0.3972	0.3843	16.97	0.5525	0.3244	1.53	2.94	2.09
Closed-source applications												
Neural Preset [44]	N/A	0.1377	19.12	0.8048	<u>0.4858</u>	<u>0.1210</u>	22.61	<u>0.8975</u>	<u>0.4392</u>	<u>4.32</u>	4.15	<u>4.35</u>
Pixelmator Pro [67]	N/A	0.1207	19.43	0.8825	0.3824	0.1526	21.60	0.8684	0.3121	4.41	3.69	3.81
Nano Banana [29]	N/A	0.2817	16.91	0.6509	0.4927	0.3069	18.42	0.6287	0.3836	3.74	<u>4.31</u>	4.47
PETAL	<u>0.2306</u>	0.1027	21.80	0.8710	0.4578	0.1021	24.63	0.9272	0.5220	4.65	4.71	<u>4.44</u>

4.1 Quantitative Comparisons

Photographic Embedding. As discussed earlier, pretrained image encoders can extract image semantics, which is further confirmed by the results in Tab. 1. Note that their R@1 scores are still significantly worse than that of the proposed photographic embedding, yet they all achieve a high R@5 score on the PPR10K [54] subset, because that dataset contains only 3 kinds of photographic style. In contrast, the FiveK [11] dataset is more challenging for retrieval, as each scene is aggregated from images with similar content rather than the exact same scene. For the two style embeddings, CSD [80] and FSR [49] perform similarly to the other embeddings, indicating the gap between artwork style and photographic style. On INSRANK, the proposed photographic embedding also shows the highest alignment with human preference in both the top-1 prediction and the ranking, measured by Spearman’s ρ [83] and Kendall’s τ [45]. Crucially, even when we fine-tune strong pretrained backbones on our training data with the same objectives (DINOv3* and CLIP*), their R@1 scores improve substantially (*e.g.*, 59.30 and 54.42 on PPR10K) yet remain below the 74.27 achieved by the proposed photographic embedding, suggesting that the network design is important for capturing photographic style. These results confirm that previous methods target semantics rather than photographic style, highlighting the efficacy of the proposed photographic embedding.

Photographic Style Transfer. In Tab. 2, we first evaluate the running time by transferring content images with the resolution of 1080×1920 , using reference images of size 512×512 as guidance. PETAL processes a 1080×1920 input in 0.23s, including photographic-embedding extraction and the LUT forward pass, and is the second-fastest method among the baselines.

We then report the quantitative performance of each method on the two benchmark datasets. PETAL attains the best PSNR and LPIPS on both datasets,

together with the best SSIM and H-Corr [28] on PST50 [28], confirming accurate, content-preserving transfer with superior perceptual quality [107]. The second-best Neural Preset [44] lags in PSNR, while Deep Preset [36] matches PSNR but trails in H-Corr. We note that H-Corr, being a purely global histogram statistic, can be inflated without faithful transfer: on the PPR10K [54] subset, StyTr² [21] and Nano Banana Pro [29] report higher H-Corr than PETAL yet severely alter the content, as their much lower PSNR and SSIM and the qualitative results in Fig. 5 reveal. Additional perceptual and VLM-based evaluations are provided in the supplementary material.

4.2 User Study

To assess perceptual quality, we conducted a user study with 34 volunteers. For each input, participants viewed the stylized results produced by each method and assigned a score from 1 (worst) to 5 (best) for three aspects: (i) content preservation, (ii) style consistency, and (iii) overall visual quality, as listed in Tab. 2. PETAL ranks first in both content preservation (tied with Deep Preset [36]) and style consistency, where it leads the runner-up by a clear margin, and ranks second in overall quality, narrowly behind Nano Banana Pro [29]. The competing methods’ failure modes are reflected in their ratings: Nano Banana Pro [29] scores highest on overall quality but modifies the scene content, yielding a much lower content-preservation rating; Deep Preset [36] preserves content well but leaves the original style largely intact, giving the lowest style consistency among the top methods; and Neural Preset [44] is competitive overall but may misinterpret the style, limiting its style consistency.

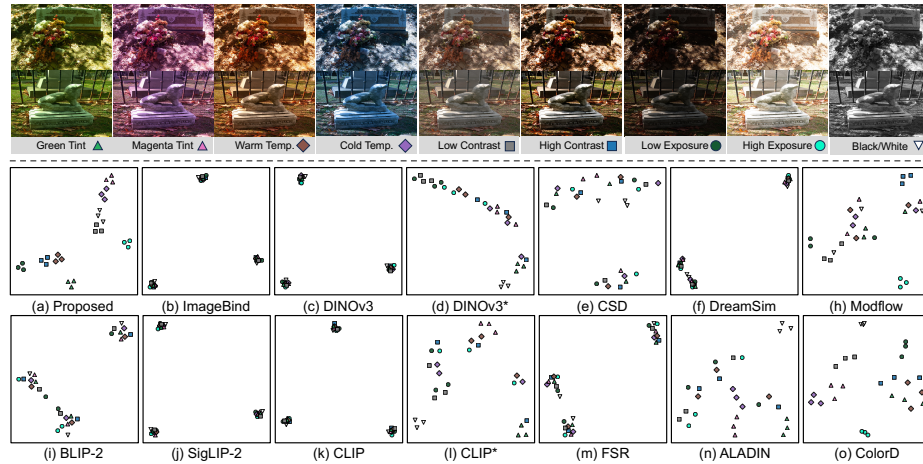


Fig. 4: A t-SNE [58] visualization of the proposed photographic embedding and baseline methods, conducted on the FiveK-Concept [11] dataset. Ideally, photographic-style-aware representations should have 9 clusters, each sharing one kind of shape, where each shape denotes one photographic adjustment.



Fig. 5: Qualitative comparisons for photographic style transfer. Reference images are shown in the bottom-right red boxes of the first column.

4.3 Qualitative Comparisons

First, we visualize the photographic embedding space in Fig. 4. An ideal photographic embedding should reflect the distance in photographic style: with images adjusted by different photographic concepts, it should map those with similar adjustments closer to each other and form clusters in the t-SNE [58] visualization. The FiveK-Concept [11] subset contains three content-similar source images, each retouched with the same set of 9 photographic adjustment types; thus, a photographic-style-aware embedding should form 9 style-driven clusters, each containing the three images sharing one adjustment. As shown in Fig. 4(a), the proposed photographic embedding follows this behavior. In comparison, semantic embeddings such as DINO [61, 78] are mainly driven by content and are unaware of photographic style, while the specially trained color discriminator ColorD [44] still fails to distinguish some fine-grained photographic styles.

We then compare PETAL with baselines across landscape, portrait, stylized lighting, architectural, and indoor examples in Fig. 5, where the reference for each row is provided as a bottom-right inset in the first column. As shown, PETAL faithfully transfers the reference exposure, color temperature, saturation, and contrast while preserving scene geometry and fine textures. In contrast, Neural Preset [44] and Deep Preset [36] consistently under-transfer the reference style (*e.g.*, missing the cool tones in the first three rows), and SALUT [28] often yields

incorrect contrast levels. Modflow [47] applies coarse global shifts that introduce unnatural color casts, failing to maintain local tonal balance. Furthermore, CAP-VSTNet [97], PhotoWCT² [18], and StyTr² [21] severely distort the content, introducing spatial color bleeding, structural artifacts, and texture leakage.

In Fig. 6, we interpolate the conditioning photographic embedding, and find that the output shows a linear interpolation between the two reference photographic styles, and the content is kept throughout the interpolation.

In Fig. 7, we additionally test a sequence of transfer, using the previous output as the next input. After multiple rounds, the content structure and the target style remain stable, and reapplying the same reference in later steps yields only minimal changes, indicating that the LUT behaves near-identity when content and reference share similar photographic style.

An additional PPR10K embedding-space visualization, more qualitative comparisons, and failure-case analysis are provided in the supplementary material.



Fig. 6: Interpolating the photographic embedding leads to the interpolation of photographic style between the two shown reference images, and the content is kept throughout the process.

4.4 Ablation Study

We first verify the proposed photographic embedding network design and the self-supervision scheme in Tab. 3. Changing the color space from CIE-Lab to RGB degrades the representation, as the former disentangles luminance and chromaticity and thus allows easier extraction of photographic style. Removing the cross-attention module also removes the information from the histogram encoders; judging from the performance, the global statistical information remains necessary and cannot be learned directly by a ViT. Removing the adaptive pooling or replacing it with mean pooling degrades the performance, showing that the adaptive pooling aggregates the important features in an image. As for the supervision, we find the intra-sample loss to be the most important, and the combination of all three loss terms yields the best performance.

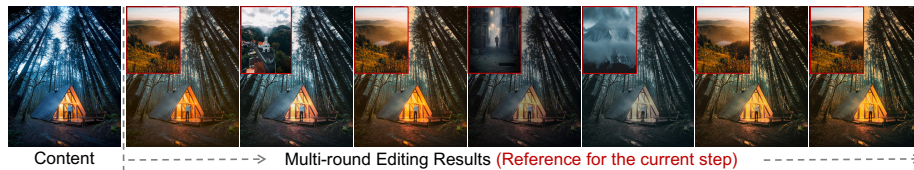


Fig. 7: Sequential photographic style transfer results (from left to right), with the reference for each round shown in the top-left red box.

Table 3: An ablation study of photographic embeddings on retrieval performance and quantitative results for photographic style transfer.

Setting	PPR10K [54] subset				FiveK-Concept [11]				PST50 [28]			
	R@1	R@5	mAP	F1@5	R@1	R@5	mAP	F1@5	LPIPS↓	PSNR↑	SSIM↑	H-Corr↑
RGB color space	70.34	96.77	69.16	53.61	51.21	78.29	<u>49.37</u>	36.58	<u>0.1068</u>	22.22	<u>0.9120</u>	0.4034
W/o cross attn	68.93	96.66	66.44	51.21	50.29	75.91	<u>48.30</u>	36.19	0.1120	22.24	0.9005	0.4436
[CLS] only	<u>73.74</u>	96.88	<u>70.01</u>	<u>54.36</u>	49.17	<u>80.14</u>	48.64	35.46	0.1131	23.11	0.9086	0.4355
[CLS]+ mean pool	68.91	<u>97.25</u>	66.53	51.25	50.48	79.24	47.61	34.78	0.1149	22.49	0.9001	0.4342
W/o $\mathcal{L}_{\text{rec}}$	66.98	96.32	66.06	51.13	45.86	78.14	47.58	34.04	0.1354	21.69	0.8967	0.4048
W/o $\mathcal{L}_{-}$	36.09	89.79	45.42	33.38	18.64	51.67	21.62	14.56	0.1249	<u>23.47</u>	0.9063	<u>0.4506</u>
W/o $\mathcal{L}_{+}$	67.53	96.75	65.03	49.79	<u>51.79</u>	81.51	49.36	<u>36.69</u>	0.1071	22.41	0.9045	0.4194
Complete	74.27	98.40	71.99	56.60	52.94	78.56	51.84	39.01	0.1021	24.63	0.9272	0.5220

For the ablation of the photographic style transfer part, please refer to the supplementary material, where we train the LUT after substituting the photographic embedding with baseline image embeddings. This further validates the proposed method’s awareness of photographic style, and shows that the LUT is enlightened to transfer it.

5 Conclusion

In this paper, we present PETAL, Photographic Embedding for Transfer with an Adaptive LUT. We propose to extract a photographic embedding to reflect the photographic style, optimized with self-supervised contrastive learning. We apply the stop-gradient operation [16] and hard negative examples generated through photographic style augmentation to stabilize the optimization. This tailored image representation makes it possible to capture continuous styles, and conditions a simple neural LUT to transfer the photographic style without any test-time optimization.

Limitations. Although the photographic embedding considers spatial information, object-level style, such as color adjustment applied to faces, is not addressed in this work, as it requires more domain-specific data (*e.g.*, portraits in various styles). We empirically find that PETAL remains stable on the portrait-focused PPR10K benchmark, but the most challenging cases typically involve skin-tone white balance and local, face-specific retouching, which we leave for future work. In addition, photographic style sometimes involves artistic expression beyond the adjustment of color or lightness, such as adding noise or grain, calling for future work to learn such unique styles.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No. 62136001, 62276007) and Beijing Major Science and Technology Project (Grant No. Z251100008125009).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023) [3](#)
2. Adobe Inc.: Adobe lightroom classic. <https://www.adobe.com/products/photoshop-lightroom-classic.html> (2024), version 2024 [9](#)
3. Afifi, M., Brown, M.S.: Sensor-independent illumination estimation for dnn models. In: Proc. of British Machine Vision Conference (2019) [5](#)
4. Afifi, M., Brubaker, M.A., Brown, M.S.: HistoGAN: Controlling colors of gan-generated and real images via color histograms. In: Proc. of Computer Vision and Pattern Recognition (2021) [5](#), [7](#)
5. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Proc. of European Conference on Computer Vision. Springer (2004) [3](#)
6. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval. In: Proc. of International Conference on Learning Representations (2023) [10](#)
7. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proc. of Computer Vision and Pattern Recognition (2018) [3](#)
8. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) [3](#)
9. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025) [9](#)
10. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021) [4](#)
11. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: Proc. of Computer Vision and Pattern Recognition (2011) [9](#), [10](#), [11](#), [12](#), [13](#), [15](#)
12. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Proc. of Advances in Neural Information Processing Systems (2020) [7](#)
13. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. of International Conference on Computer Vision (2021) [10](#)
14. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) [3](#)
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. of International Conference on Machine Learning (2020) [4](#)
16. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proc. of Computer Vision and Pattern Recognition (2021) [3](#), [4](#), [6](#), [15](#)
17. Chesser, L., Carbone, T., Zahid, A.: Unsplash full dataset 1.2.2 (2020), unsplash.com/data, Last accessed on 2024-10-31 [8](#)

18. Chiu, T.Y., Gurari, D.: PhotoWCT²: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In: Proc. of Winter Conference on Applications of Computer Vision (2022) [2](#), [9](#), [11](#), [14](#)
19. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proc. of Computer Vision and Pattern Recognition (2005) [4](#)
20. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of Computer Vision and Pattern Recognition (2005) [3](#)
21. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: StyTr²: Image style transfer with transformers. In: Proc. of Computer Vision and Pattern Recognition (2022) [9](#), [11](#), [12](#), [14](#)
22. Duan, Z.P., Zhang, J., Lin, Z., Jin, X., Wang, X., Zou, D., Guo, C., Li, C.: DiffRetouch: Using diffusion to retouch on the shoulder of experts. In: Proc. of Association for the Advancement of Artificial Intelligence (2025) [4](#)
23. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: Improving visual-semantic embeddings with hard negatives. In: Proc. of British Machine Vision Conference (2018) [3](#)
24. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: DreamSim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023) [9](#), [10](#)
25. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. of Computer Vision and Pattern Recognition (2016) [1](#), [9](#), [10](#)
26. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (2017) [4](#)
27. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: ImageBind: One embedding space to bind them all. In: Proc. of Computer Vision and Pattern Recognition (2023) [9](#), [10](#)
28. Gong, Z., Wu, Z., Tao, Q., Li, Q., Loy, C.C.: SA-LUT: Spatial adaptive 4D look-up table for photorealistic style transfer. In: Proc. of International Conference on Computer Vision (2025) [2](#), [3](#), [9](#), [10](#), [11](#), [12](#), [13](#), [15](#)
29. Google Inc.: Nano Banana Pro. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf> (2025), model card [10](#), [11](#), [12](#)
30. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: Proc. of Advances in Neural Information Processing Systems (2020) [3](#), [4](#)
31. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proc. of International Conference on Artificial Intelligence and Statistics (2010) [4](#)
32. Hasselblad, V.: Hasselblad masters competition (2025), <https://www.hasselblad.com/inspiration/masters/2023/>, Last accessed on 2025-05-13 [9](#)
33. He, J., Wang, Z., Wang, L., Liu, T.I., Fang, Y., Sun, Q., Ma, K.: Multiscale sliced wasserstein distances as perceptual color difference measures. In: Proc. of European Conference on Computer Vision. pp. 425–442. Springer (2024) [10](#)
34. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proc. of Computer Vision and Pattern Recognition (2020) [4](#)

35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of Computer Vision and Pattern Recognition (2016) [3](#)
36. Ho, M.M., Zhou, J.: Deep Preset: Blending and retouching photos with color style transfer. In: Proc. of Winter Conference on Applications of Computer Vision (2021) [4](#), [9](#), [11](#), [12](#), [13](#)
37. Hong, K., Jeon, S., Yang, H., Fu, J., Byun, H.: Domain-aware universal style transfer. In: Proc. of International Conference on Computer Vision (2021) [2](#), [9](#), [11](#)
38. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. ACM Transactions on Graphics (2018) [4](#)
39. Hu, Z., Jia, J., Liu, B., Bu, Y., Fu, J.: Aesthetic-aware image style transfer. In: Proc. of ACM International Conference on Multimedia (2020) [5](#)
40. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proc. of International Conference on Computer Vision (2017) [2](#)
41. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1) (2010) [10](#)
42. Johnson, G.M., Fairchild, M.D.: A top down description of S-CIELAB and CIEDE2000. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur **28**(6), 425–435 (2003) [5](#)
43. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. In: Proc. of Advances in Neural Information Processing Systems (2020) [3](#)
44. Ke, Z., Liu, Y., Zhu, L., Zhao, N., Lau, R.W.: Neural preset for color style transfer. In: Proc. of Computer Vision and Pattern Recognition (2023) [2](#), [4](#), [9](#), [10](#), [11](#), [12](#), [13](#)
45. Kendall, M.G.: A new measure of rank correlation. Biometrika (1938) [10](#), [11](#)
46. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc. of Advances in Neural Information Processing Systems. vol. 25 (2012) [3](#)
47. Larchenko, M., Lobashev, A., Guskov, D., Palyulin, V.V.: Color transfer with modulated flows. In: Proc. of Association for the Advancement of Artificial Intelligence (2025) [2](#), [9](#), [10](#), [11](#), [14](#)
48. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proc. of European Conference on Computer Vision (2018) [3](#)
49. Li, H., Jia, C., Jin, P., Cheng, Z., Li, K., Sui, J., Liu, C., Yuan, L.: FreestyleRet: Retrieving images from style-diversified queries. In: Proc. of European Conference on Computer Vision (2024) [2](#), [9](#), [10](#), [11](#)
50. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. of International Conference on Machine Learning (2023) [2](#), [9](#), [10](#)
51. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. of International Conference on Machine Learning (2022) [2](#), [9](#), [10](#)
52. Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pre-training for

- vision-language tasks. Proc. of European Conference on Computer Vision (2020) [3](#)
53. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photo-realistic image stylization. In: Proc. of European Conference on Computer Vision (2018) [2](#)
54. Liang, J., Zeng, H., Cui, M., Xie, X., Zhang, L.: PPR10K: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In: Proc. of Computer Vision and Pattern Recognition (2021) [9](#), [10](#), [11](#), [12](#), [15](#)
55. Liu, C., Yang, H., Fu, J., Qian, X.: 4D LUT: learnable context-aware 4d lookup table for image enhancement. IEEE Transactions on Image Processing (2023) [2](#), [3](#), [4](#)
56. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. of International Conference on Computer Vision (1999) [3](#)
57. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proc. of Computer Vision and Pattern Recognition (2017) [2](#), [9](#)
58. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9** (2008) [12](#), [13](#)
59. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-RNN). Proc. of International Conference on Learning Representations (2015) [3](#)
60. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proc. of Computer Vision and Pattern Recognition (2016) [4](#), [10](#)
61. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. arXiv:2304.07193 (2023) [9](#), [10](#), [13](#)
62. Ouyang, W., Dong, Y., Kang, X., Ren, P., Xu, X., Xie, X.: RSFNet: A white-box image retouching approach using region-specific color filters. In: Proc. of International Conference on Computer Vision (2023) [4](#)
63. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. of Computer Vision and Pattern Recognition (2007) [10](#)
64. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proc. of Computer Vision and Pattern Recognition (2008) [10](#)
65. Pitié, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Proc. of International Conference on Computer Vision (2005) [2](#)
66. Pitié, F., Kokaram, A.C., Dahyot, R.: Automated colour grading using colour distribution transfer. Computer Vision and Image Understanding **107** (2007) [2](#)
67. Pixelmator Team Ltd.: Pixelmator Pro. <https://support.pixelmator.com/pixelmator-pro-user-guide/automatically-edit-images/automatically-match-image-colors> (2025), color match AI tool [9](#), [11](#)
68. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) [3](#)

69. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proc. of Computer Vision and Pattern Recognition (2018) [10](#)
70. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. of International Conference on Machine Learning (2021) [2](#), [3](#), [6](#), [9](#), [10](#)
71. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21** (2001) [2](#)
72. Risheng, L., Long, M., Jiaao, Z., Xin, F., Zhongxuan, L.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proc. of Computer Vision and Pattern Recognition (2021) [4](#)
73. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020) [3](#)
74. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of Computer Vision and Pattern Recognition (2022) [3](#)
75. Ruta, D., Motiian, S., Faieta, B., Lin, Z., Jin, H., Filipkowski, A., Gilbert, A., Collomosse, J.: Aladin: All layer adaptive instance normalization for fine-grained style similarity. In: Proc. of International Conference on Computer Vision (2021) [9](#), [10](#)
76. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: Proc. of Computer Vision and Pattern Recognition (2015) [4](#)
77. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) [3](#)
78. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv:2508.10104 (2025) [9](#), [10](#), [13](#)
79. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. In: Proc. of Advances in Neural Information Processing Systems (2016) [4](#)
80. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Investigating style similarity in diffusion models. In: Proc. of European Conference on Computer Vision (2024) [2](#), [9](#), [10](#), [11](#)
81. Song, Y., Liu, X., Shou, M.Z.: DiffSim: Taming diffusion models for evaluating visual similarity. In: Proc. of International Conference on Computer Vision (2025) [9](#), [10](#)
82. Song, Y., Qian, H., Du, X.: StarEnhancer: Learning real-time and style-aware image enhancement. In: Proc. of International Conference on Computer Vision (2021) [4](#)
83. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology (1904) [10](#), [11](#)
84. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proc. of Computer Vision and Pattern Recognition (2015) [3](#)
85. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) [3](#), [10](#)

86. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) [3](#)
87. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) [3](#)
88. Tschanen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) [6](#), [7](#), [9](#), [10](#)
89. Tseng, E., Zhang, Y., Jebe, L., Zhang, X., Xia, Z., Fan, Y., Heide, F., Chen, J.: Neural photo-finishing. ACM Transactions on Graphics (2022) [4](#)
90. Vera, H.S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S.R., Smoot, S., Naim, I., Zou, J., Chen, F., et al.: Embeddinggemma: Powerful and lightweight text representations. arXiv preprint arXiv:2509.20354 (2025) [9](#)
91. Wang, H., Zhang, J., Liu, M., Wu, X., Zuo, W.: Learning diverse tone styles for image retouching. IEEE Transactions on Image Processing (2023) [4](#)
92. Wang, W., Zhou, W., Bao, J., Chen, D., Li, H.: Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In: Proc. of International Conference on Computer Vision (2021) [3](#)
93. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) [3](#)
94. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: LAVIE: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision (2024) [3](#)
95. Wang, Y., Li, X., Xu, K., He, D., Zhang, Q., Li, F., Ding, E.: Neural color operators for sequential image retouching. In: Proc. of European Conference on Computer Vision (2022) [4](#)
96. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4) (2004) [10](#)
97. Wen, L., Gao, C., Zou, C.: CAP-VSTNet: Content affinity preserved versatile style transfer. In: Proc. of Computer Vision and Pattern Recognition (2023) [9](#), [11](#), [14](#)
98. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In: Proc. of International Conference on Machine Learning. pp. 54015–54029 (2024) [10](#)
99. Yang, S., Huang, B., Cao, M., Ji, Y., Guo, H., Wong, N., Yang, Y.: Taming lookup tables for efficient image retouching. In: Proc. of European Conference on Computer Vision (2024) [2](#), [3](#), [4](#)
100. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) [3](#)
101. Yao, M., Huang, J., Jin, X., Xu, R., Zhou, S., Zhou, M., Xiong, Z.: Generalized lightness adaptation with channel selective normalization. In: Proc. of International Conference on Computer Vision (2023) [4](#)

102. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proc. of International Conference on Computer Vision (2019) [2](#)
103. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) [2](#), [3](#), [4](#)
104. Zhang, F., Zeng, H., Zhang, T., Zhang, L.: CLUT-Net: Learning adaptively compressed representations of 3DLUTs for lightweight image enhancement. In: Proc. of ACM International Conference on Multimedia (2022) [2](#), [3](#), [4](#)
105. Zhang, F., Zhang, L., Zhang, T., Wang, D.: Adaptively hashing 3DLUTs for lightweight real-time image enhancement. In: Proc. of International Conference on Multimedia and Expo (2023) [2](#), [3](#), [4](#)
106. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Making visual representations matter in vision-language models. In: Proc. of Computer Vision and Pattern Recognition (2021) [3](#)
107. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of Computer Vision and Pattern Recognition (2018) [8](#), [10](#), [12](#)
108. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proc. of Computer Vision and Pattern Recognition (2023) [10](#)

Enlightening Photographic Style Transfer with a Self-Supervised Photographic Embedding

Supplementary Material

Chengxuan Zhu¹, Jiacong Fang^{1,2}, Shuchen Weng^{1*}, Youwei Lyu²,
Jiajun Tang², Qingnan Fan², Chao Xu¹, and Boxin Shi^{1*}

¹ Peking University ² vivo BlueImage Lab
{peterzhu, shuchenweng, shiboxin}@pku.edu.cn, jiacong_fang@stu.pku.edu.cn

A	More Demonstrations and Comparisons	1
A.1	Visualization of Photographic Embedding Space	1
A.2	More Qualitative Comparisons	2
A.3	Additional Perceptual and VLM Evaluation	5
A.4	Ablation Study	5
B	Photographic Style Augmentation Detail	8
C	Mathematical Definition for Evaluation Metrics	8
D	Training and Inference Details	11
E	Dataset Details	11
F	Failure Cases	12
G	Discussion	14
G.1	Scope of Photographic Style	14
G.2	Potential Positive Societal Impacts	14
G.3	Potential Negative Societal Impacts	14
G.4	Relationship with Professional Softwares	14
G.5	Future Work	15

A More Demonstrations and Comparisons

A.1 Visualization of Photographic Embedding Space

We visualize the embedding space of the proposed photographic embedding and other baselines on the PPR10K [18] subset dataset, as shown in Fig. 8. We denote the works of each expert as green triangles, blue squares, and red circles respectively. The 2D visualization is performed by first applying principal component analysis (PCA) to reduce the features to 15 dimensions, and then using t-SNE [20] to reduce the dimensions to 2 for visualization. As the photographic styles within each expert’s works are similar, while different between different

Equal contribution. * Corresponding authors.

¹ Chengxuan Zhu and Chao Xu are with National Key Lab of General AI, School of Intelligence Science and Technology, Peking University. Jiacong Fang, Shuchen Weng, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

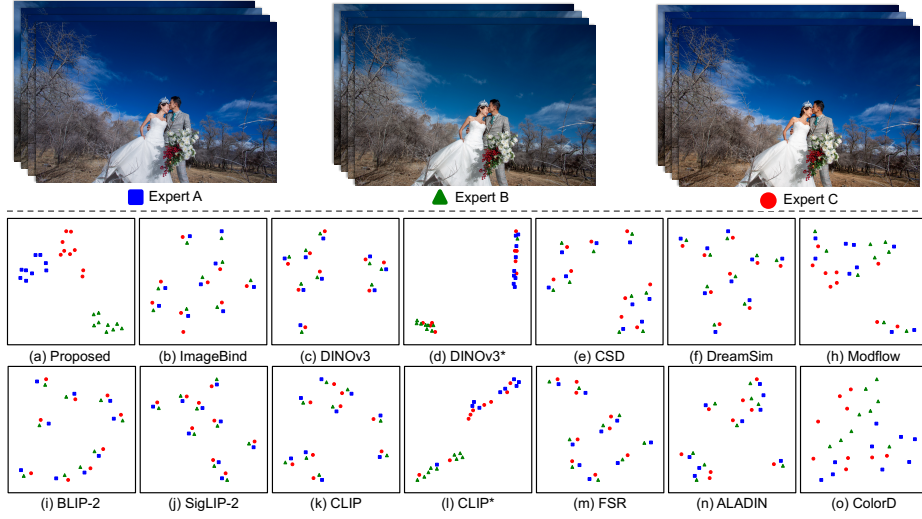


Fig. 8: A t-SNE [20] visualization of the proposed photographic embedding and baseline methods, conducted on the PPR10K [18] subset dataset. The images are represented by the experts who retouched them, as denoted by the blue squares for Expert A, green triangles for Expert B, and red circles for Expert C.

experts, we can see that the proposed photographic embedding is able to cluster the images more effectively than the baselines. Semantic embeddings (*e.g.*, CLIP [22] and DINO [24]) primarily capture content similarity, and existing style embeddings (*e.g.*, CSD [25] and Modflow [16]) still struggle to separate such photographic styles. Notably, the specially trained color discriminator ColorD [13] also fails to clearly separate the fine-grained style differences.

A.2 More Qualitative Comparisons

As shown in Fig. 9, we first perform spherical interpolation from a monochrome style with moderate exposure and contrast (Reference A) to another three difference target styles (References B, C, and D). By utilizing these interpolated photographic embeddings, PETAL smoothly blends the image from the source style to the target styles, while consistently preserving the underlying content structure and details throughout the process.

As illustrated in Fig. 10, we further present more qualitative examples of the photographic style transfer on test images from open-source DPST [19] and PST50 [7] datasets, as well as our collected INSRANK and Hasselblad Masters datasets [9].

Overall, PETAL can transfer a wide range of photographic styles from the references to content images across diverse scenes (*e.g.*, portrait, landscape, street, and nature), while preserving the underlying content structure. In contrast, several open-source baselines fail in characteristic ways. StyTr² [6], CAP-



Fig. 9: Interpolation transfer results of photographic style between the Reference A (monochrome with moderate exposure and contrast, etc.) and three reference styles: Reference B (high saturation, teal and orange tone, bright exposure, etc.), Reference C (monochrome with high contrast, etc.), and Reference D (low contrast and exposure, dim highlights, cold tint, etc.). The content is kept throughout the process.

VSTNet [31], PhotoWCT² [5], and DSTN [12] often introduce unnatural artifacts or textures in the stylized results. Deep Preset [11] frequently fails to faithfully apply the reference style. SALUT [7] and Modflow [16] either produce unwanted distortions (*e.g.*, the sky in the third column) or fail to capture key stylistic cues from the reference (*e.g.*, the light-blue sky and the yellow-green grass in the first column).

We further compare against three closed-source applications. Neural Preset [13] tends to apply global tone and color adjustments from the reference, but fails to produce a more dedicated, reference-faithful retouch. For example, it only partially desaturates the portrait case where the reference is monochrome (second column), and in the landscape example (first column) it shifts the scene toward a cooler, darker rendering, missing the vivid high-saturation spring palette of the reference. The commercial Pixelmator Pro [21] shows a similar trend. While it captures coarse attributes, it often misses fine-grained stylistic signals, such as highly saturated tones (first column) and local contrast distribution (fifth column). Finally, Nano Banana Pro [8] (prompt in Fig. 17) can produce visually strong edits but sometimes sacrifices content fidelity and fine-grained style adherence, introducing distortions or hallucinated details, such as the scene change in the fourth column and unnatural lighting in the last column.



Fig. 10: More qualitative comparisons of PETAL and the baselines on test images from open-sourced DPST [19] and PST50 [7], as well as our collected INS RANK and Hasselblad Masters datasets [9].

Table 4: Additional perceptual and VLM-as-a-judge evaluation on PST50 [7]. LIQE [34] and Q-Align [32] measure reference-free perceptual quality, MS-SWD [10] measures tonal-style distribution similarity, and Gemini-3.1-Pro [27] judges content preservation (CP), style similarity (SS), and overall quality.

Method	LIQE↑	Q-Align↑	MS-SWD↓	CP↑	SS↑	Overall↑
CAP-VSTNet [31]	<u>2.4362</u>	3.6574	5.6058	4.90	<u>4.58</u>	4.40
StyTr ² [6]	1.0429	3.3114	5.4223	3.78	4.10	3.08
Deep Preset [11]	2.3577	3.6067	4.6861	<u>4.98</u>	2.86	3.16
SALUT [7]	2.0362	3.0940	<u>4.2559</u>	4.64	3.74	3.54
DSTN [12]	1.6471	2.6037	7.8048	2.08	2.52	1.12
PhotoWCT ² [5]	2.2718	3.3488	5.9969	4.82	4.42	4.16
Modflow [16]	2.3195	3.4032	5.6159	5.00	4.52	<u>4.52</u>
Neural Preset [13]	2.2851	<u>3.7617</u>	4.4059	4.92	4.28	4.48
Nano Banana Pro [8]	1.9676	3.7097	5.2761	4.08	3.08	3.10
PETAL	2.5772	3.7813	3.8384	5.00	4.74	4.80

A.3 Additional Perceptual and VLM Evaluation

We use LIQE [34] and Q-Align [32] as reference-free perceptual quality metrics, and MS-SWD [10] as a perceptual color-distribution distance between the generated result and the reference style. For the VLM-as-a-judge evaluation, each source/content image, reference style image, and stylized output is fed to Gemini-3.1-Pro [27] with the prompt shown in Fig. 17; Gemini returns 1–5 JSON scores, from which we report content preservation (CP), style similarity (SS, from tone/style fidelity), and overall quality. As shown in Tab. 4, PETAL ranks best on LIQE, Q-Align, MS-SWD, style similarity, and overall quality, and ties for the best content preservation score. These results support the same trend as the reference-based metrics and the user study in the main paper.

A.4 Ablation Study

We first verify the proposed photographic embedding network design and the self-supervision, and the complete retrieval performance on PPR10K [18] subset, FiveK-Concept [3] and INSRANK are shown in Tab. 5. Changing the color space from CIE-Lab to RGB worsens the representation in both evaluation dataset, as the former disentangles luminance and chromaticity and allows for an easier extraction of photographic style. Without the cross attention module, the information from histogram encoders is also removed, and judging from the performance, the global statistical information is still necessary, and cannot be directly learned with a single ViT. Without the cross-attention module, the information from the histogram encoders is also removed. The retrieval performance drop, especially on PPR10K subset and FiveK-concept, shows that such global statistical information remains necessary and cannot be directly learned by a ViT alone. The adaptive pooling combines the important features in an image, while

Table 5: An ablation study of photographic embeddings on retrieval performance. All metrics are shown as percentages. For the proposed INS RANK dataset, Spearman’s ρ [26] and Kendall’s τ [14] measure the correlation between model rankings and photographers’ preferences. In this table, higher numbers indicate better performance. Best values are in **bold**, and second best are underlined.

Setting	PPR10K [18] subset				FiveK-Concept [3]				INS RANK		
	R@1	R@5	mAP	F1@5	R@1	R@5	mAP	F1@5	R@1	ρ [26]	τ [14]
RGB color space	70.34	96.77	69.16	53.61	51.21	78.29	<u>49.37</u>	36.58	37.50	17.50	14.37
W/o cross attn	68.93	96.66	66.44	51.21	50.29	75.91	48.30	36.19	50.00	30.00	23.75
[CLS] only	<u>73.74</u>	96.88	<u>70.01</u>	<u>54.36</u>	49.17	<u>80.14</u>	48.64	35.46	43.75	29.69	23.47
[CLS] + mean pool	68.91	<u>97.25</u>	66.53	51.25	50.48	79.24	47.61	34.78	62.50	36.87	30.62
W/o $\mathcal{L}_{rec}$	66.98	96.32	66.06	51.13	45.86	78.14	47.58	34.04	<u>53.12</u>	<u>40.00</u>	<u>35.62</u>
W/o $\mathcal{L}_-$	36.09	89.79	45.42	33.38	18.64	51.67	21.62	14.56	43.75	13.75	13.12
W/o $\mathcal{L}_+$	67.53	96.75	65.03	49.79	<u>51.79</u>	81.51	49.36	<u>36.69</u>	<u>53.12</u>	26.87	21.88
Complete	74.27	98.40	71.99	56.60	52.94	78.56	51.84	39.01	62.50	52.19	46.88

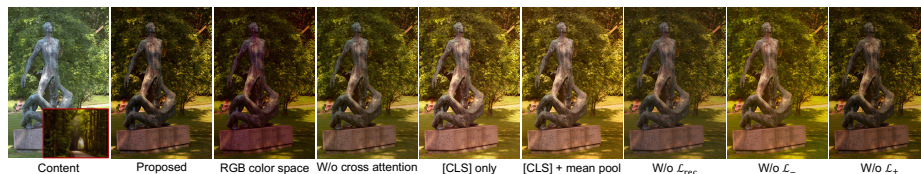


Fig. 11: A qualitative comparison between the proposed photographic embeddings and ablated settings in Tab. 5. The reference is shown in the bottom-right red box.

removing it or substituting it with mean pooling deteriorates the performance on PPR10K, FiveK-Concept and INS RANK. As for the supervision, we found the intra-sample loss the most important, and a combination of the three loss terms have the best overall performance.

As illustrated in Fig. 11, we further present a qualitative comparison of the proposed photographic embedding and the ablated settings mentioned above. PETAL inherits the warm highlight and high contrast, and renders the desired color and shadows on the trees consistent with the reference. The embedding in RGB color space unable to extract the luminance and chromaticity features, producing an unnatural purple tone on the statue and failing to transfer the warm highlight and high contrast. Without the cross-attention module, the model cannot reproduce the high contrast and the desired color and shadows on the trees, indicating the necessity of the global statistical information. Removing the adaptive pooling, or substituting it with mean pooling, also degrades the visual quality, tending to render the content image to global high exposure. A combination of the three loss terms yields the best qualitative results, while removing any single term leads to noticeable degradation, especially in the shadows, tree colors, and the yellow sunlit highlights.

We then replace the photographic embedding with each previous image embedding and train the photographic style transfer network using the same loss

Table 6: Quantitative comparisons for photographic style transfer on the PPR10K [18] subset and PST50 [7] dataset. $\uparrow$ ($\downarrow$) denote higher (lower) is better. All the methods are trained with the same lightweight photographic style transfer network and loss functions as PETAL. The superscript * indicates that the baseline embeddings are fine-tuned with the proposed self-supervised framework, data and objectives first.

Method	PPR10K [18] subset				PST50 [7]			
	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	H-Corr $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	H-Corr $\uparrow$
CSD [25]	0.1131	20.45	0.8544	0.3237	0.1353	21.36	0.8848	0.3786
CLIP [22]	0.1369	19.65	0.8396	0.3436	0.1712	20.52	0.8493	0.3172
DINOv3 [24]	0.3240	14.84	0.6662	0.1527	0.3150	17.26	0.7557	0.1893
BLIP [17]	0.3200	15.36	0.7450	0.1787	0.3390	16.81	0.7827	0.1814
SigLIP-2 [28]	0.1273	20.39	0.8537	0.3737	0.1702	20.87	0.8672	0.3239
DINOv3* [24]	<u>0.1101</u>	<u>21.57</u>	0.8743	<u>0.3948</u>	0.1167	22.39	0.8988	0.3757
CLIP* [22]	0.1215	20.81	0.8491	0.3793	<u>0.1122</u>	<u>23.70</u>	<u>0.9191</u>	<u>0.4719</u>
PETAL	0.1027	21.80	<u>0.8710</u>	0.4578	0.1021	24.63	0.9272	0.5220



Fig. 12: A qualitative comparison of PETAL and the ablation that using the previous embeddings, with images sampled from the unpaired test set of PST50 [7].

functions as in PETAL. As shown in Tab. 6, all metrics, especially H-Corr, drop when using previous image embeddings, as these embeddings primarily encode compressed semantic features rather than detailed photographic style representation. We further present qualitative comparisons using images sampled from the unpaired test set of PST50 [7]. As illustrated in Fig. 12, using previous image embeddings leads to clear visual degradation. CSD [25], CLIP [22], and SigLIP-2 [28] fail to render the desired tone and contrast. The transfer network also performs poorly with BLIP [17] and DINOv3 [24], introducing visible artifacts or unnatural appearances. Notably, substituting the backbone with strong pretrained models (DINOv3* and CLIP*) and fine-tuning them with our self-supervised framework still yields incorrect photographic stylistic rendering, especially for the skin tones. This suggests that the proposed embedding network design is essential for effectively capturing photographic style.

B Photographic Style Augmentation Detail

For reproducibility, we implement the photographic style augmentation function $f(I, \{\xi_i\})$ as a composition of differentiable, lightweight photo-edit operators and an optional 3D color lookup table (LUT) transform implemented by trilinear interpolation (`grid_sample`). Given an input image $I \in [0, 1]^{3 \times H \times W}$, the augmentation samples a random subset of operators and applies them sequentially. Each operator is controlled by a scalar parameter $\xi \sim \mathcal{U}([-1, 1])$ (or an equivalent factor sampled from a conservative range), reaching its maximum strength when $|\xi| = 1$. We intentionally keep the ranges moderate to avoid out-of-distribution artifacts while still producing hard negatives in photographic style. For the contrastive views used in training, we do not treat arbitrary distant patches from one image as positives. An anchor crop is jittered by small spatial shifts to form overlapping local views under the same PSA parameters. These local views share most content, illumination context, and smooth tone/-color rendering, making them reliable positives within our scope. Conversely, two opposite PSA views of the same patch preserve content and layout but deliberately differ in photographic tone/color, which discourages content-only shortcuts and provides hard negatives for learning a style-sensitive embedding.

As shown in Tab. 7, we select these operators to align with the most common controls used by photographers in Adobe Lightroom [1]/Adobe Camera Raw: setting the luminance anchor (Exposure/Contrast), redistributing dynamic range (Highlights/Shadows), and shaping overall color mood (Saturation/Temperature/Tint). Vignette serves as a lightweight finishing adjustment to guide attention and alter atmosphere without changing scene semantics. Finally, an optional 3D LUT compactly models preset-like, coupled color remapping, enabling diverse yet realistic style variations while remaining differentiable. In practice, we build a LUT bank of 100 professional LUTs; when a LUT is used, we randomly sample one of four application strategies to increase diversity: single, pre-mix (blend LUTs before applying), post-mix (blend LUT outputs), and cascade (apply two LUTs sequentially with reduced intermediate strength).

We further visualize random samples of the proposed photographic style augmentation, where the leftmost column shows the original images from the Unsplash [4] training dataset. As shown in Fig. 13, our augmentation generates a wide spectrum of realistic style variants across diverse scenes (*e.g.*, portrait, architecture, landscape, and nature), while preserving the underlying content. Such rich and diverse style perturbations provide the essential training views for our self-supervised learning framework.

C Mathematical Definition for Evaluation Metrics

As we mention in the paper that the datasets of PPR10K [18] and FiveK-Concept [3] are divided into multiple scenes, to guarantee similar content during evaluation, we denote the number of scenes as $M_{\mathcal{U}}$, and the dataset as $\mathcal{U}$. For each scene $\mathcal{S} = \{I_j\}_{j=1}^{N_{\mathcal{S}}} \subset \mathcal{U}$ in the dataset which contains $N_{\mathcal{S}}$ images retouched

Table 7: Photographic style augmentation operators used to instantiate $f(I, \{\xi_i\})$. All operations are applied in CIE-Lab space with values normalized to $[0, 1]$ for L and to a standardized range for a, b .

Operator	Mapping (max strength at $ \xi = 1$)
Exposure	$L \leftarrow \text{clip}(L \cdot 2^{0.5\xi}) \ (\pm 0.5 \text{ EV})$
Contrast	$L \leftarrow \text{clip}((L - \bar{L})(1 + 0.30\xi) + \bar{L})$
Saturation	$(a, b) \leftarrow (a, b) \cdot (1 + 0.35\xi)$
Temperature	$a \leftarrow a + 8\xi$ (warm/cool shift)
Tint	$b \leftarrow b + 8\xi$ (green/magenta shift)
Highlights	$L \leftarrow L - 0.25\xi \cdot \max(L - 0.7, 0)^2 / (0.3^2)$
Shadows	$L \leftarrow L + 0.25\xi \cdot \max(0.3 - L, 0)^2 / (0.3^2)$
Vignette	$I \leftarrow I \cdot (1 - 0.35\xi \cdot r^2)$ where $r \in [0, 1]$ is the normalized radius
3D LUT	$I \leftarrow (1 - \xi)I + \xi \cdot \text{LUT}(I), \ \xi \sim \mathcal{U}([0.5, 1])$

by N_P photographers, we retrieve the top- k nearest neighbors of each query image based on cosine similarity. We then define the metrics used in the paper mathematically. Denote $\text{rel}(I, k)$ as an indicator function that equals 1 if the k -th retrieved image shares the same label as I , and 0 otherwise.

The R@ k metric is given by

$$\text{R@}k = \frac{1}{M_{\mathcal{U}}} \sum_{\mathcal{S}} \sum_{j \in \mathcal{S}} \frac{1}{N_{\mathcal{S}}} \bigvee_{i=1}^K \text{rel}(I_j, k), \quad (7)$$

where $\bigvee$ is the logical “or” function. Basically, $\bigvee_{i=1}^K \text{rel}(I_j, k)$ measures if any of the top- k neighbor matches the ground truth label of I_j .

The metric of mAP is the average of the area under the precision-recall curve [29], defined by

$$P(I_j, k) = \frac{1}{k} \sum_{i \leq k} \text{rel}(I_j, i), \quad (8)$$

$$\text{AP}(I_j) = \frac{\sum_{k=1}^{N_{\mathcal{S}}} P(k) \text{rel}(I_j, k)}{M(\mathcal{S}, I_j)}, \quad (9)$$

where $P(I_j, k)$ is the precision at k -th position in the retrieval list of I_j . $M(\mathcal{S}, I_j)$ is the total number of images in $\mathcal{S}$ retouched by the same photographer as I_j . The metric of mAP is the average $\text{AP}(\mathcal{S})$ over all the scenes in the dataset, namely

$$\text{AP}(\mathcal{S}) = \frac{1}{N_{\mathcal{S}}} \sum_{j=1}^{N_{\mathcal{S}}} \text{AP}(I_j), \text{mAP} = \frac{1}{M_{\mathcal{U}}} \sum_{\mathcal{S}} \text{AP}(\mathcal{S}). \quad (10)$$

F1@ k computes the harmonic mean of precision and recall at cutoff k , providing a balanced measure of a model’s retrieval performance.

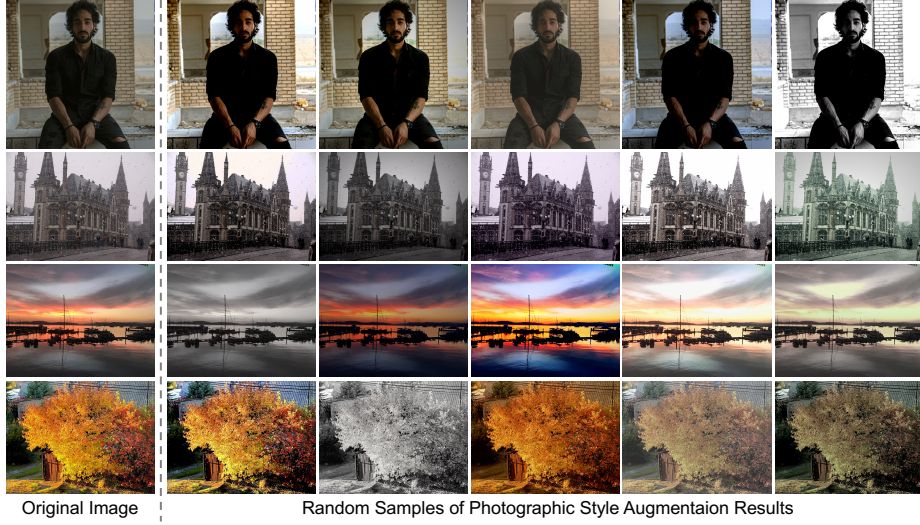


Fig. 13: Random samples of the proposed photographic style augmentation results. The original images are sourced from the Unsplash [4] training dataset.

As for the INSRANK dataset, we treat the evaluation as a ranking problem. Each query q_i is associated with $M = 5$ candidates, having a ground-truth ranking π_i^* and a predicted ranking $\hat{\pi}_i$. We report R@1 to measure the top-1 accuracy, along with Kendall’s τ [14] and Spearman’s ρ [26] for ranking correlation.

The R@1 metric here checks if the predicted top-1 image c_i^* matches the expert-annotated best:

$$\text{R@1} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{\pi}_i(c_i^*) = 1), \quad (11)$$

where N is the total number of queries and $\mathbb{I}(\cdot)$ is the indicator function. To evaluate global ranking consistency, Kendall’s τ [14] and Spearman’s ρ [26] are defined as

$$\tau = \frac{1}{N} \sum_{i=1}^N \frac{2(N_c^i - N_d^i)}{M(M-1)}, \quad (12)$$

$$\rho = \frac{1}{N} \sum_{i=1}^N \left(1 - \frac{6 \sum_{j=1}^M d_{i,j}^2}{M(M^2 - 1)} \right), \quad (13)$$

where N_c^i and N_d^i denote the number of concordant and discordant pairs between π_i^* and $\hat{\pi}_i$, respectively. $d_{i,j}$ represents the rank difference for the j -th candidate. Both metrics range from $[-1, 1]$, with higher values indicating better alignment with human photographers’ aesthetics.

D Training and Inference Details

The training of the proposed method is conducted on 2 NVIDIA RTX PRO 6000 GPUs. The profiling and inference are performed on a single NVIDIA RTX A6000. The running time of the photographic style transfer is evaluated by processing content images with a resolution of 1080×1920 , using reference images of size 512×512 as guidance.

To train the photographic style encoder, we download 43w high quality photograph from the Unsplash [4] dataset. We use a batch size of 12, where each batch contains 2 augmented versions of 4 patches per image. We use the Adam [15] optimizer with a learning rate of $2e-4$ and cosine annealing. It takes about 20 hours to train for 80k iterations. The Vision Transformer is composed of 24 layers of transformer blocks with 16 attention heads, an embedding dimension of 1024, and a patch size of 16. Following previous work [24], we utilize RMSNorm [33] for normalization and SwiGLU [23] for the FFN layers. The total number of parameters is 362M.

For the style transfer module, the photographic style encoder is fixed. We train for 80k iterations (approximately 18 hours) with a batch size of 12, using the Adam [15] optimizer with a learning rate of $4e-4$ and cosine annealing. The MLPs g_μ and g_σ , which map the photographic style embedding to the mean and variance, consist of two layers connected by Group Normalization. The total number of parameters is only 1.727M.

E Dataset Details

In the paper, we mentioned that the FiveK-Concept [3] dataset is divided into scenes first, and we describe the details here. We first categorize images from the original FiveK [3] dataset into subsets based on the “Location” and “Subject” metadata. Here we exclude images with missing metadata. Next, we employ Qwen3-VL-32B-Instruct [2] to annotate each image with detailed descriptions of the main subjects and scene content (prompt in Fig. 17). Then, we perform K-means clustering within each subset using the annotations encoded by the EmbeddingGemma [30] text embedding model, and finally we get 623 groups averaging 7 images each. A professional photographer edits each group’s RAW files in Adobe Lightroom [1], varying exactly one concept (*e.g.*, contrast, exposure, temperature, or tint) for a single style while maintaining within-group consistency.

As for the INS RANK dataset, we collect photos posted by photographers on Instagram, and curate them into 32 groups. Each group contains one query image

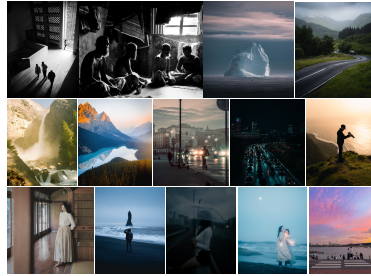


Fig. 14: Sample images from the proposed INS RANK dataset.



Fig. 15: A double-exposure failure case of photographic style transfer. PETAL transfers the warm amber tone and candlelight contrast, but cannot reproduce the semi-transparent floral/portrait overlay. Reference taken by Annie Spratt [4]

and five candidate images. Two photographers are invited to rank the candidates independently according to style similarity to the query. Their preferences are used as pseudo ground truth to validate the alignment with human preferences. The 32 query images are divided into three categories: natural landscapes (50%), cityscape/architecture (25%), and human/cultural subjects (25%). We selected these categories as they represent common scenarios where photographers typically employ post-processing to establish distinct styles. Conversely, we excluded genres like documentary and still-life, where the visual aesthetic is intrinsically tied to the subject content itself rather than artistic retouching. And some sample images are shown in Fig. 14.

F Failure Cases

As mentioned in the main-paper limitations, PETAL focuses on retouching-level tone/color transfer and smooth spatial effects, while local object- or face-specific retouching and artistic effects beyond color or lightness adjustment remain challenging. We demonstrate these two out-of-scope cases below: the double-exposure reference in Fig. 15 and the face-specific local retouching example in Fig. 16.

As shown in Fig. 15, the reference image is composed of highly saturated orange color, which the proposed method manages to transfer to the input image. More specifically, PETAL preserves the two people and candle highlights while moving the result toward the reference’s warm amber illumination. However, it treats the reference mainly as a tone and lighting cue, so the transparent floral

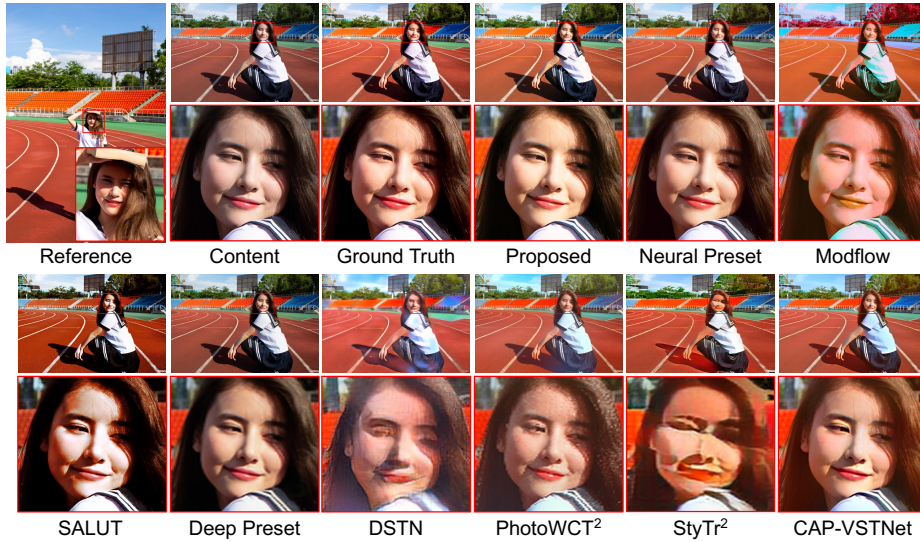


Fig. 16: A face-specific local-retouching failure case on PPR10K [18]. PETAL transfers the bright exposure, warm tone, and saturated red track from the reference, but the zoomed face region still shows imperfect local skin-tone retouching.

texture, ghosted face silhouette, and layered exposure boundary that define the double-exposure effect are absent. Other methods greatly brighten the input and cause over-exposed highlights, transfer incorrect colors, or introduce unnatural artifacts. For example, Neural Preset [13] and CAP-VSTNet [31] tend to amplify the candle region, SALUT [7] and Deep Preset [11] under-transfer the reference appearance, while DSTN [12], PhotoWCT² [5], and StyTr² [6] introduce stronger blur, color bleeding, or texture artifacts. However, none of the baselines, nor the proposed method, is able to transfer the unique double exposure effect.

We further show a portrait example from PPR10K [18] in Fig. 16. PETAL better follows the global photographic style than most baselines, especially in the bright outdoor exposure, warm tone, and red-track saturation. However, the face crop reveals a remaining limitation: local skin-tone retouching is not fully matched to the ground truth, with a slight yellow cast and imperfect face-specific adjustment. This distinction explains why PPR10K [18] remains a valid benchmark for our scope: its labels mainly evaluate whole-image global and smooth regional retouching, while identity-aware facial beautification and fine local skin edits require region-specific or face-specific supervision beyond the current PETAL design.

G Discussion

G.1 Scope of Photographic Style

In our setting, photographic style means retouching-level tone/color rendering and smooth spatial effects while preserving scene content, geometry, and texture. It covers exposure, contrast, white balance, saturation, tint, highlight/shadow redistribution, preset-like color remapping, vignetting, and spatial tone/color falloff. It does not target object-level semantic edits, identity-specific facial retouching, geometric deformation, artistic textures, film grain, double exposure, or painterly effects. This scope explains why PPR10K [18] is a valid portrait-focused benchmark for global and smooth regional retouching, while face-specific local skin retouching remains a limitation.

G.2 Potential Positive Societal Impacts

The proposed method can be used to enhance the visual quality of photos taken by amateur photographers, making them look more professional. This can help to promote the work of these users and increase their visibility. It also helps photographers learn to master the photographic style from professional photographers' works, by developing a before-after pair with their own images, so that they can learn how professional photographers retouch on their photos.

If the users are the photographers of the reference images, the proposed method can help them retouch their photos more efficiently. For the journalists reporting live events, the critical delay in retouching can be reduced, allowing them to publish their work more quickly. Since the proposed method can run real-time, it may even be possible to serve as a preset in the camera, allowing the users to create customized presets for their own unique photographic style.

G.3 Potential Negative Societal Impacts

From the perspective of the users, the proposed method may indulge them in the use of automatic tools to retouch their photos, which may lead to a lack of understanding of the photographic style and techniques, hindering their learning process.

For the photographers of the reference images, the copyright may be infringed if the generated images are used for commercial purposes without their permission. However, since there is no regulation on this issue, it is difficult to enforce copyright laws in this case. Were we to release the model, we would notify the user to respect the copyright of the photographer, and that the proposed method is designed for non-commercial use only.

G.4 Relationship with Professional Softwares

It is possible to achieve a similar result in Adobe Lightroom [1] or Adobe Photoshop, but to retouch the images with a reference, or even by the photographers

themselves, is a time-consuming task. Usually the process takes several minutes, while the proposed method is real-time. If integrated as a built-in preset in cameras, the proposed method can be used to render the previews for photographers, allowing them to see the world with their own photographic style, fostering their creativity.

G.5 Future Work

1. Raw file support. The proposed method is designed to work with JPEG images, but it can easily be extended to support raw files which holds more color information and allow for more flexibility in retouching.
2. Whiteness inpainting. As the input image may contain over-exposed area, to transfer a photographic style with less contrast or less exposure, the information in the over-exposed area is desired. Generative models can be used to inpaint the over-exposed area, but this is out of the scope of this work, due to the complicated nature of the task.
3. Photographic style transfer for videos. While it is easy to transfer the photographic style to a video instead of an image, to transfer from one video to another is a more complicated task. Extending the proposed method to work with videos is a future work.

Prompt for Qwen3-VL You are given an image. Generate ONE English sentence caption that begins with "A photo of ...". Content rules: - Describe the main photographic subjects that are visually prominent and compositionally important (e.g., person, flower, mountain, animal, building). - Describe each subject with specific, visually grounded adjectives (e.g., colors, shapes, states, or conditions). - Examples: "tall green tree", "small stone house", "bright red flower", "clear blue sky", "shiny glass building". - Include environmental context if visually clear (e.g., indoors, outdoors, in a garden, in the city). - Add a description of time of day (e.g., morning, afternoon, night, sunset) or lighting conditions (e.g., soft natural light, golden sunlight, dim light, dramatic shadows) if evident. - Focus on concise, scene-defining details that photographers would note. Ignore: - Small, cluttered, or repetitive objects. - Minor accessories or textures (e.g., hats, glasses, fences, bricks). - Actions, relations, abstract concepts, emotions, or atmosphere words not visually explicit. Output rules: - Must start with "A photo of ...". - Use concise noun phrases and natural descriptive adjectives. - Use commas or "with" to connect subjects, environment, and lighting. - Only one sentence. - If no clear subject is visible, output exactly: "A photo of nothing." Example outputs: - "A photo of white amaryllis flowers in bloom indoors with soft natural light." - "A photo of a mountain landscape under golden sunset light." - "A photo of a person standing in a city street at night with bright neon lights." - "A photo of a tall green pine tree beside a calm blue lake under morning sunlight." - "A photo of a small wooden cabin on a snowy mountain at sunset."
Prompt for Nano Banana Pro Preserve the content and detail of the first image, and transfer the photographic concept from the second image onto it, to get the result image with the same content as the first image, but with the photographic style of the second image.
Prompt for Gemini-3.1-Pro Evaluation You are an expert evaluator for photographic tone and color style transfer. You will be given three images: Image 1: the source/content image. Image 2: the reference tone/style image. Image 3: the edited/stylized output. Evaluate Image 3 by comparing it with both Image 1 and Image 2. Criteria: 1. Content Preservation: The main scene structure, objects, identity, geometry, and semantic content should be preserved from Image 1. Do not penalize expected color and tone changes. 2. Tone/Style Fidelity: The output should match the reference tone/style in Image 2, including exposure, contrast, brightness distribution, highlights, shadows, saturation, color temperature, color palette, and overall photographic mood. Do not reward copying objects or content from Image 2. 3. Photorealism and Naturalness: The output should look like a natural photograph, with plausible colors, lighting, contrast, and local details. 4. Artifact Severity: Check for unnatural color casts, over-saturation, clipping, banding, halos, texture damage, structural distortion, or other visible artifacts. Please give scores from 1 to 5: 1 = very poor 2 = poor 3 = acceptable 4 = good 5 = excellent Return the result in JSON format: <pre> { "content_preservation": { "score": 1-5, "reason": "brief explanation" }, "tone_style_fidelity": { "score": 1-5, "reason": "brief explanation" }, "photorealism": { "score": 1-5, "reason": "brief explanation" }, "artifact_control": { "score": 1-5, "reason": "brief explanation" }, "overall": { "score": 1-5, "reason": "brief explanation" } } </pre>

Fig. 17: Prompts used for Qwen3-VL [2] annotation, Nano Banana Pro [8] photographic style transfer, and Gemini-3.1-Pro-based VLM evaluation [27].

References

1. Adobe Inc.: Adobe lightroom classic. <https://www.adobe.com/products/photoshop-lightroom-classic.html> (2024), version 2024
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: Proc. of Computer Vision and Pattern Recognition (2011)
4. Chesser, L., Carbone, T., Zahid, A.: Unsplash full dataset 1.2.2 (2020), unsplash.com/data, Last accessed on 2024-10-31
5. Chiu, T.Y., Gurari, D.: PhotoWCT²: Compact autoencoder for photorealistic style transfer resulting from blockwise training and skip connections of high-frequency residuals. In: Proc. of Winter Conference on Applications of Computer Vision (2022)
6. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: StyTr²: Image style transfer with transformers. In: Proc. of Computer Vision and Pattern Recognition (2022)
7. Gong, Z., Wu, Z., Tao, Q., Li, Q., Loy, C.C.: SA-LUT: Spatial adaptive 4D look-up table for photorealistic style transfer. In: Proc. of International Conference on Computer Vision (2025)
8. Google Inc.: Nano Banana Pro. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf> (2025), model card
9. Hasselblad, V.: Hasselblad masters competition (2025), <https://www.hasselblad.com/inspiration/masters/2023/>, Last accessed on 2025-05-13
10. He, J., Wang, Z., Wang, L., Liu, T.I., Fang, Y., Sun, Q., Ma, K.: Multiscale sliced wasserstein distances as perceptual color difference measures. In: Proc. of European Conference on Computer Vision. pp. 425–442. Springer (2024)
11. Ho, M.M., Zhou, J.: Deep Preset: Blending and retouching photos with color style transfer. In: Proc. of Winter Conference on Applications of Computer Vision (2021)
12. Hong, K., Jeon, S., Yang, H., Fu, J., Byun, H.: Domain-aware universal style transfer. In: Proc. of International Conference on Computer Vision (2021)
13. Ke, Z., Liu, Y., Zhu, L., Zhao, N., Lau, R.W.: Neural preset for color style transfer. In: Proc. of Computer Vision and Pattern Recognition (2023)
14. Kendall, M.G.: A new measure of rank correlation. *Biometrika* (1938)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Larchenko, M., Lobashev, A., Guskov, D., Palyulin, V.V.: Color transfer with modulated flows. In: Proc. of Association for the Advancement of Artificial Intelligence (2025)
17. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. of International Conference on Machine Learning (2022)
18. Liang, J., Zeng, H., Cui, M., Xie, X., Zhang, L.: PPR10K: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In: Proc. of Computer Vision and Pattern Recognition (2021)
19. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proc. of Computer Vision and Pattern Recognition (2017)

20. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. *Journal of Machine Learning Research* **9** (2008)
21. Pixelmator Team Ltd.: Pixelmator Pro. <https://support.pixelmator.com/pixelmator-pro-user-guide/automatically-edit-images/automatically-match-image-colors> (2025), color match AI tool
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *Proc. of International Conference on Machine Learning* (2021)
23. Shazeer, N.: GLU variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020)
24. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. *arXiv:2508.10104* (2025)
25. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Investigating style similarity in diffusion models. In: *Proc. of European Conference on Computer Vision* (2024)
26. Spearman, C.: The proof and measurement of association between two things. *The American Journal of Psychology* (1904)
27. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023)
28. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmen, J., Steiner, A., Zhai, X.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025)
29. Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM (2006)
30. Vera, H.S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S.R., Smoot, S., Naim, I., Zou, J., Chen, F., et al.: Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354* (2025)
31. Wen, L., Gao, C., Zou, C.: CAP-VSTNet: Content affinity preserved versatile style transfer. In: *Proc. of Computer Vision and Pattern Recognition* (2023)
32. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-ALIGN: teaching LMMs for visual scoring via discrete text-defined levels. In: *Proc. of International Conference on Machine Learning*. pp. 54015–54029 (2024)
33. Zhang, B., Sennrich, R.: Root mean square layer normalization. In: *Proc. of Advances in Neural Information Processing Systems* (2019)
34. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: *Proc. of Computer Vision and Pattern Recognition* (2023)