

MirrorPPR: Exemplar-Based Portrait Photo Retouching

Zhihong Liu¹, Zheng Li¹, Jiachun Jin¹, Siqi Kou^{1,2}, Yitao Jian^{1,2},
Fengpei Yu², and Zhijie Deng¹^{*}

¹ Shanghai Jiao Tong University, China

² Triverse AI, China

{lzh15841259307,zhijied}@sjtu.edu.cn

Abstract. While text-guided image editing has made remarkable progress, it remains limited in structural portrait retouching. Textual descriptions struggle to convey fine-grained changes to facial features and body proportions. To address this gap, we introduce Exemplar-Based Portrait Photo Retouching, where the model is given an exemplar pair and tasked with inferring and applying the same retouching operations to a new query image. Existing exemplar-based editing methods primarily focus on tasks with pronounced visual transformations. In contrast, structural portrait retouching involves extremely delicate and localized modifications, making accurate extraction and transfer of these edits challenging. To tackle this, we propose MirrorPPR, a novel framework specifically designed to capture and transfer subtle structural retouching operations. Our method uses a Retouching Operation Extractor to capture the subtle differences from the exemplar pair. The extracted representations are then injected into a pre-trained Diffusion Transformer (DiT) through a connector and Low-Rank Adaptation (LoRA) modules. Furthermore, constructing perfectly aligned cross-identity training pairs is severely hindered by operation misalignment. To overcome this, we propose an advanced data self-augmentation paradigm that ensures strictly aligned retouching operations. To alleviate data scarcity and support this novel task, we introduce MirrorPPR47M, a large-scale dataset with over 47 million retouched pairs. By structuring the dataset into simulated and professional subsets, we enable progressive curriculum learning to smoothly optimize the network. Extensive experiments demonstrate that MirrorPPR significantly outperforms existing baselines in both retouching quality and identity preservation. The project page is available at <https://sjtu-deng-lab.github.io/MirrorPPR>.

Keywords: Portrait Photo Retouching · Exemplar-based Editing · Diffusion Models

1 Introduction

Text-guided image editing [5, 27, 35, 51, 55, 56] has achieved remarkable success. However, it reveals intrinsic limitations when applied to portrait photo retouch-

^{*} Corresponding author.



Fig. 1: Results of MirrorPPR on various samples. MirrorPPR successfully extracts fine-grained retouching operations from the exemplar and transfers them to the query. The operation types for each sample are: (a) top-left: “Shrink mouth”; (b) top-right: “Enlarge eyes, Plump lips, Narrow nasal alae”; (c) bottom-left: “Slim legs”; (d) bottom-right: “Narrow nasal alae, Square shoulders”.

ing. Portrait retouching encompasses appearance-level (e.g., color grading, skin smoothing) and structural-level adjustments (e.g., refining jawlines, enlarging eyes, or modifying body proportions). While appearance retouching can be efficiently handled by existing parametric filters, structural retouching remains highly challenging and relies heavily on manual expertise. Natural language is inherently ambiguous for automating such tasks, as it struggles to quantitatively specify the exact spatial scale, direction, and magnitude of fine-grained structural manipulations. Consequently, text-guided models frequently suffer from misinterpretations, leading to insufficient adjustments or unnatural over-editing.

To bypass the expressive limitations of text, exemplar-based image editing [11, 16, 28, 31, 36, 41, 49, 53, 58, 60, 62] has emerged as a promising alternative, allowing users to convey their intent intuitively via before-and-after image pairs. Yet, adapting this paradigm to structural portrait retouching introduces a unique challenge: unlike general image editing tasks characterized by substantial visual transformations, structural retouching operations are remarkably subtle. Consequently, existing exemplar-based models often lack the sensitivity to perceive and transfer such minute differences.

Furthermore, exemplar-based structural retouching suffers from a severe scarcity of suitable datasets. Existing datasets [3, 4, 9, 30, 42, 43, 45] primarily focus on appearance-level adjustments or lack operation diversity. More critically, constructing cross-identity training pairs—where two distinct portraits undergo the exact same combination of retouching operations—is extremely difficult. Variations in image content, such as shot scales, partial occlusions, and head poses, mean that specific local edits are rarely universally applicable (e.g., applying leg thinning to a facial close-up, or modifying eye spacing on a profile face). There-

fore, curating strictly paired, operation-aligned cross-identity datasets is highly unscalable and poses a critical data bottleneck.

To tackle these challenges, we introduce **MirrorPPR**, a framework that accurately captures and transfers subtle structural retouching operations. The framework consists of two core modules. A Retouching Operation Extractor captures the subtle differences between the source and target exemplars, producing a retouching operation representation. This representation is then injected into a pre-trained Diffusion Transformer (DiT) [55] backbone as a conditioning signal, guiding the DiT to transfer the structural retouching to a new query image. To optimize the framework, we employ a progressive two-stage training strategy: the extractor is first pre-trained, and subsequently linked to the DiT via a connector for joint LoRA [21] fine-tuning.

To successfully optimize our framework and overcome the aforementioned cross-identity misalignment challenge, we propose an advanced data self-augmentation paradigm. Instead of using a different identity for the query, we construct the training quadruplet directly from the exemplar pair by applying a randomized set of spatial augmentations to the source and target exemplars synchronously. This self-augmented formulation ensures strict alignment of the retouching operations between the exemplar pair and the query-ground truth pair, effectively reducing optimization difficulty and accelerating model convergence.

Benefiting from the data self-augmentation strategy, we eliminate the need for strictly paired cross-identity examples, requiring only a comprehensive retouching dataset. To this end, we construct **MirrorPPR47M**, comprising over 47 million retouched pairs. Because authentic structural edits are very subtle to learn from scratch, we structure the dataset to support curriculum learning [2]: it includes a simulated subset with pronounced deformations for pre-training, and a professional subset for fine-tuning on highly realistic operations.

Extensive experiments validate that our approach significantly outperforms existing baselines, achieving state-of-the-art performance in both retouching quality and identity preservation.

In summary, our main contributions are threefold:

- We introduce the novel task of *Exemplar-Based Portrait Photo Retouching* that focuses on structural reshaping, and propose **MirrorPPR**, an innovative framework that accurately captures and transfers subtle retouching operations.
- We design an advanced data self-augmentation paradigm that effectively resolves the cross-identity misalignment challenge.
- We construct **MirrorPPR47M**, a large-scale dataset comprising over 47 million pairs for comprehensive structural portrait reshaping, establishing a solid foundation for mastering intricate real-world retouching operations.

2 Related Work

Text-guided Image Editing. Diffusion-based models [20, 39, 47, 48] have revolutionized text-guided image editing by leveraging natural language instruc-

tions [5, 13, 35, 46, 51, 56]. Early works like P2P [19] and InstructPix2Pix [6] handle general edits, while MasaCtrl [10] enables tuning-free non-rigid editing. Recent state-of-the-art models further support complex tasks like image composition and style transfer via multi-reference editing [8, 23, 27, 29, 50, 55]. Despite their versatility, these methods remain inadequate for structural portrait retouching, as natural language lacks the precision to specify the exact spatial scale, direction, and magnitude required for subtle adjustments.

Exemplar-Based Image Editing. To address the semantic ambiguity of text, a growing body of research has shifted toward exemplar-based editing [16, 17, 28, 33, 36, 49, 53, 59, 60]. Early attempts cast this as an inversion problem, mapping visual examples to textual embeddings or discrete instructions [41, 58, 62]. Recent works advance toward In-Context Learning (ICL), with methods like EditTransfer [11] and VIRAL [31] addressing non-rigid transformations and heterogeneous tasks. However, as these frameworks largely target general-purpose editing, extending them to extract and transfer the fine-grained edits required for structural portrait retouching remains an underexplored challenge.

Portrait Photo Retouching. Traditional portrait retouching pipelines [9, 22, 25, 26, 37, 40] primarily focus on global, appearance-level enhancements like color grading and skin smoothing. While generative models such as StyleGAN [24, 38, 52] enable structural facial editing, they still struggle with fine-grained, localized adjustments. A major bottleneck for AI-driven structural retouching is the lack of suitable training data. Existing datasets are either dominated by appearance adjustments [7, 30, 45] or lack operation diversity [3, 42, 43]. To resolve these issues, we carefully construct a large-scale structural portrait retouching dataset.

3 Method

3.1 Task Formulation and Method Overview

Given an exemplar pair (X_s, X_t) representing the source and its retouched counterpart, alongside a new query image X_q , the objective of exemplar-based portrait photo retouching is to generate a target image $\hat{Y}_q$ by applying the identical retouching operations demonstrated in the exemplar pair. The ground-truth retouched image corresponding to X_q is denoted as Y_q .

As shown in Figure 2, our proposed MirrorPPR framework consists of two core components: (1) a **Retouching Operation Extractor** that captures subtle retouching operations from the exemplar pair, and (2) a pre-trained **Diffusion Transformer (DiT)** backbone that transfers these operations to the query image via a dedicated connector and LoRA modules.

3.2 Retouching Operation Extractor

Portrait retouching involves extremely subtle modifications. Inspired by Moto [12], we introduce the Retouching Operation Extractor to accurately capture subtle retouching operations. It consists of a frozen Masked Autoencoder (MAE) [18]

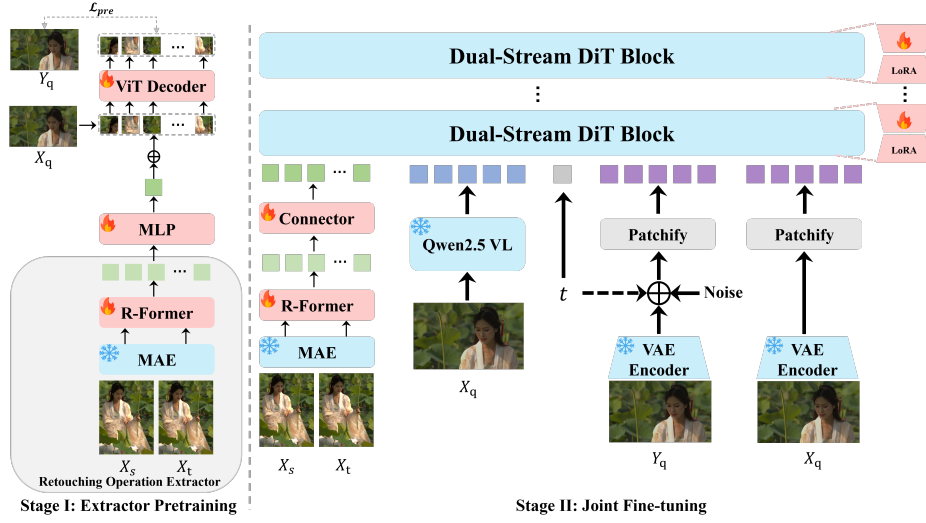


Fig. 2: Overall architecture of the proposed MirrorPPR framework. Our framework is trained in a progressive two-stage pipeline. Left: A Retouching Operation Extractor, comprising a frozen MAE and a trainable R-Former, extracts subtle retouching operations from the exemplar pair (X_s, X_t) . It is pre-trained via an auxiliary reconstruction task using a temporary MLP and ViT decoder. Right: The pre-trained R-Former is integrated with a frozen dual-stream DiT. The extracted operation features are processed by a trainable connector and injected into the DiT blocks. The R-Former, connector, and newly added LoRA modules are jointly fine-tuned to transfer the operations. The snowflake and fire icons denote frozen and trainable modules, respectively.

and a Transformer-based network, termed **R-Former**, which is specifically designed to extract retouching operations from the MAE features. MAE is particularly suitable for this extractor, as it preserves local spatial structures and provides representations favorable to geometry, localization, and fine-grained differences [57], thereby facilitating the capture of subtle retouching operations.

Retouching Operation Extraction. The patch-level features extracted by the frozen MAE from the exemplar pair are passed into the R-Former. Within the network, a set of internal learnable query tokens is concatenated with the incoming image features along the sequence dimension. As they pass through the self-attention layers, the query tokens interact with the image features, enabling them to extract the subtle retouching operations from the image features. We retain only the output representations corresponding to these queries, denoted as $\mathbf{H}_{edit}$, which serve as the raw retouching operation representation.

Extractor Pre-training. To ensure that $\mathbf{H}_{edit}$ accurately encapsulates the precise retouching intent, we introduce an auxiliary reconstruction proxy task to pre-train the extractor. Specifically, a Multi-Layer Perceptron (MLP) projects $\mathbf{H}_{edit}$ into a compact edit embedding $\mathbf{e}_{edit}$. For a given query image X_q , its patch embeddings are extracted, and $\mathbf{e}_{edit}$ is added to each patch token. A lightweight

ViT decoder then processes these fused tokens to reconstruct the retouched result $\hat{Y}_q$. This pre-training stage is optimized using a combination of Mean Squared Error (MSE) reconstruction loss and LPIPS perceptual loss [61]:

$$\mathcal{L}_{pre} = \|\hat{Y}_q - Y_q\|_2^2 + \lambda \mathcal{L}_{lips}(\hat{Y}_q, Y_q), \quad (1)$$

where λ is a balancing scalar. After pre-training, the extractor learns robust operation representations, and the auxiliary MLP and ViT decoder are discarded.

3.3 Operation Transfer via Diffusion Transformer

After pre-training the extractor, we integrate it with a pre-trained image editing diffusion model to apply the extracted retouching operations to the query image. **Backbone.** We adopt Qwen-Image-Edit-2511 [55] as our backbone, which features a dual-stream DiT. One stream leverages the Multimodal Large Language Model (MLLM) Qwen2.5-VL [1] to extract high-level visual representations, while the other stream performs diffusion denoising on the target latents conditioned on image VAE latents. Since our exemplar-based task is fundamentally driven by visual demonstrations rather than textual descriptions, we omit text instructions entirely. We retain the frozen Qwen2.5-VL to extract the visual feature embeddings $\mathbf{c}_{img}$ from X_q , and preserve the frozen VAE encoder $\mathcal{E}$ to obtain the image latent representation $\mathbf{z}_{cond} = \mathcal{E}(X_q)$.

Operation Injection. To inject the retouching operations into the diffusion process, the extracted operation representations $\mathbf{H}_{edit}$ are processed by a trainable connector. The connector maps $\mathbf{H}_{edit}$ into the instruction conditioning space of the DiT backbone, producing the final retouching operation condition $\mathbf{c}_{edit}$.

Joint Fine-tuning. Let $\mathbf{z}_0 = \mathcal{E}(Y_q)$ be the encoded ground-truth target latent, and $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a standard Gaussian noise vector. For a given timestep $t \in [0, 1]$, the intermediate noisy latent $\mathbf{z}_t$ is constructed via linear interpolation, and the target velocity $\mathbf{v}_t$ is defined as the trajectory from noise to data:

$$\mathbf{z}_t = t\mathbf{z}_0 + (1 - t)\mathbf{z}_1, \quad \mathbf{v}_t = \mathbf{z}_0 - \mathbf{z}_1. \quad (2)$$

The flow matching loss is formulated as the MSE between the predicted velocity $\mathbf{v}_\theta$ and the ground-truth velocity $\mathbf{v}_t$:

$$\mathcal{L}_{flow} = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_0, t} \left[\|\mathbf{v}_\theta([\mathbf{z}_t, \mathbf{z}_{cond}], t, \mathbf{c}_{img}, \mathbf{c}_{edit}) - \mathbf{v}_t\|_2^2 \right]. \quad (3)$$

To efficiently adapt the powerful prior of the pre-trained DiT to our task, we freeze its original parameters and add trainable LoRA [21] modules into its attention blocks. The pre-trained R-Former, the connector, and the LoRA modules are end-to-end jointly optimized via $\mathcal{L}_{flow}$, enabling the network to accurately transfer the subtle retouching operations.

3.4 Data Self-Augmentation Paradigm

As discussed in Section 1, constructing valid training quadruplets for exemplar-based portrait retouching is non-trivial. A straightforward strategy is to build

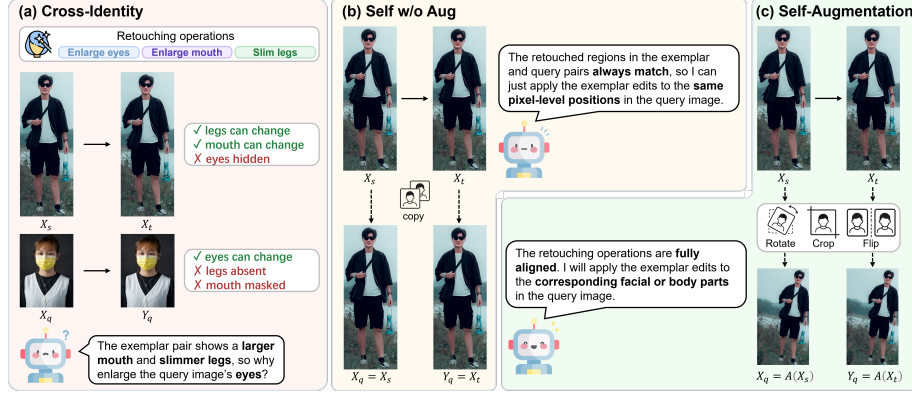


Fig. 3: Illustration of the proposed data self-augmentation paradigm. The Cross-Identity setting suffers from operation misalignment between the exemplar pair and the query pair, while Self w/o Aug causes pixel-level shortcut learning. Our Self-Augmentation applies the same spatial augmentation A to both X_s and X_t , which preserves operation alignment while breaking absolute coordinate correspondence.

Cross-Identity quadruplets, where the exemplar pair (X_s, X_t) and query pair (X_q, Y_q) come from different identities and are expected to share the same intended retouching operations. However, as shown in Figure 3 (a), due to variations in shot scale, pose, occlusion, and portrait composition, the operation between the two pairs may not strictly correspond. Such misalignment makes the supervision ambiguous and hinders the model from learning a consistent operation-transfer mapping from the exemplar pair to the query pair.

A natural workaround is to construct the training quadruplet directly from a single exemplar pair. We denote the naive version as *Self w/o Aug*, where the query image is set to the source exemplar and the ground truth is set to the target exemplar, i.e., $X_q = X_s$ and $Y_q = X_t$. Although this construction avoids cross-identity operation misalignment, it introduces severe *shortcut learning*. Since the exemplar pair and the query pair are perfectly aligned at the pixel level, the model can trivially copy coordinate-wise differences rather than infer the underlying retouching semantics. Consequently, it struggles to generalize to query images with different identities, spatial layouts, or scales.

To break this spatial coupling while retaining exact operation consistency, we introduce *Self-Augmentation*. Specifically, we synthesize the query pair by applying a shared random spatial augmentation A , such as scaling, cropping, rotation, and horizontal flipping, to both the source and target exemplars. The query pair is therefore constructed as $X_q = A(X_s)$ and $Y_q = A(X_t)$. Because the same augmentation is applied to both images, the retouching operations in the synthesized query pair remain exactly consistent with those in the exemplar pair. Meanwhile, the transformed query pair no longer shares the same absolute pixel coordinates with the original exemplar pair (X_s, X_t) , preventing shortcut learning. By preserving operation alignment while breaking absolute

pixel-level correspondence, this construction encourages the model to transfer the retouching operations in the exemplar pair according to the query image’s own spatial layout, thereby overcoming the cross-identity data bottleneck and enabling robust generalization of the learned operations.

4 MirrorPPR47M Dataset

To overcome the scarcity of paired geometric retouching data, we construct **MirrorPPR47M**, a large-scale portrait dataset encompassing comprehensive structural reshaping. These operations extensively cover facial features, face contours, and body proportions across diverse shot scales. Since real-world retouching is very subtle, training a network from scratch is highly challenging. To mitigate this, MirrorPPR47M is designed to support an easy-to-hard *curriculum learning strategy*. It comprises a **Simulated Retouching Subset** featuring pronounced deformations and a **Professional Retouching Subset** providing authentic, complex operations. We construct the dataset through the following pipeline.

4.1 Data Curation and Filtering

We source raw high-resolution images and apply strict filtering criteria based on head pose, facial occlusion, and portrait area ratio. Detailed criteria are provided in Appendix A.1. For the *Simulated Retouching Subset*, we retain 30,171 high-quality images from the FFHQ dataset [24]. For the *Professional Retouching Subset*, we select 3,789 4K-8K portraits from the PPR10K dataset [32].

4.2 Retouching Execution

Simulated Retouching Generation. To synthesize pronounced geometric deformations, we propose the Landmark-Guided Local Warping (LLW) algorithm. Detailed formulations and implementations of this algorithm are provided in Appendix A.2. We simulate 8 types of facial operations, each with two opposite directions. By randomly combining 1 to 8 operations per identity, we synthesize 808,439 retouched pairs.

Professional Retouching Generation. To approximate real-world authenticity, we process the source images from PPR10K using a commercial retouching API. We select 27 professional operations, comprising 18 for facial features, 4 for face shapes, and 5 for body proportions. Applying random combinations of 1 to 7 operations per image yields 46,642,845 finely retouched pairs.

Appendix A.4 provides more detailed operation taxonomy and statistics.

4.3 Data Processing and Self-Augmentation

Following the Self-Augmentation paradigm in Section 3.4, we implement a concrete data processing pipeline to generate spatially decoupled yet operation-aligned training quadruplets (X_s, X_t, X_q, Y_q) .

We first crop and resize raw images to a total area of approximately 4 MP. This cropping process utilizes facial bounding boxes and human body masks to strictly ensure the completeness of the portrait subject. Detailed cropping procedures are provided in Appendix A.3. Subsequently, we synchronously apply randomized spatial augmentations to the source and target images. These augmentations include mirroring, rotation, and dynamic cropping with varied spatial offsets. Consequently, the retouching operations in the query pair (X_q, Y_q) appear at entirely different spatial coordinates and scales compared to the exemplar pair (X_s, X_t) . This strategy effectively prevents shortcut learning and expands each original pair into approximately 13.3 spatially decoupled variations.

4.4 Progressive Curriculum Learning

We use the two subsets to design a progressive training strategy. First, the Retouching Operation Extractor is pre-trained on the Simulated Retouching Subset to capture fundamental structural variations, then further trained on the Professional Retouching Subset to adapt to intricate, real-world retouching subtleties. In the final fine-tuning stage, all trainable modules are exclusively optimized on the professional data. This tailored data recipe ensures that the network can stably capture highly complex structural retouching operations.

5 Experiments

5.1 Experimental Setting

Benchmarks. To rigorously evaluate cross-identity operation transfer capabilities, we construct two benchmarks where the exemplar pair and query image feature entirely different identities—a crucial distinction from our self-augmented training phase. **SimFace-100** comprises 100 retouching combinations randomly paired from 12 distinct 1024×1024 facial images using 8 operations supported by LLW, testing the preliminary perception of prominent simulated retouching. **ProPortrait-500** involves 500 combinations of 27 professional operations applied to 40 high-quality portraits using the same commercial API as the Professional Retouching Subset, with images standardized to an area equivalent to $1.5K \times 1.5K$ pixels. This benchmark poses a severe challenge for capturing highly subtle and fine-grained professional edits.

Evaluation Metrics. We employ PSNR, SSIM [54], and LPIPS [61] to evaluate pixel-level fidelity, structural reconstruction accuracy, and perceptual similarity, respectively. Following MoFRR [34], we introduce Face Similarity to evaluate identity preservation during retouching. This metric leverages the face recognition model ArcFace [14] to extract feature embeddings from the generated result and the ground truth, and computes their cosine similarity for assessment.

Table 1: Quantitative comparison results with baselines on SimFace-100.

Category	Model	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Face Similarity $\uparrow$
Multi-reference Image Editing	Qwen-Image-Edit-2511	9.06	0.468	0.745	0.207
	FLUX.2-dev	9.28	0.481	0.698	0.110
	Nano Banana 2	16.72	0.784	0.329	0.556
	Seedream 4.5	13.01	0.709	0.501	0.351
Exemplar-based	Qwen-Image-Edit-2511-ICedit-LoRA	9.21	0.533	0.640	0.300
	RelationAdapter	16.57	0.698	0.543	0.204
	EditTransfer	15.68	0.691	0.492	0.464
Text-guided	Qwen-Image-Edit-2511	25.80	0.862	0.260	0.463
	FLUX.2-dev	22.44	0.804	0.301	0.531
	Nano Banana 2	24.25	0.860	0.239	0.601
	Seedream 4.5	18.01	0.788	0.368	0.600
Ours	MirrorPPR-Face	32.25	0.909	0.186	0.937

Implementation Details. Corresponding to our benchmarks, we train two model variants. **MirrorPPR-Face** is trained in two stages on 1024×1024 simulated facial crops extracted using FFHQ facial bounding boxes. **MirrorPPR-Pro** undergoes the progressive curriculum learning proposed in Section 4.4. Specifically, the extractor is sequentially pre-trained on the simulated and professional subsets, followed by joint fine-tuning of the entire framework exclusively on the professional data. Detailed network configurations and training settings are provided in Appendices B and C.

5.2 Comparison with Other Methods

To comprehensively evaluate the effectiveness of our proposed MirrorPPR, we compare it against powerful baselines across three distinct categories. For multi-reference image editing, we choose Qwen-Image-Edit-2511 [55], FLUX.2-dev [27], Nano Banana 2 [50], and Seedream 4.5 [44]. For exemplar-based image editing, we select three recent DiT-based methods: Qwen-Image-Edit-2511-ICedit-LoRA [15], RelationAdapter [16], and EditTransfer [11]. Additionally, we evaluate the text-guided editing capabilities of Qwen-Image-Edit-2511, FLUX.2-dev, Nano Banana 2, and Seedream 4.5. For multi-reference and exemplar-based methods, we uniformly use the prompt: "Infer the portrait retouching operations applied from Image 1 (original) to Image 2 (retouched), then apply the same retouching operations to Image 3." For text-guided methods, we provide clear text instructions. More evaluation details are provided in Appendix D.

Quantitative Analysis. Quantitative results on the SimFace-100 and ProPortrait-500 benchmarks are summarized in Table 1 and Table 2, respectively. From the results, we draw the following conclusions:

Multi-reference and exemplar-based methods completely fail on this task. As shown in the tables, existing multi-reference editing models (with the exception of Nano Banana 2) and baseline exemplar-based methods exhibit

Table 2: Quantitative comparison results with baselines on ProPortrait-500.

Category	Model	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Face Similarity $\uparrow$
Multi-reference Image Editing	Qwen-Image-Edit-2511	10.23	0.538	0.645	0.413
	FLUX.2-dev	9.36	0.466	0.728	0.220
	Nano Banana 2	17.72	0.835	0.250	0.811
	Seedream 4.5	12.12	0.689	0.436	0.705
Exemplar-based	Qwen-Image-Edit-2511-ICEdit-LoRA	12.06	0.631	0.564	0.606
	RelationAdapter	15.74	0.709	0.586	0.283
	EditTransfer	18.32	0.748	0.481	0.457
Text-guided	Qwen-Image-Edit-2511	20.85	0.732	0.387	0.501
	FLUX.2-dev	19.94	0.748	0.345	0.616
	Nano Banana 2	27.45	0.904	0.183	0.667
	Seedream 4.5	16.43	0.770	0.378	0.782
Ours	MirrorPPR-Pro	32.65	0.927	0.200	0.960

extremely poor performance. All evaluation metrics fall far below the normal ranges expected for portrait retouching tasks.

Text-guided methods perform better in reconstruction but struggle with identity preservation. Compared to the first two categories, text-guided models demonstrate significantly better pixel-level reconstruction and structural fidelity, achieving much higher PSNR and SSIM scores. However, a prominent weakness across these models is their noticeably low Face Similarity scores. For instance, although the text-guided Nano Banana 2 achieves a highly competitive LPIPS of 0.183 on the ProPortrait-500 benchmark, its Face Similarity drops to merely 0.667. This sharp contrast indicates that text-guided models severely compromise the biometric characteristics of the portrait subject.

MirrorPPR demonstrates comprehensive superiority. Both MirrorPPR-Face and MirrorPPR-Pro consistently yield excellent scores across the evaluation metrics on both benchmarks. On the highly challenging ProPortrait-500 dataset, MirrorPPR-Pro achieves a PSNR of 32.65, an SSIM of 0.927, and an unprecedented Face Similarity of 0.960. These superior metrics robustly demonstrate the effectiveness of our method for exemplar-based portrait retouching.

Qualitative Comparison. We manually inspected the test outputs to explain the metric results and summarize the main causes of baseline failures. Specifically, multi-reference and exemplar-based models typically misinterpret operation transfer as image blending or face swapping. Meanwhile, text-guided models lack precise geometric control due to language ambiguity, frequently producing anatomically exaggerated, "over-edited" features that corrupt the original identity. In contrast, our MirrorPPR achieves precise structural reshaping while perfectly preserving the unedited details. Representative examples illustrating these observations are presented in Figure 4 and Figure 5.

User Study. To complement automatic metrics and visual comparisons, we further conduct a user study in which participants select the anonymized candidate

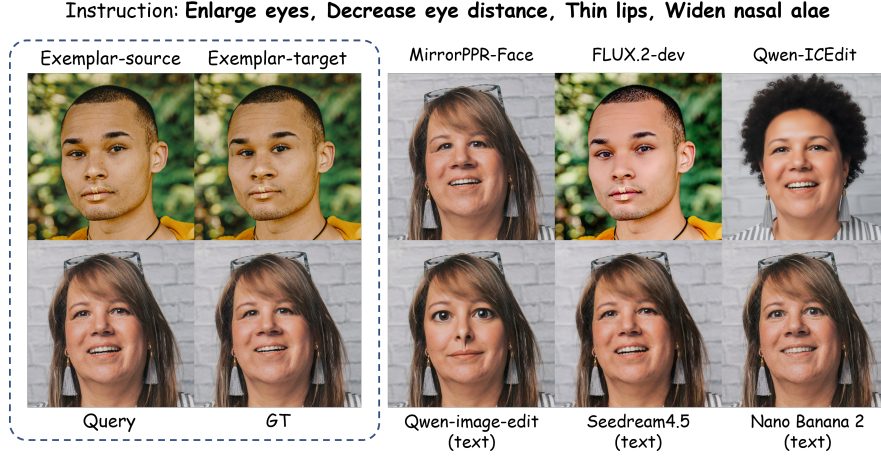


Fig. 4: Qualitative comparison on SimFace-100. MirrorPPR-Face accurately captures and faithfully transfers all operations, while other models suffer from various issues.

that best transfers the demonstrated retouching operation while preserving identity and avoiding changes to unrelated regions. MirrorPPR-Pro receives 79.0% overall preference, showing that its advantage is also clearly perceived by human users. More details about the user study are provided in Appendix E.

5.3 Latent Space Analysis of the Extracted Operations

To gain a deeper understanding of the operation representations learned by our framework, we conduct further analyses on the edit embeddings $\mathbf{e}_{edit}$. These embeddings are obtained from the Retouching Operation Extractor and subsequently mapped through the auxiliary MLP. We evaluate these embeddings on the SimFace-100 benchmark, revealing several compelling properties of the learned latent space, including robust operation transfer consistency, distinct operation clustering, and vector additivity.

Operation Transfer Consistency. A key requirement of our framework is to ensure that the exact retouching intent demonstrated in the exemplar pair is faithfully transferred to the target query image. To quantitatively verify this, we feed the exemplar pair (X_s, X_t) into the pre-trained extractor and the MLP to obtain the edit embedding $\mathbf{e}_{edit}^{exemplar}$. Concurrently, we extract the edit embedding $\mathbf{e}_{edit}^{query}$ from the query image and our generated retouching result $(X_q, \hat{Y}_q)$.

On SimFace-100, the average cosine similarity between $\mathbf{e}_{edit}^{exemplar}$ and $\mathbf{e}_{edit}^{query}$ is remarkably high at 0.950. This exceptionally high similarity rigorously ensures that the retouching operations in the output perfectly align with the exemplar.

Clustering of Retouching Operations. We further investigate whether our pre-trained extractor learns distinct representations for different retouching operations. We construct a complete single-operation subset comprising 192 image

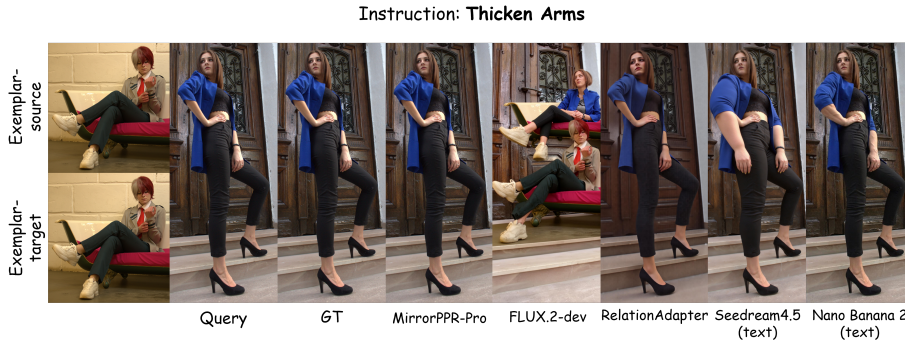


Fig. 5: Qualitative comparison on ProPortrait-500. MirrorPPR-Pro accurately captures and transfers all operations, while other models suffer from various issues.

Table 3: Quantitative evaluation of vector additivity on SimFace-100. Simply adding the individual operation vectors directly in the latent space surpasses all baselines.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Face Similarity $\uparrow$
MirrorPPR-Face	32.25	0.909	0.186	0.937
Latent Vector Additivity	30.85	0.892	0.196	0.869

pairs (12 distinct face images, where 16 different simulated single-operation edits are applied to each face) and extract their corresponding edit embeddings $\mathbf{e}_{edit}$.

Raw edit embeddings from the pre-trained model inevitably retain instance-specific image characteristics. To isolate the pure retouching directions, we apply *instance-level mean centering*: for each image, we subtract the mean vector of its 16 operations from the individual embeddings. As shown in Figure 6, these centered directional vectors exhibit clear operation-based clustering.

Vector Additivity. Based on the well-clustered embeddings, we further examine their additivity by testing whether they support direct vector addition for multiple operations. Using the identical 192-sample single-operation dataset, we attempt to composite multiple operations in the latent space.

For a specific image requiring a composite edit (e.g., Operation A + Operation B), we first isolate the pure editing directions, denoted as Δ_A and Δ_B , by applying the aforementioned mean centering. The composite edit embedding is then computed as $\mathbf{e}_{composite} = \mathbf{e}_{mean} + \Delta_A + \Delta_B$, where $\mathbf{e}_{mean}$ is the mean base vector of that instance. This embedding is then directly fed into the pre-trained ViT decoder to generate the retouched result.

We evaluate this vector addition approach on SimFace-100. As presented in Table 3, it surpasses all compared baselines in Table 1. This compelling finding suggests that our learned latent space possesses a robust additive property, providing a reliable foundation for representing complex combined operations.

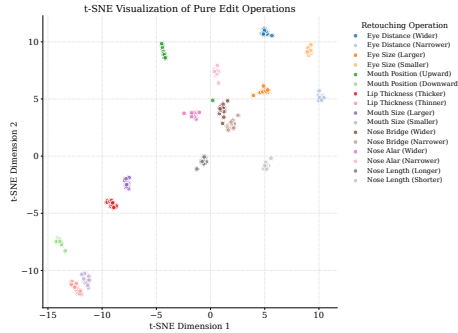


Fig. 6: t-SNE visualization of the re-touching embeddings. The clear clustering confirms our extractor’s ability to learn distinct, identity-agnostic representations for different retouching operations.

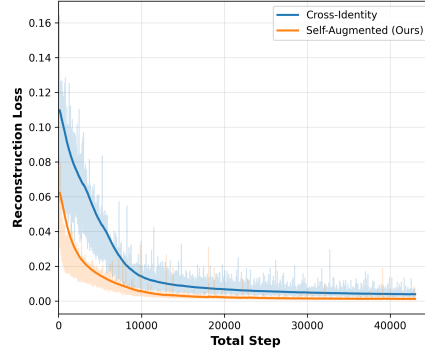


Fig. 7: Training loss curves of the Cross-Identity and Self-Augmentation settings.

Table 4: Ablation results on SimFace-100 for evaluating whether training with Self-Augmentation weakens cross-identity transfer.

Setting	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Face Similarity $\uparrow$
Self w/o Aug	28.63	0.906	0.213	0.680
Cross-Identity	32.08	0.909	0.186	0.937
Self-Augmentation (Ours)	32.25	0.909	0.186	0.937

5.4 Ablation Study

We conduct two ablation studies to validate the proposed Data Self-Augmentation paradigm. The first study tests whether training with Self-Augmentation causes a training-inference mismatch and weakens cross-identity operation transfer. The second study investigates whether Self-Augmentation brings clearer benefits when cross-identity training data is difficult to keep operation-aligned.

Self-Augmentation does not weaken cross-identity transfer. To verify this, we conduct an ablation on the simulated retouching subset and evaluate the models on SimFace-100. We evaluate the three data construction strategies defined in Section 3.4 and illustrated in Figure 3: Cross-Identity, Self w/o Aug, and Self-Augmentation. In this ablation, the Cross-Identity setting is a relatively ideal scenario with good operation alignment, since the data are face-centric and the retouching operations are controlled by LLW.

As shown in Table 4, “Self w/o Aug” performs the worst, especially in Face Similarity, confirming that directly reusing the exemplar pair causes severe spatial shortcut learning. Qualitative failure cases are provided in Appendix F.1. In contrast, Self-Augmentation achieves comparable performance to Cross-Identity.

Table 5: Ablation results on ProPortrait-500 for evaluating the benefit of self-augmentation when cross-identity training data is difficult to align.

Setting	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Face Similarity $\uparrow$
Cross-Identity	30.52	0.914	0.212	0.916
Self-Augmentation (Ours)	32.65	0.927	0.200	0.960

This shows that using the same-identity self-augmented quadruplets during training does not harm cross-identity operation transfer.

Self-Augmentation is more advantageous when operation alignment is difficult. To verify this, we conduct an ablation on the professional retouching subset and evaluate on ProPortrait-500. Compared with the simulated subset, the professional subset involves more diverse identities, shot scales, poses, portrait compositions, and locally applicable operations, making cross-identity training data much harder to keep operation-aligned. We therefore compare the two relevant data construction strategies: Cross-Identity and Self-Augmentation.

As shown in Table 5, Self-Augmentation clearly outperforms Cross-Identity across all metrics. These results demonstrate that self-augmentation provides a more reliable training paradigm when cross-identity training data suffers from imperfect operation alignment.

We also compare the training efficiency between Self-Augmentation and Cross-Identity. As shown in Figure 7, Self-Augmentation converges faster than Cross-Identity. This suggests that Self-Augmentation reduces the learning difficulty by preserving exact operation consistency while avoiding unnecessary interference from cross-identity pairing.

More ablation studies are provided in Appendix F.

6 Conclusion

In this paper, we introduce the novel task of Exemplar-Based Portrait Photo Retouching and propose MirrorPPR, which integrates a Retouching Operation Extractor with a Diffusion Transformer. To address operation misalignment and data scarcity, we develop a data Self-Augmentation paradigm and construct MirrorPPR47M, a large-scale dataset with over 47 million pairs for progressive curriculum learning. Extensive experiments demonstrate that MirrorPPR achieves state-of-the-art performance in retouching quality and identity preservation, establishing a solid foundation for real-world structural portrait retouching.

Acknowledgments

This work was supported by Shanghai Key Technology R&D Program “New Generation of Information Technology” (No. 25511103700), NSF of China (Nos. 62306176, 92470118), CCF-ALIMAMA TECH Kangaroo Fund (NO. CCF-ALIMAMA OF 2025010), and Ant Group.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), <https://arxiv.org/abs/2502.13923>
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
3. Bharati, A., Singh, R., Vatsa, M., Bowyer, K.W.: Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security* **11**(9), 1903–1913 (2016)
4. Bharati, A., Vatsa, M., Singh, R., Bowyer, K.W., Tong, X.: Demography-based facial retouching detection using subclass supervised sparse autoencoder. In: 2017 IEEE international joint conference on biometrics (IJCB). pp. 474–482. IEEE (2017)
5. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledit++: Limitless image editing using text-to-image models (2024), <https://arxiv.org/abs/2311.16711>
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
7. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR 2011. pp. 97–104. IEEE (2011)
8. Cai, H., Wang, X., Bai, Y., Zhou, T., Xu, S., Hao, Y., Cui, Z., Yang, Y., Zhu, W., Chen, Y., Tang, X., Hu, Y., Li, Z.: Idglow: Dynamic identity modulation for multi-subject generation (2026), <https://arxiv.org/abs/2603.00607>
9. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. *IEEE transactions on image processing* **27**(4), 2049–2062 (2018)
10. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
11. Chen, L., Mao, Q., Gu, Y., Shou, M.Z.: Edit transfer: Learning image editing via vision in-context relations. *arXiv preprint arXiv:2503.13327* (2025)
12. Chen, Y., Ge, Y., Tang, W., Li, Y., Ge, Y., Ding, M., Shan, Y., Liu, X.: Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19752–19763 (2025)
13. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance (2022), <https://arxiv.org/abs/2210.11427>
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
15. DiffSynth-Studio: Qwen-image-edit-2511-icedit-lora. Hugging Face Model Repository (2025), <https://huggingface.co/DiffSynth-Studio/Qwen-Image-Edit-2511-ICEdit-LoRA>, accessed on June 24, 2026

16. Gong, Y., Song, Y., Li, Y., Li, C., Zhang, Y.: Relationadapter: Learning and transferring visual relation with diffusion transformers. arXiv preprint arXiv:2506.02528 (2025)
17. Gu, Z., Yang, S., Liao, J., Huo, J., Gao, Y.: Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)* **43**(4), 1–15 (2024)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 16000–16009 (2022)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *Iclr* **1**(2), 3 (2022)
22. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)* **37**(2), 1–17 (2018)
23. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4401–4410 (2019)
25. Kim, Y.T.: Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE transactions on Consumer Electronics* **43**(1), 1–8 (1997)
26. Kosugi, S., Yamasaki, T.: Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34, pp. 11296–11303 (2020)
27. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. <https://bf1.ai/blog/flux-2> (2025), accessed on June 24, 2026
28. Lai, B., Juefei-Xu, F., Liu, M., Dai, X., Mehta, N., Zhu, C., Huang, Z., Rehg, J.M., Lee, S., Zhang, N., et al.: Unleashing in-context learning of autoregressive models for few-shot image manipulation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18346–18357 (2025)
29. Li, S., Singh, H., Grover, A.: Instructany2pix: Flexible visual editing via multi-modal instruction following. arXiv preprint arXiv:2312.06738 (2023)
30. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: *Proceedings of the 26th ACM international conference on Multimedia*. pp. 645–653 (2018)
31. Li, Z., Duan, Z., Ye, J., Chen, C., Chen, D., Li, Y., Chen, Y.: Viral: Visual in-context reasoning via analogy in diffusion transformers. arXiv preprint arXiv:2602.03210 (2026)
32. 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: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 653–661 (2021)
33. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. arXiv preprint arXiv:1705.01088 (2017)

34. Liu, J., Ying, Q., Qian, Z., Li, S., Zhang, R., Liu, J., Zhang, X.: Mofrr: Mixture of diffusion models for face retouching restoration. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12842–12851 (2025)
35. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761* (2025)
36. Lu, H., Chen, J., Yang, Z., Gnanha, A.T., Wang, F.L., Qing, L., Mao, X.: Pairedit: Learning semantic variations for exemplar-based image editing. *arXiv preprint arXiv:2506.07992* (2025)
37. Mantiuk, R., Daly, S., Kerofsky, L.: Display adaptive tone mapping. In: *ACM SIGGRAPH 2008 papers*, pp. 1–10 (2008)
38. Medin, S.C., Egger, B., Cherian, A., Wang, Y., Tenenbaum, J.B., Liu, X., Marks, T.K.: Most-gan: 3d morphable stylegan for disentangled face image manipulation. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 36, pp. 1962–1971 (2022)
39. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021)
40. Mukherjee, J., Mitra, S.K.: Enhancement of color images by scaling the dct coefficients. *IEEE Transactions on Image processing* **17**(10), 1783–1794 (2008)
41. Nguyen, T., Li, Y., Ojha, U., Lee, Y.J.: Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems* **36**, 9598–9613 (2023)
42. Rathgeb, C., Botaljev, A., Stockhardt, F., Isadskiy, S., Debiasi, L., Uhl, A., Busch, C.: Prnu-based detection of facial retouching. *IET Biometrics* **9**(4), 154–164 (2020)
43. Rathgeb, C., Satnoianu, C.I., Haryanto, N.E., Bernardo, K., Busch, C.: Differential detection of facial retouching: A multi-biometric approach. *IEEE Access* **8**, 106373–106385 (2020)
44. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427* (2025)
45. Shafaei, A., Little, J.J., Schmidt, M.: Autoretouch: Automatic professional face retouching. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 990–998 (2021)
46. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8871–8879 (2024)
47. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International conference on machine learning*. pp. 2256–2265. *pmlr* (2015)
48. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019)
49. Srivastava, A., Menta, T.R., Java, A., Jadhav, A., Singh, S., Jandial, S., Krishnamurthy, B.: Reedit: Multimodal exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2411.03982* (2024)
50. 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)

51. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., Cai, X., Guan, Y., Hu, J.: Longcat-image technical report. arXiv preprint arXiv:2512.07584 (2025)
52. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6142–6151 (2020)
53. Wang, Z., Jiang, Y., Lu, Y., He, P., Chen, W., Wang, Z., Zhou, M., et al.: In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems* **36**, 8542–8562 (2023)
54. 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), 600–612 (2004)
55. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), <https://arxiv.org/abs/2508.02324>
56. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
57. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14475–14485 (2023)
58. Xu, P., Fan, Q., Kou, F., Qin, S., Gu, H., Zhao, R., Ling, C., Wang, B.: Textualize visual prompt for image editing via diffusion bridge. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 21779–21787 (2025)
59. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18381–18391 (2023)
60. Yang, Y., Peng, H., Shen, Y., Yang, Y., Hu, H., Qiu, L., Koike, H., et al.: Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems* **36**, 48723–48743 (2023)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
62. Zhao, R., Fan, Q., Kou, F., Qin, S., Gu, H., Wu, W., Xu, P., Zhu, M., Wang, N., Gao, X.: InstructBrush: Learning Attention-based Instruction Optimization for Image Editing (2024). <https://doi.org/10.48550/arXiv.2403.18660>