
Self-Supervised Selective-Guided Diffusion Model for Old-Photo Face Restoration

Wenjie Li¹, Xiangyi Wang¹, Heng Guo^{1*}, Guangwei Gao², Zhanyu Ma¹

¹Beijing University of Posts and Telecommunications

²Nanjing University of Science and Technology

{cswjli, guoheng, mazhanyu}@bupt.edu.cn csggao@gmail.com

Abstract

Old-photo face restoration poses significant challenges due to compounded degradations such as breakage, fading, and severe blur. Existing pre-trained diffusion-guided methods either rely on explicit degradation priors or global statistical guidance, which struggle with localized artifacts or face color. We propose Self-Supervised Selective-Guided Diffusion (SSDiff), which leverages pseudo-reference faces generated by a pre-trained diffusion model under weak guidance. These pseudo-labels exhibit structurally aligned contours and natural colors, enabling region-specific restoration via staged supervision: structural guidance applied throughout the denoising process and color refinement in later steps, aligned with the coarse-to-fine nature of diffusion. By incorporating face parsing maps and scratch masks, our method selectively restores breakage regions while avoiding identity mismatch. We further construct VintageFace, a 300-image benchmark of real old face photos with varying degradation levels. SSDiff outperforms existing GAN-based and diffusion-based methods in perceptual quality, fidelity, and regional controllability. Code link: <https://github.com/PRIS-CV/SSDiff>.

1 Introduction

Recent image restoration methods [1, 2, 3] excel in generic scenes but struggle with facial regions, may causing misaligned features. To address this, face restoration [4, 5] has been extensively studied, particularly in blind face restoration (BFR) [6, 7, 8, 9, 10, 11] for real scenarios. However, the restoration of old face photographs remains relatively underexplored, as such low-quality (LQ) images often suffer from compounded degradations, including fading, breakage, and severe blurring, which pose challenges for constructing large-scale paired datasets for supervised learning.

Although existing learning-based BFR methods, such as GPEN [7], CodeFormer [9], DiffFace [10], and DiffBIR [11], can alleviate blurring in old face photos, they struggle to address degradations like breakage and color fading that are not present in the training data, as illustrated in Fig. 1. Recently, the powerful generative capability of diffusion priors [12, 13] has opened up new possibilities for face restoration. By designing effective guidance strategies for pre-trained diffusion models, the denoising trajectory can be steered toward task-specific objectives, enabling adaptation to various zero-shot restoration tasks [12, 14, 15, 16] in a train-free manner. Such flexibility and low cost make pre-trained diffusion-guided methods well-suited for old-photo face restoration.

Nevertheless, existing pre-trained diffusion-guided methods [14, 16] for old-photo face restoration are limited by rigid, pre-defined priors. For example, DDNM [14] constructs a closed-form solution space using the pseudo-inverse of a linear degradation model and adjusts the denoising trajectory through backpropagation. However, as shown in Fig. 1, degradation in old photos is complex and

*Corresponding Author.

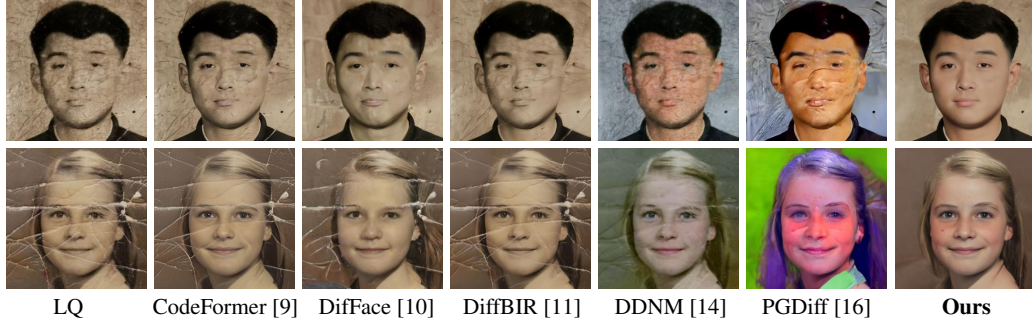


Figure 1: Existing face restoration methods face challenges with complex degradations in old face photos, which may result in residual breakage or unnatural facial colors. In contrast, our method restores sharp facial structures and natural facial color, and no visible breakage regions.

nonlinear, resulting in DDNM only being able to process old photos with relatively simple degradation. PGDiff [16] improves generalization by introducing pre-defined global attribute priors, such as color statistics and smoothing semantics. However, it faces two key limitations: *i)* smoothing semantic priors is less robust to large breakage regions and may cause the model to collapse facial structure consistency during early sampling, leading to noticeable artifacts. *ii)* enforcing global color statistics (*i.e.*, matching the average mean and variance of color channels in FFHQ [17] dataset) may induce sampling bias toward local extrema, resulting in significant image color inhomogeneity.

Motivated by these limitations, we investigate the generative behaviors of pre-trained diffusion models under different guidance and discover a key insight. **Our Insight:** As shown in Fig. 2, we observe that, under weak guidance from a degraded input with appropriate s , a pre-trained generative face diffusion model can produce face images that, despite deviating from the original identity (low feature IOU, *e.g.*, eyes, mouth, *etc.*), exhibit structurally similar facial contours (high contour IOU), natural color tones (saturation distribution close to HQ faces in FFHQ), and effective restoration in breakage regions (low edge strength variation). This motivates us to use such images as pseudo-references to guide the restoration of facial colors and breakage regions. Meanwhile, facial components like eyes and mouth are not involved in guiding features to prevent low-fidelity facial results.

Building on this insight, we introduce Self-Supervised Selective-Guided Diffusion (SSDiff), a training-free framework that performs self-supervision by using pseudo-references generated under weak guidance. To fully exploit the guidance potential of pseudo-references, SSDiff introduces a staged guidance scheme aligned with the coarse-to-fine nature of diffusion: restoration-oriented guidance, including covering both structure-aware and breakage completion, is applied throughout the process, while color refinement is introduced in later steps, as discussed in Section 3.2. Additionally, we integrate face parsing maps and scratch masks obtained from inputs to selectively guide breakage completion and face coloring without disturbing identity-sensitive features. Through our design, our restorations exhibit high IOU of face contour and components, smooth breakage regions, and consistent color saturation, as shown in Fig. 2. Therefore, SSDiff produces high-quality results even under complex degradations, as shown in Fig. 1. Moreover, benefiting from face parsing-based region selection, SSDiff can also perform region-specific stylized restoration, such as hair and lips.

Contributions. **i)** We propose SSDiff, a training-free framework that introduces a reference learning paradigm, where pseudo-references generated under weak guidance from a pre-trained face diffusion model are suitable to steer the reverse process for high-fidelity old-photo face restoration. **ii)** We design a staged, region-specific guidance scheme that combines structure-aware and breakage-aware restoration and late-stage color guidance, enabling robust restoration across diverse degradation types. **iii)** We construct VintageFace, a real-world benchmark of 300 old face photos with varied degradations, and demonstrate that SSDiff consistently outperforms existing GAN-based and zero-shot diffusion methods in perceptual quality and fidelity of restorations.

2 Related Work

Blind Face Restoration. Blind face restoration aims to recover HQ face images from LQ face images with complex degradations. To enhance visual quality, previous approaches integrate pre-trained

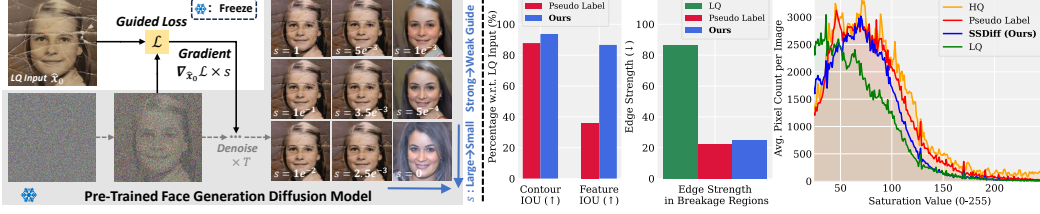


Figure 2: **(Left)** LQ inputs provides supervision via a loss gradient scaled by s , aligning reverse diffusion outputs toward it. Strong guidance improves fidelity but harms perceptual quality, while weaker guidance enhances perceptual quality but reduces fidelity. **(Right)** Statistical analysis on our dataset shows that *pseudo-labels generated under weak guidance* ($s = 1e^{-3}$) preserve facial contours similar to LQ inputs, exhibit low similarity in semantic regions (e.g., eyes, mouth, nose), and achieve smooth breakage regions with color saturation distributions close to HQ faces.

generative adversarial networks [18] trained on HQ faces into the restoration pipeline. Methods such as GFPGAN [6], GPEN [7], GLEAN [19], SGPN [20], and FReX [21] leverage facial structure cues extracted from degraded inputs and inject them into a pre-trained StyleGAN [17] for restoration. To mitigate the uncertainty of StyleGAN in continuous feature space, approaches like VQFR [8], CodeFormer [9], and DAEFR [22] adopt pre-trained VQGAN [23] to retrieve semantically similar facial features from a discrete codebook. However, GAN-based methods suffer from mode collapse. With the rise of diffusion probabilistic models [24], diffusion-based methods like DR2 [25] employ diffusion models to resist complex degradations, while DiffFace [10] models the posterior distribution from LQ to HQ images to generate high-fidelity results. PFStorer [26], PLTrans [27], and Wave-Face [28] further introduce high-frequency features to constrain the denoising process for improved fidelity. Nevertheless, these methods remain limited in handling old-photo face restoration, which may involve breakages and fading that don't appear in training. Our method, situated within the blind face restoration paradigm, is more robust to such challenging degradations in old face photos.

Diffusion Prior for Restoration. Methods leveraging diffusion priors for restoration can be broadly categorized into two groups. The first group, including DiffBIR [11], StableSR [29], OSDDiff [30], and OSDFace [31], incorporates pre-trained diffusion models such as Stable Diffusion [13] into the restoration pipeline and fine-tunes controllable modules for generative restoration. These approaches perform well on degradation types seen during training but fail to generalize to unseen ones. The second group focuses on zero-shot restoration by developing efficient guidance strategies for pre-trained diffusion models. Among them, DDRM [32], DDNM [14], GDP [15], and T2I [33] adjust pre-trained diffusion models by estimating a degradation process at each iteration, using fixed linear operators [32, 14, 33] or parameterized degradation models [15] to steer intermediate outputs towards the input LQ images. PGDiff [16] and AGLLDiff [34] guide restoration by modeling desired attributes of HQ images, such as color statistics and structural information from clean datasets. Our method belongs to the second category but differs by using pseudo-references generated through weak guidance instead of explicit attribute modeling. Furthermore, unlike applying global priors, SSDiff selectively guides facial sub-regions in stages, enabling finer control under complex degradation.

3 Methodology

3.1 Preliminaries

Denosing Diffusion Probabilistic Models (DDPM). DDPM generative models [24] consist of two processes: **i)** the forward process progressively adds Gaussian noise \mathcal{N} to the input x_0 with a predefined variance schedule $\{\beta_i\}_{i=0}^T$, producing a noisy sample x_t at timestep t according to:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \quad (1)$$

$$x_t = \sqrt{1 - \beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \quad (2)$$

ii) The reverse process is parameterized by a learned denoising network ϵ_θ , typically UNet [35] architecture, which iteratively parameterize the mean value $\mu_\theta(x_t, t)$ by network ϵ_θ :

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad \Sigma_\theta(x_t, t) = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t I, \quad (3)$$

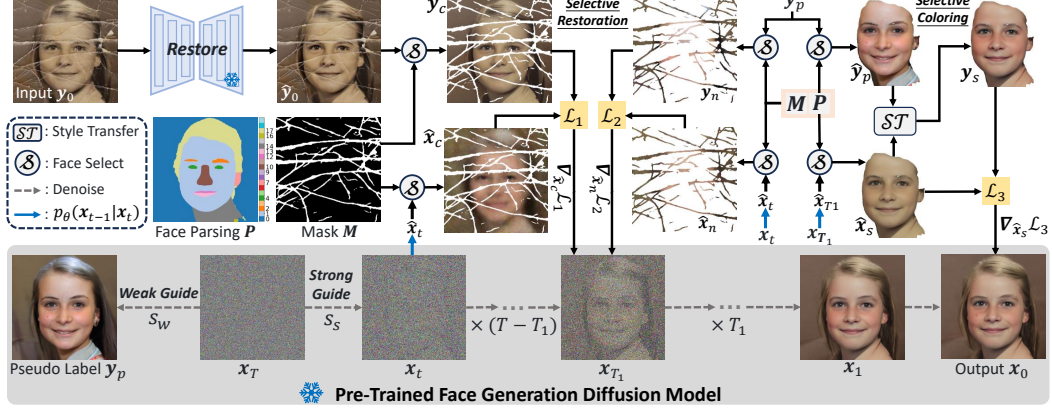


Figure 3: Overview of our method. Face parsing maps P and scratch masks M are estimated from input y_0 using a pre-trained face parsing network [36] and a pre-trained scratch detection network [37]. "Weak Guide" refers to the case shown in Fig. 2 at $s = 1e^{-3}$, i.e., $s_w = 1e^{-3}$.

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \epsilon_\theta(x_t, t) \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \right), \quad \alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=0}^t \alpha_i. \quad (4)$$

In practice, we can directly estimate $\hat{x}_0 = x_t / \sqrt{\bar{\alpha}_t} - \epsilon_\theta(x_t, t) \sqrt{(1 - \bar{\alpha}_t) / \bar{\alpha}_t}$ from ϵ_θ .

Classifier Guidance. Classifier guidance [12] introduces semantic control into the reverse diffusion process by leveraging gradients from a pre-trained classifier $c_\phi(y|x)$. Instead of training a conditional diffusion model, this approach modifies the denoising step of an unconditional model by shifting the predicted mean based on the classifier gradient $\nabla_x \log c_\phi(y|x)$, evaluated at the intermediate prediction. This guides the generation trajectory toward regions in image space that are more likely to match the target semantics:

$$p_{\theta, \phi}(x_{t-1}|x_t, y) \approx \mathcal{N}(\epsilon_\theta(x_t, t) + \Sigma_\theta(x_t, t) \nabla_x \log c_\phi(y|x)|_{x=\mu_\theta(x_t, t)}, \Sigma_\theta(x_t, t)), \quad (5)$$

where the gradient term $g = \nabla_x \log c_\phi(y|x)$ acts as a guidance signal that biases the sampling distribution toward the target class y , $\Sigma_\theta(x_t, t)$ denotes the degree of diffusion of the sample.

3.2 Self-Supervised Selective-Guided Diffusion Model (SSDiff)

As shown in Fig.3, we aim to restore an old face photo LQ input $y_0 \in \mathbb{R}^{C \times H \times W}$, which may suffer from breakage, fading, and blurring, into a HQ face photo output $x_0 \in \mathbb{R}^{C \times H \times W}$. To achieve this, we unify self-supervised loss gradient guidance under a generalized reverse diffusion update, where task-specific losses \mathcal{L}_1 , \mathcal{L}_2 and \mathcal{L}_3 as pseudo classifiers to replace traditional classifier gradients, i.e., $\nabla_x \log c_\phi(y|x) \propto (y - x) \approx -\nabla_x \|y - x\|_2^2$ in Eq.(5). To enable stage-aware supervision, SSDiff integrates two parts: Selective Restoration Guidance, which is applied across the entire reverse trajectory to recover face and breakages, and Selective Coloring Guidance, which is applied in later steps to refine facial color. This temporal separation is empirically motivated by the gradient behavior observed in Fig.4, where image gradient magnitude remains steady in early steps but drops sharply in later steps, indicating a shift from structural restoration to detail refinement. Our design naturally aligns with the coarse-to-fine denoising curriculum [38, 39] in diffusion models, where early reverse steps reconstruct spatial structures while later steps refine details. Combined with Section 3.1, Algorithm 1 summarizes the full sampling process of SSDiff. Below, we elaborate on its two key components: *Selective Restoration Guidance* and *Selective Coloring Guidance*.

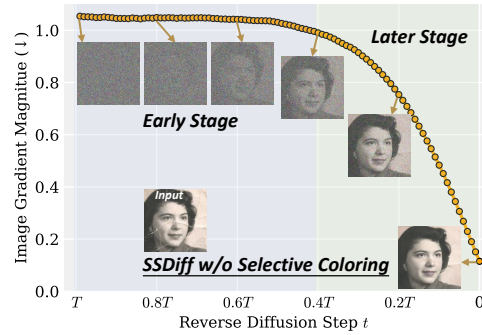


Figure 4: In the early stage, gradient drops are minor, as updates focus on coarse structures like facial contours, which occupy a small region. In the later stage, gradients drop sharply due to refinements in face details, which cover most of a face.

Algorithm 1 : Staged Self-Supervised Sampling in our SSDiff.

Input: Old face photo \mathbf{y}_0 , face parsing map \mathbf{P} and mask \mathbf{M} obtained from \mathbf{y}_0 .
Require: Pre-trained diffusion model $(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$, face selector \mathcal{S} , style transfer \mathcal{ST} , pseudo label \mathbf{y}_p under weak guidance s_w , edge magnitude \mathcal{D} , strong guidance s_s for restoration.
Output: High-quality face photo \mathbf{x}_0 .
 $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$; $\mathbf{y}_c \leftarrow \mathcal{S}(\mathbf{y}_p, \mathbf{M})$, $\mathbf{y}_n \leftarrow \mathcal{S}(\mathbf{y}_p, \mathbf{M}, \mathbf{P})$, $\hat{\mathbf{y}}_p \leftarrow \mathcal{S}(\mathbf{y}_p, \mathbf{P})$
for t from T to 1 **do**
 $\hat{\mathbf{x}}_t \leftarrow \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \sqrt{\frac{1-\alpha_t}{\alpha_t}} \epsilon_\theta(\mathbf{x}_t, t)$
 $\hat{\mathbf{x}}_c \leftarrow \mathcal{S}(\hat{\mathbf{x}}_t, \mathbf{M})$, $\hat{\mathbf{x}}_n \leftarrow \mathcal{S}(\hat{\mathbf{x}}_t, \mathbf{M}, \mathbf{P})$
 repeat
 if $t > T_1$ **then** ▷ Stage I: Restoration Guidance
 $\nabla \mathcal{L} = \nabla_{\hat{\mathbf{x}}_c} \mathcal{L}_1 + \nabla_{\hat{\mathbf{x}}_n} \mathcal{L}_2 = \nabla_{\hat{\mathbf{x}}_c} \|\mathbf{y}_c - \hat{\mathbf{x}}_c\|_2^2 + \nabla_{\hat{\mathbf{x}}_n} \|\mathcal{D}(\mathbf{y}_n) - \mathcal{D}(\hat{\mathbf{x}}_n)\|_2^2$
 $\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t) - s_w \Sigma_\theta(\mathbf{x}_t, t) \nabla \mathcal{L}, \Sigma_\theta(\mathbf{x}_t, t))$
 else ▷ Stage II: Restoration & Coloring Guidance
 $\hat{\mathbf{x}}_{T_1} \leftarrow \frac{1}{\sqrt{\alpha_{T_1}}} \mathbf{x}_{T_1} - \sqrt{\frac{1-\alpha_{T_1}}{\alpha_{T_1}}} \epsilon_\theta(\mathbf{x}_{T_1}, t)$; $\hat{\mathbf{x}}_s \leftarrow \mathcal{S}(\hat{\mathbf{x}}_{T_1}, \mathbf{P})$, $\mathbf{y}_s \leftarrow \mathcal{ST}(\hat{\mathbf{x}}_s, \hat{\mathbf{y}}_p)$
 $\nabla \mathcal{L} = \nabla \mathcal{L} + (\nabla_{\hat{\mathbf{x}}_s} \mathcal{L}_3 = \nabla_{\hat{\mathbf{x}}_s} \|\mathbf{y}_s - \hat{\mathbf{x}}_s\|_2^2)$
 $\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t) - s_s \Sigma_\theta(\mathbf{x}_t, t) \nabla \mathcal{L}, \Sigma_\theta(\mathbf{x}_t, t))$
 end if
 until N steps
 $\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t) - s_s \Sigma_\theta(\mathbf{x}_t, t) \nabla \mathcal{L}, \Sigma_\theta(\mathbf{x}_t, t))$
end for
return \mathbf{x}_0

Selective Restoration Guidance. Selective restoration guidance is applied throughout the reverse diffusion process, with its main role being to accurately restore the face image and fill in the breakage areas. Given the input \mathbf{y}_0 , we first use an arbitrary pre-trained face restoration model [9, 7] to obtain an initially restored semantics, resulting in an output $\hat{\mathbf{y}}_0$ that remains broken. We treat non-broken regions \mathbf{y}_c of $\hat{\mathbf{y}}_0$ as plausible and use them to guide the frozen pre-trained face generation model to perform reverse diffusion in non-broken regions of faces. We define this part loss \mathcal{L}_1 between non-masked face regions \mathbf{y}_c of $\hat{\mathbf{y}}_0$ and the reverse diffusion-guided output $\hat{\mathbf{x}}_c$, formulated as $\mathcal{L}_1 = \|(1 - \mathbf{M}) \odot (\mathbf{y}_c - \hat{\mathbf{x}}_c)\|_2^2$. Its loss gradient $\nabla_{\hat{\mathbf{x}}_c} \mathcal{L}_1$ provides supervision to the reverse diffusion process, ensuring faithful restoration in non-broken face regions. $\mathbf{M} \in \{0, 1\}$ is a mask obtained by a pre-trained scratch detector [37] from \mathbf{y}_0 , \odot is a Hadamard product.

For breakage regions, previous methods [14, 16] typically set gradients to zero and rely on the pre-trained face generation model to complete them automatically. However, this often results in incomplete completion, especially in areas such as background and skin. To address this issue, we introduce a smoothness-guided loss \mathcal{L}_2 which explicitly enforces surface smoothness consistency between pseudo-labels and guided features within breakage regions, mitigating color inconsistency caused by direct guidance. We first use face parsing maps [36] \mathbf{P} and scratch masks [37] \mathbf{M} derived from input \mathbf{y}_0 to select plausible breakage guide regions $\hat{\mathbf{x}}_n$ and \mathbf{y}_n within breakage parts from reverse diffusion and pseudo-labels. This selection focuses on a part of the breakage regions $\mathbf{M}_{guide} \subset \mathbf{M}$ such as background, skin, and hair, while excluding semantically sensitive areas (*e.g.*, eyes, eyebrows, mouth) that may differ from original identity. Furthermore, we extend (\mathcal{E}) the mask of breakages along the horizontal and vertical dimensions of \mathbf{M}_{guide} to alleviate incomplete scratch detection [37]. The loss \mathcal{L}_2 for the breakage regions is defined as:

$$\mathcal{L}_2 = \|\mathcal{D}(\mathbf{y}_n, \mathcal{E}(\mathbf{M}_{guide})) - \mathcal{D}(\hat{\mathbf{x}}_n, \mathcal{E}(\mathbf{M}_{guide}))\|_2^2, \quad (6)$$

where the function $\mathcal{D}(\mathbf{y}, \mathbf{M}) \in \mathbb{R}^{H \times W}$ computes the average per-pixel gradient magnitude within the masked area. Let $\mathbf{y}_{i,j}$ denote the pixel value at the spatial location (i, j) , then \mathcal{D} is defined as:

$$\mathcal{D}(\mathbf{y}, \mathbf{M}) = |\mathbf{y}_{i,j} - \mathbf{y}_{i,j+1}| (\mathbf{M}_{i,j} \odot \mathbf{M}_{i,j+1}) + |\mathbf{y}_{i,j} - \mathbf{y}_{i+1,j}| (\mathbf{M}_{i,j} \odot \mathbf{M}_{i+1,j}). \quad (7)$$

The above operations ensure smooth consistency of gradients in breakage regions to restore breakages. Then, we leverage the loss gradient $\nabla_{\hat{\mathbf{x}}_n} \mathcal{L}_2$ as a guidance signal during the reverse diffusion process, enabling our model to actively refine breakage regions. For the remaining small portion of breakage, we still allow models to complete them automatically, as pre-trained face generation models generally perform well in restoring semantically critical facial regions such as eyes, eyebrows, and mouth.

Table 1: Quantitative comparison of old-photo face images, which are categorized as simple, medium, and hard based on degradation degree. **Bold** and underlined indicate best and second best results. *Our Appendix provides results on public test sets, including WebPhoto-Test [6], and CelebA-Chlid [6].*

Type	Metric	GAN-based		Diffusion-based (Learning)		Diffusion-based (Train-free)		
		BOPB [37]	Code [9]	DiffFace [10]	DiffBIR [11]	DDNM [14]	PGDiff [16]	Ours
Simple	FID↓	177.19	151.01	<u>132.14</u>	163.47	177.82	136.73	129.13
	BRISQUE↓	22.76	6.69	23.13	17.08	51.74	8.77	<u>6.71</u>
	TOPIQ↑	0.4444	0.6005	0.5954	0.5474	0.3976	<u>0.6216</u>	0.6494
	MAN-IQA↑	0.3124	<u>0.3876</u>	0.3676	0.3568	0.2742	0.3646	0.4005
Medium	FID↓	176.11	189.37	<u>142.22</u>	198.84	173.96	142.56	128.31
	BRISQUE↓	22.54	<u>8.29</u>	23.45	12.27	52.78	10.22	7.29
	TOPIQ↑	0.4337	<u>0.6123</u>	0.5917	0.5684	0.3929	0.6060	0.6412
	MAN-IQA↑	0.2975	<u>0.3789</u>	0.3511	0.3509	0.2721	0.3456	0.3949
Hard	FID↓	218.35	173.89	<u>145.85</u>	200.38	203.13	166.42	122.59
	BRISQUE↓	28.91	<u>10.51</u>	26.41	14.36	54.17	13.60	8.95
	TOPIQ↑	0.3865	0.5670	0.5450	0.5424	0.3447	<u>0.5747</u>	0.5977
	MAN-IQA↑	0.2454	<u>0.3379</u>	0.3159	0.3140	0.2243	0.3159	0.3575

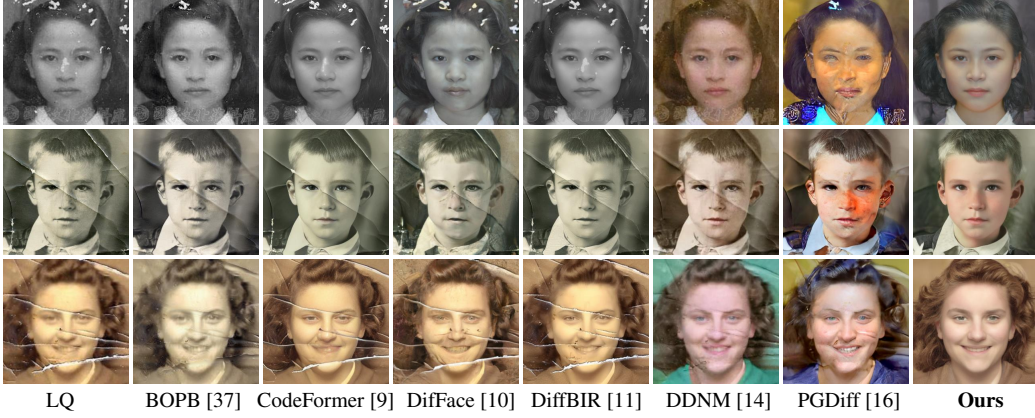


Figure 5: Qualitative comparisons of old-photo face images with existing methods.

Selective Coloring Guidance. Selective coloring guidance is applied only after facial identity features such as eyes, nose, and mouth have stabilized during the reverse diffusion process. At this stage, as shown in Fig.4, identity features are essentially fixed, and remaining steps focus on refining details, which reduces the risk of introducing identity inconsistency through color guidance. Given intermediate images x_{T_1} from the reverse diffusion and pseudo-label y_p , we first extract regions related to the skin, including face and neck, using face parsing maps M , resulting in \hat{x}_s and \hat{y}_p . However, direct supervision from \hat{y}_p to \hat{x}_s is suboptimal due to potential variations in facial components. To address this, we employ a pre-trained color style transfer network [40] to transfer the color style of \hat{y}_p to \hat{x}_s to obtain a color-adapted intermediate result y_s , which subsequently guides the remaining reverse diffusion steps for faithful and coherent face coloring. To supervise this process, we define a color consistency loss \mathcal{L}_3 between \hat{x}_s and y_s , which enforces global color coherence while allowing structural flexibility in regions sensitive to identity. Specifically, \mathcal{L}_3 is defined as:

$$\mathcal{L}_3 = \|y_s - \hat{x}_s\|_2^2, \quad y_s = \text{ColorTransfer}(\hat{x}_s, \hat{y}_p). \quad (8)$$

In this phase, by using the loss gradient $\nabla_{\hat{x}_s} \mathcal{L}_3$ to guide tones and shadows of skin regions, and combining this with the earlier selective restoration guidance, facial coloring realism, and identity preservation are improved while supporting further fine-grained facial detail reconstruction.

Overall Guidance. Our guidance strategy is staged to align with the progression of the reverse diffusion. In the early phase of denoising, where $t \in [T_1, T]$, models focus solely on structural restoration. We apply guidance losses $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, where \mathcal{L}_1 encourages fidelity in intact regions and \mathcal{L}_2 enforces smoothness in breakage regions. Once the facial identity stabilizes, in the later phase $t \in [0, T_1]$, we introduce a selective coloring loss \mathcal{L}_3 and formulate the full loss as $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$.

In this framework, pseudo-labels generated under a weak guidance scale s_w are used to supervise the actual restoration process, which is conducted under a stronger guidance scale s_s . To ensure coherent interaction between the two stages, we adopt a joint optimization perspective:

$$\mathcal{L} = \arg \min \left(\mathbb{E} \|y_c(s_w^*) - \hat{x}_c(s_s^*)\|_2^2 + \mathbb{E} \|y_n(s_w^*) - \hat{x}_n(s_s^*)\|_2^2 + \mathbb{E} \|y_s(s_w^*) - \hat{x}_s(s_s^*)\|_2^2 \right). \quad (9)$$

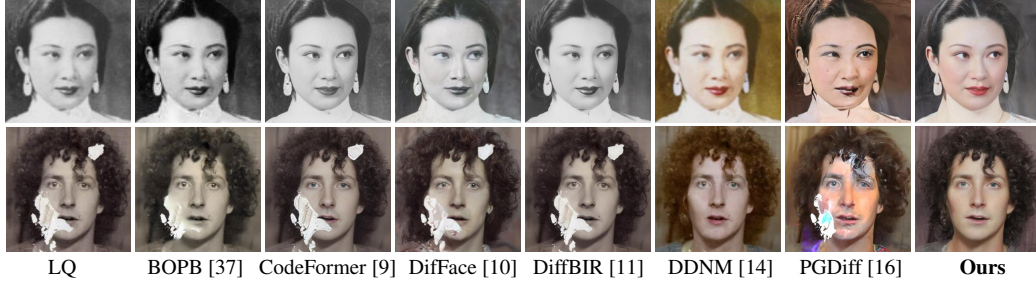


Figure 6: Quantitative comparisons of results on no breakage and large-region breakage images.

Theoretically, an optimal pair of weak and strong guidance scales (s_w^*, s_s^*) achieves an equilibrium, where neither under- nor over-guidance dominates, enabling effective perceptual restoration.

4 Experiments

4.1 Experiments Setting

Datasets and Metrics. Our method is **training-free** and does not require a training dataset. For evaluation, we randomly collect 300 old face photographs from the Internet as our benchmark, called VintageFace. All images are cropped and aligned to 512×512 using the open-source FaceLib library¹. We further categorize them into three levels of degradation: simple, medium, and hard, based on image quality, 100 images each. Details of the benchmark are provided in our *Appendix*. Since ground-truth is unavailable, we follow previous works [10, 11] and adapt no-reference metrics, like FID [41], BRISQUE [42], TOPIQ [43], and MAN-IQA [44] to evaluate perceptual quality. Furthermore, we provide qualitative results and CLIP-based [45] identity distance to assess restoration fidelity.

Implementation Details. We use the pre-trained real-time model BiSeNet [36] and the scratch detection model from [37] to obtain face parsing maps and scratch masks from inputs, respectively. Our pre-trained diffusion model is an unconditional denoising network trained on FFHQ [17] datasets, which learns to reconstruct high-quality faces from pure noise over $T = 1000$ steps. We restore breakage facial regions during the first 600 steps, and apply color migration in the remaining 400 steps ($T_1 = 400$). The strong gradient factor of our SSDiff is $s_s = 3.5e^{-3}$. PGDiff [16] is also set to this value for fairness. For the style transfer process shown in Fig. 3, we adopt a pre-trained lightweight model CAP-VSTNet [40], without using its built-in segmentation module for boundary invariance. All experiments are implemented in PyTorch framework on an NVIDIA RTX 4090 GPU.

4.2 Comparisons with Existing Methods

We compare our SSDiff with BOPB [37], which is specialized for old photo restoration; CodeFormer [9], DiffFace [10], and DiffBIR [11], which are designed for blind face restoration; and DDNM [14] and PGDiff [16], which guide pre-trained diffusion for zero-shot old-photo restoration.

As shown in Table 1, a quantitative comparison of our SSDiff with the above methods demonstrates that our approach not only significantly outperforms existing methods in FID, BRISQUE, and TOPIQ, which reflect image features quality, but also achieves more natural image quality in the cross-modal human perceptual metric MAN-IQA. However, these metrics alone do not fully capture the fidelity of the results. Therefore, Figure. 5 presents qualitative comparisons across three types of old face photographs: simple, medium, and hard. Our method produces face images with natural color, clear texture, and minimal damage, while maintaining high fidelity and preserving the identity of the input face images. Furthermore, we provide more old-face photo restoration results in our *Appendix*.

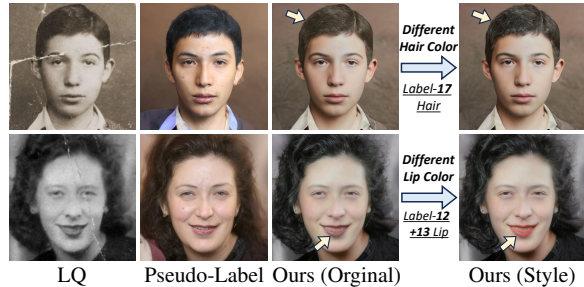


Figure 7: Our method allows stylized restoration of facial components specified by reference to pseudo-labels on inputs.

¹Facelib: <https://github.com/sajjjadayobi/FaceLib>

Evaluation on Diverse Old Face Photos. As shown in Fig. 6, we validate the generality of our SSDiff on a broader set of old face photos. This includes no-breakage images where the mask is all zeros, and large-region breakage images where the mask indicates missing regions to be filled. In all scenarios, our method performs robustly, achieving high-quality face coloring and region completion while restoring sharp facial contours. It significantly outperforms existing methods that are limited to a single function, such as coloring, completion, or suffer from overall poor results.

Region-Specific Stylized Restoration. Since the face parsing map is insensitive to scratches and highly structured features of faces, we can reliably select specific facial components for targeted restoration using fixed semantic labels. As shown in Fig. 7, our restored faces exhibit unnatural violet around the lips. This issue can be mitigated by explicitly selecting the lip region via the parsing map and increasing the guidance strength for this component. Furthermore, the same strategy can be applied to adjust hair color, for example, modifying yellowish hair to a dark tone by reference to pseudo-labels. This property allows us to fine-tune facial local components with suboptimal details and achieve more visually appealing face restoration results.

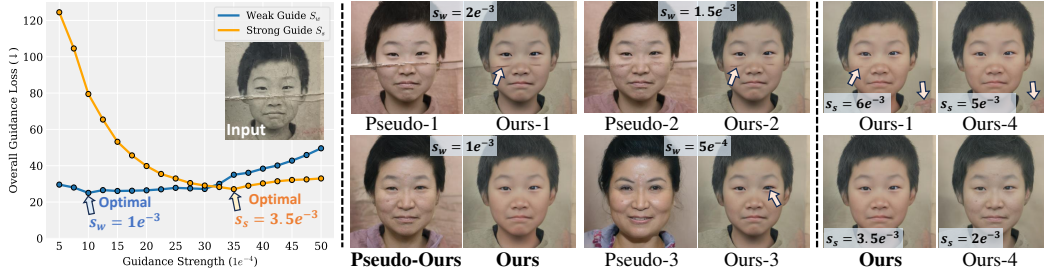


Figure 8: Ablation studies on the effect of different weak guidance factors s_w for generating pseudo labels and strong factors s_s in restoration on results. When $s_s = 1e-3$ and $s_w = 3.5e-3$, our results show the best guidance loss, facial tones, facial details, and fidelity of identity.

4.3 Ablation Studies

Self-Supervised Strategy. When the self-supervised strategy is removed, namely when the pseudo-label y_p is no longer used, we disable both the selective restoration guidance and selective coloring guidance associated with y_p in breakage regions. As shown in Table 2, results show a significant drop in visual quality without the self-supervision strategy. Furthermore, based on Fig. 2, we explore the appropriate range for weak and strong guide factors s_w and s_s , which determines the reliability of generated pseudo-labels and results. As shown in Fig. 8, an excessively large s_w leads to incomplete removal of pseudo-label scratches, leaving visible artifacts in restorations. Conversely, an overly small s causes some identity feature mismatch, resulting in distorted eyes and dark facial tones. Thus, we set $s_w = 1e-3$ and $s_s = 3.5e-3$ to obtain best results.

Selective Guidance. We first evaluate the effectiveness of the three guided losses: \mathcal{L}_1 , \mathcal{L}_2 , and \mathcal{L}_3 . As shown in Table 3, all three gradient-based losses contribute to improved restoration. As shown in Fig. 10, visualization results further reveal that \mathcal{L}_1 and \mathcal{L}_2 enhance the sharpness of facial structures, reducing visible breakage artifacts in old photos. In contrast, \mathcal{L}_3 plays a critical role in facial coloring, ensuring a natural tone in results. Furthermore, since \mathcal{L}_3 is applied after the facial structure has stabilized, we investigate the appropriate value for T_1 . As shown in Fig. 9, letting T denote the total number of diffusion steps, setting $T_1 = 0.4T$ (fluctuating with degradations) stabilizes the facial structure and enables the guided coloring to produce more natural skin tones.

Next, we justify the design of selective restoration. The pseudo-label y_p is not used to guide high-semantic facial components such as the eyes, nose, or mouth, as these regions are identity-sensitive and pseudo-label inconsistencies may degrade fidelity, as shown in Fig. 10. For selective coloring,

Table 2: Ablation studies of our self-supervised strategy on “medium” type data.

Self-Supervise	FID↓	TOPIQ↑	BRISQUE↓
✗	145.7	0.618	7.92
✓	128.3	0.641	7.29

Table 3: Ablation studies of our selective guidance strategy on “medium” type data.

Guidance	FID↓	TOPIQ↑	BRISQUE↓
Ours	128.3	0.641	7.29
w/o \mathcal{L}_1	142.8	0.602	8.18
w/o \mathcal{L}_2	132.1	0.630	7.43
w/o \mathcal{L}_3	136.3	0.624	7.65

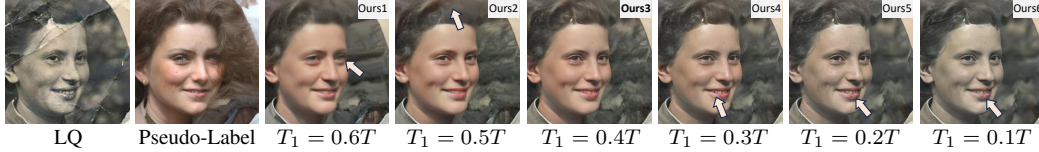


Figure 9: Ablation study on different T_1 values in reverse diffusion, *i.e.*, when to start guiding face coloring, on restoration results. Best facial tones and details are observed at $T_1 = 0.4T$.

we examine why guidance is applied only to skin regions, rather than full faces, or why we don't just color transfer. As shown in Fig. 10, due to the limited performance of the pre-trained color transfer network and large color variations of overall faces, the overall direct migration or full face guidance leads to unsatisfactory results. Therefore, we restrict color guidance to skin regions.

Fidelity of Restoration. Due to the presence of severely degraded facial features in a part of inputs, it is challenging to reliably evaluate identity consistency across all samples. To ensure meaningful measurement, we select 50 LQ inputs with relatively well-preserved facial features and compute CLIP-based [45] identity feature distance between the restorations and inputs. As shown in Table 4, our method achieves low identity distances, indicating good fidelity in restoring identity-relevant features.

Table 4: Quantitative comparison on identity distances (\downarrow) with LQ inputs.

BOPB [9]	DiffFace [10]	DiffBIR [11]
0.2348	0.2451	0.2223
DDNM [14]	PGDiff [16]	Ours
0.2386	0.2859	<u>0.2280</u>

4.4 Limitations and Future Works

Our method improves the robustness of face restoration for old photos and performs well even under complex breakage. However, as shown in Fig. 11, when large stained regions are present, our method may mistakenly restore these stains as part of the face, such as skin, leading to unclear results. This issue stems from the diffusion model's tendency to over-rely on surrounding context in severely degraded regions, leading to incorrect restorations when visual cues are ambiguous. In the future, we plan to incorporate fine-grained region-level annotations for such cases to reduce failure modes.

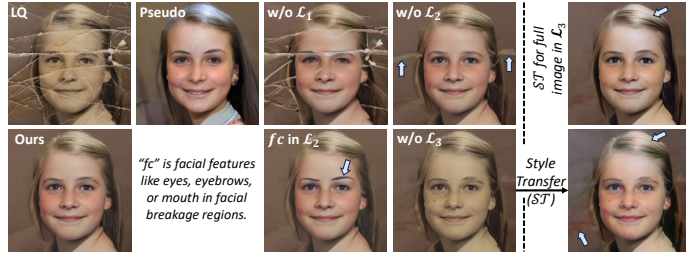


Figure 10: **(Left)** Ablations of \mathcal{L}_1 , \mathcal{L}_2 , \mathcal{L}_3 , and why we can't guide "fc" in breakages in \mathcal{L}_2 . **(Right)** Ablations of color-guided areas, and why don't we just color style transfer (ST) instead of guiding coloring?

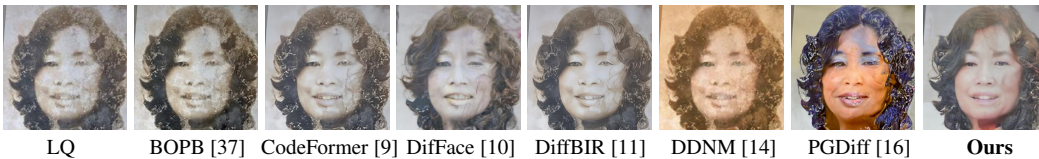


Figure 11: Our Failures. When large stains appear, they may be misinterpreted as a part of facial or hair structures and color by our method, leading to unnatural artifacts and distorted local details.

5 Conclusion

In this work, we present SSDiff, a specialized framework for old-photo face restoration that leverages the generative capacity of pre-trained diffusion models through self-supervised pseudo-reference guidance. Unlike existing approaches that rely on explicit degradation models or pre-defined high-quality face attributes, SSDiff guides the denoising process using features selected from generated pseudo-references and facial geometry priors such as parsing maps and scratch masks at different sampling stages. This design enables the selective restoration of plausible facial regions of old photos while maintaining high visual fidelity. Extensive experiments demonstrate that SSDiff not only handles complex old-photo degradations effectively but also enables region-specific stylization of faces, offering a flexible and robust solution for challenging old-photo face restoration tasks.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62472044, U24B20155, 62225601, U23B2052), Beijing-Tianjin-Hebei Basic Research Funding Program No. F2024502017, Hebei Natural Science Foundation Project No. 242Q0101Z, Beijing Natural Science Foundation Project No. L242025.

References

- [1] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *ECCV*, pages 111–129. Springer, 2024.
- [2] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pages 25456–25467, 2024.
- [3] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *ECCV*, pages 1–21. Springer, 2024.
- [4] Wenjie Li, Mei Wang, Kai Zhang, Juncheng Li, Xiaoming Li, Yuhang Zhang, Guangwei Gao, Weihong Deng, and Chia-Wen Lin. Survey on deep face restoration: From non-blind to blind and beyond. *arXiv preprint arXiv:2309.15490*, 2023.
- [5] Wenjie Li, Heng Guo, Xuannan Liu, Kongming Liang, Jiani Hu, Zhanyu Ma, and Jun Guo. Efficient face super-resolution via wavelet-based feature enhancement network. In *ACM MM*, pages 4515–4523, 2024.
- [6] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pages 9168–9178, 2021.
- [7] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, pages 672–681, 2021.
- [8] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, pages 126–143. Springer, 2022.
- [9] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, volume 35, pages 30599–30611, 2022.
- [10] Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [11] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pages 430–448. Springer, 2024.
- [12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021.
- [13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.
- [14] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023.
- [15] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023.
- [16] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdif: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023.
- [17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.
- [18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014.
- [19] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, pages 14245–14254, 2021.

- [20] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *CVPR*, pages 7662–7671, 2022.
- [21] Zhengrui Chen, Liying Lu, Ziyang Yuan, Yiming Zhu, Yu Li, Chun Yuan, and Weihong Deng. Blind face restoration under extreme conditions: leveraging 3d-2d prior fusion for superior structural and texture recovery. In *AAAI*, volume 38, pages 1263–1271, 2024.
- [22] Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin CK Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. *ICLR*, 2024.
- [23] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021.
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020.
- [25] Zhixin Wang, Ziyang Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *CVPR*, pages 1704–1713, 2023.
- [26] Tuomas Varanka, Tapani Toivonen, Soumya Tripathy, Guoying Zhao, and Erman Acar. Pfstorer: Personalized face restoration and super-resolution. In *CVPR*, pages 2372–2381, 2024.
- [27] Lianxin Xie, Csbingbing Zheng, Wen Xue, Le Jiang, Cheng Liu, Si Wu, and Hau San Wong. Learning degradation-unaware representation with prior-based latent transformations for blind face restoration. In *CVPR*, pages 9120–9129, 2024.
- [28] Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. In *CVPR*, pages 6583–6592, 2024.
- [29] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024.
- [30] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, volume 37, pages 92529–92553, 2024.
- [31] Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. Osdface: One-step diffusion model for face restoration. In *CVPR*, 2025.
- [32] Bahjat Kavar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, volume 35, pages 23593–23606, 2022.
- [33] Kanchana Vaishnavi Gandikota and Paramanand Chandramouli. Text-guided explorable image super-resolution. In *CVPR*, pages 25900–25911, 2024.
- [34] Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Wenxue Li, Lei Zhu, and Xinghao Ding. Agldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. In *AAAI*, volume 39, pages 5307–5315, 2025.
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.
- [36] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018.
- [37] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, pages 2747–2757, 2020.
- [38] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009.
- [39] Erdi Sayar, Giovanni Iacca, Ozgur S Oguz, and Alois Knoll. Diffusion-based curriculum reinforcement learning. In *NeurIPS*, volume 37, pages 97587–97617, 2024.
- [40] Linfeng Wen, Chengying Gao, and Changqing Zou. Cap-vstnet: Content affinity preserved versatile style transfer. In *CVPR*, pages 18300–18309, 2023.
- [41] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

- [42] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.
- [43] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024.
- [44] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniq: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022.
- [45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.
- [46] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.
- [47] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021.
- [48] Wenjie Li, Heng Guo, Yuefeng Hou, Guangwei Gao, and Zhanyu Ma. Dual-domain modulation network for lightweight image super-resolution. *IEEE Transactions on Multimedia*, 2025.
- [49] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope by emphasizing our contributions and experimental results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: In Sec. 4.4 of the main paper, we discuss the limitations of the proposed method.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Sec. 4, we extensively delineate all the details of our methodology, facilitating the reproduction of all experimental content based on our method description.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: To prevent potential plagiarism before the paper is accepted, we assure that upon acceptance, we will promptly provide a link to the open-source code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Sec. 4.1, we provide a detailed description of the experimental hyperparameter settings, evaluation metrics, and the computational resources required.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This is not the experimental direction of this paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Sec. 4.1, we provide a detailed description of the experimental hyperparameter settings, evaluation metrics, and the computational resources required.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Yes, our research fully adheres to the NeurIPS Code of Ethics guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our method aids old photos restoration.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve the release of models or datasets that pose a high risk of misuse. The collection of old photographs of faces are taken from the Internet.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available methods, all of which are properly credited in the main text.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not include new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This is not the experimental direction of this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This is not the experimental direction of this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our paper does not include the usage of LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix

A Details of Our Benchmark

First, as illustrated in Fig. S12, we present our criteria for categorizing our proposed VintageFace benchmark into simple, medium, and severe degradation levels. Specifically, we employ a frozen CLIP model to compute the similarity between each old photo and a textual description of degradation severity. Images are then ranked by their similarity scores and assigned to categories accordingly. To ensure accuracy, we further manually corrected a small number of misclassified samples.

Second, as shown in Fig.S13, we display representative examples from each degradation level in the VintageFace benchmark. These examples demonstrate varying degrees of blurring, fading, and structural damage, and are largely consistent with the classification criteria established in Fig.S12.

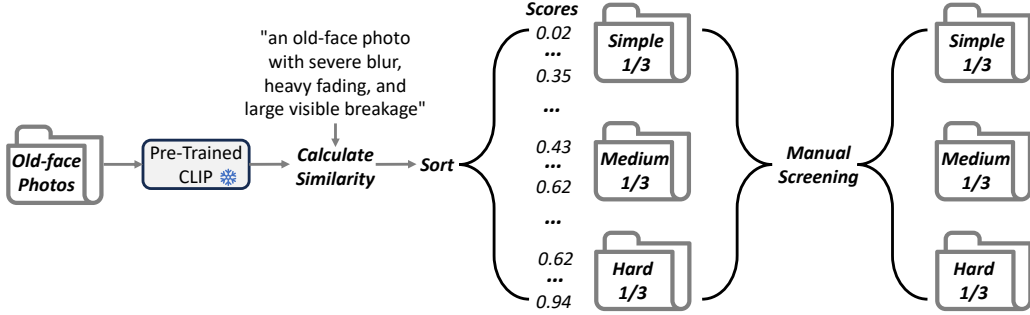


Figure S12: Method for categorizing degradation types into simple, medium, and hard levels in our benchmark VintageFace for testing.



Figure S13: A showcase of representative facial images with varying degradation types from our VintageFace benchmark. The benchmark includes faces across diverse genders, ages, and ethnicities.

B More Comparisons

First, we provide additional visual results in the appendix (*e.g.*, Fig.S16, Fig.S17, and Fig.S18) to complement the main text. These figures showcase the restoration performance on old face

Table S5: Quantitative comparison on real-world BFR benchmarks that contain old face photos, like WebPhoto-Test and CelebA-Child. **Bold** and underlined indicate best and second best results.

Dataset	Metric	GAN-based		Diffusion-based (Learning)		Diffusion-based (Train-free)		
		GPEN [7]	Code [9]	DiffFace [10]	DiffBIR [11]	DDNM [14]	PGDiff [16]	Ours
WebPhoto-Test	FID↓	101.3	83.2	89.1	91.8	165.6	96.1	<u>86.9</u>
	NIQE↓	6.326	<u>4.705</u>	4.831	6.069	9.259	5.117	4.406
CelebA-Child	FID↓	<u>113.0</u>	116.2	113.1	118.9	151.2	121.0	112.7
	NIQE↓	4.945	4.983	<u>4.818</u>	5.549	6.576	5.070	4.524

photographs across diverse genders, ethnicities (Asian, European-American, and Indian), and age groups. Our method effectively balances perceptual quality and identity preservation: the restored images exhibit minimal artifacts or breakage while maintaining faithful facial identity.

Second, we observe that widely used blind face restoration benchmarks, such as LFW-Test and CelebChild, also include a substantial number of old face photos. However, these differ from our dataset in that they primarily exhibit blurring, with no significant structural damage and limited fading. To demonstrate the generalization and effectiveness of our method, we compare it with state-of-the-art approaches on these two benchmarks. Following previous works [9, 16], we adopt FID and NIQE as evaluation metrics. As shown in Table S5, our method achieves good quantitative results on both benchmarks. Furthermore, we provide visual comparisons in Fig. S19, which reveal that our method not only effectively addresses blurring but also excels at restoring facial color. This perceptual advantage, particularly in color restoration, is not fully captured by quantitative metrics.

Thirdly, VintageFace primarily consists of frontal photos, as portrait photos decades ago were typically studio-based, focusing on clearly capturing facial features, making profile shots rare. Additionally, eyewear was less common, resulting in fewer photos with glasses. Consequently, our data has fewer such samples. Nevertheless, as shown in Fig. S14, SSDiff performs robustly across these scenarios, including glasses, profile shots, and severe degradations, consistently yielding favorable results.

C More Ablation

Robustness of Pre-trained Networks. Our SSDiff is generally robust to inaccuracies in external components (face parsing, scratch detection, style transfer). These networks only provide coarse directional signals during reverse diffusion, similar to classifier-guided diffusion, and are not strict constraints. As long as the guidance is not severely misleading, the strong generative prior of the frozen diffusion model dominates reconstruction. To quantify this robustness, as shown in the Table S6, Table S7, and Table S8, we conduct ablations on the Medium type subset:

For parsing map networks, we introduce inaccuracies by replacing the original parsing maps with pseudo-label parsing maps of different strengths s , where the resulting errors are even larger than those observed in parsing maps under severe degradations (79% IOU). The resulting IoU with the original parsing map is: for $s=2.5e-4$, IoU=76% (24% discrepancy); for $s=1e-4$, IoU=70% (30% discrepancy); For scratch detection networks, we randomly flip 10%, 20%, and 30% of the masks of breakage regions to simulate a situation where some of the breakages have not been detected; For style transfer networks, we weaken the style transfer guidance by reducing the style factor α from 0 to 0.1 and 0.2, slightly affecting color and content.

Table S6: Parsing Maps.				Table S7: Scratch Masks.				Table S8: Style Transfer.		
	Ours	76%	70%	Ours	10%flip	20%flip	30%flip	Ours ($\alpha=0$)	$\alpha=0.1$	$\alpha=0.2$
FID(↓)	128.3	131.1	133.4	128.3	129.2	130.7	132.5	128.3	129.2	128.8
MAN-IQA(↑)	0.395	0.391	0.382	0.395	0.391	0.379	0.381	0.395	0.396	0.392
Face Sim.(↑)	1	0.985	0.955	1	0.974	0.953	0.937	1	0.977	0.959

Here, Face Similarity (range [0, 1]) denotes the cosine similarity between features (extracted with ArcFace [46]) of the perturbed restoration and the original restoration (Ours). These results show that errors in the pre-trained networks are not severe, and the strong generative prior of the diffusion model can propagate the correct cues to other regions, preventing significant performance drops. This demonstrates that SSDiff is robust to these pre-trained networks.

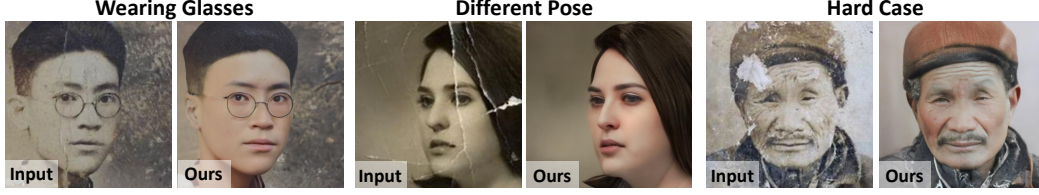


Figure S14: Visualization of SSDiff under wearing glasses, different poses, and severe degradation.

Latency. Our method is built upon existing pre-trained diffusion face generation models, where additional inference overhead mainly comes from four components: a simple restore, a face parsing network, a scratch detection network, and a style transfer network. All components are lightweight and are executed at a single denoising step rather than throughout the entire process. Moreover, except for the style migration network, the other three can optionally be pre-processed offline. When all four components are executed online, the average latency for processing a single old face photo is 95 ms on an NVIDIA GeForce RTX 4090. If the three offline-optional components are pre-processed, the average latency is reduced to 24 ms. In contrast, PGDiff [16] requires semantic information extraction at each denoising step, introducing a latency of about 10 s. Therefore, our method only introduces minimal latency to existing pre-trained generation diffusion frameworks.

Computational Cost. As shown in Table S9, we further compare the number of Params, FLOPs, and inference time of our method with existing diffusion-based face restoration methods [10, 11, 16]. We let the restore be performed offline; our method performs excellently. Our method is smaller in terms of the Params and FLOPs counts, especially compared to stable diffusion-based methods like DiffBIR [11].

Table S9: Quantitative comparison on computational costs with existing diffusion-based BFR methods.

Costs	DiffFace [10]	DiffBIR [11]	PGDiff [16]	Ours
Params↓	175.4M	1717M	47.7M	45.4M
FLOPs↓	268.8G	24234G	127.5G	120.7G

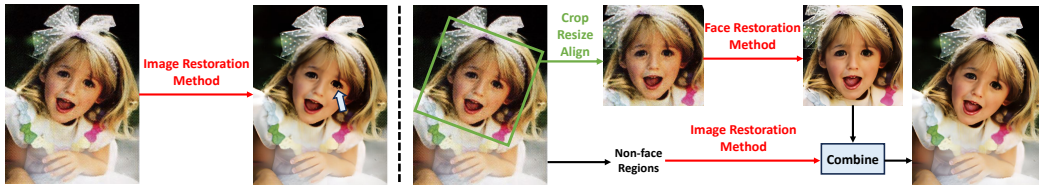


Figure S15: **(Left)** Face images are highly sensitive to artifacts, directly restoring photos containing faces with image restoration methods may results in visually disturbing results. **(Right)** A common strategy involves cropping and aligning facial regions, followed by restoration using face restoration methods, while non-facial regions are enhanced with image restoration methods to ensure visual perception. *Therefore, old photo face restoration holds practical value for old photo restoration.*

D Necessity of Old-Photo Face Restoration

While general image restoration methods [1] aim to restore the entire image holistically, we argue that dedicated face restoration [6] is necessary and beneficial, especially in the context of severely degraded old portraits. As shown in Fig. S15, directly applying general real-world image restoration models [47] to facial regions may introduce noticeable artifacts, even when these methods perform reasonably well on background areas. This is because facial regions are typically small in size, contain rich structural priors (e.g., eyes, nose, mouth), and are highly sensitive to local distortions. Artifacts in these regions are particularly perceptible and detrimental to human perception.

Similarly, old face photos suffer from unique degradation patterns such as heavy blurring, fading, and structural damage. Applying global restoration methods [47, 48] to these faces without region-specific modeling frequently leads to distorted identity features or unnatural textures. Therefore, we advocate for face-specific old photo restoration approaches [14, 16] that focus on preserving facial identity and fidelity, while allowing general old photo restoration techniques [37] to handle the surrounding non-facial regions. This targeted strategy ensures high-quality restoration where perceptual sensitivity is highest and complements broader restoration pipelines. Therefore, we respectfully believe the task of old photo face restoration holds specifically practical value.

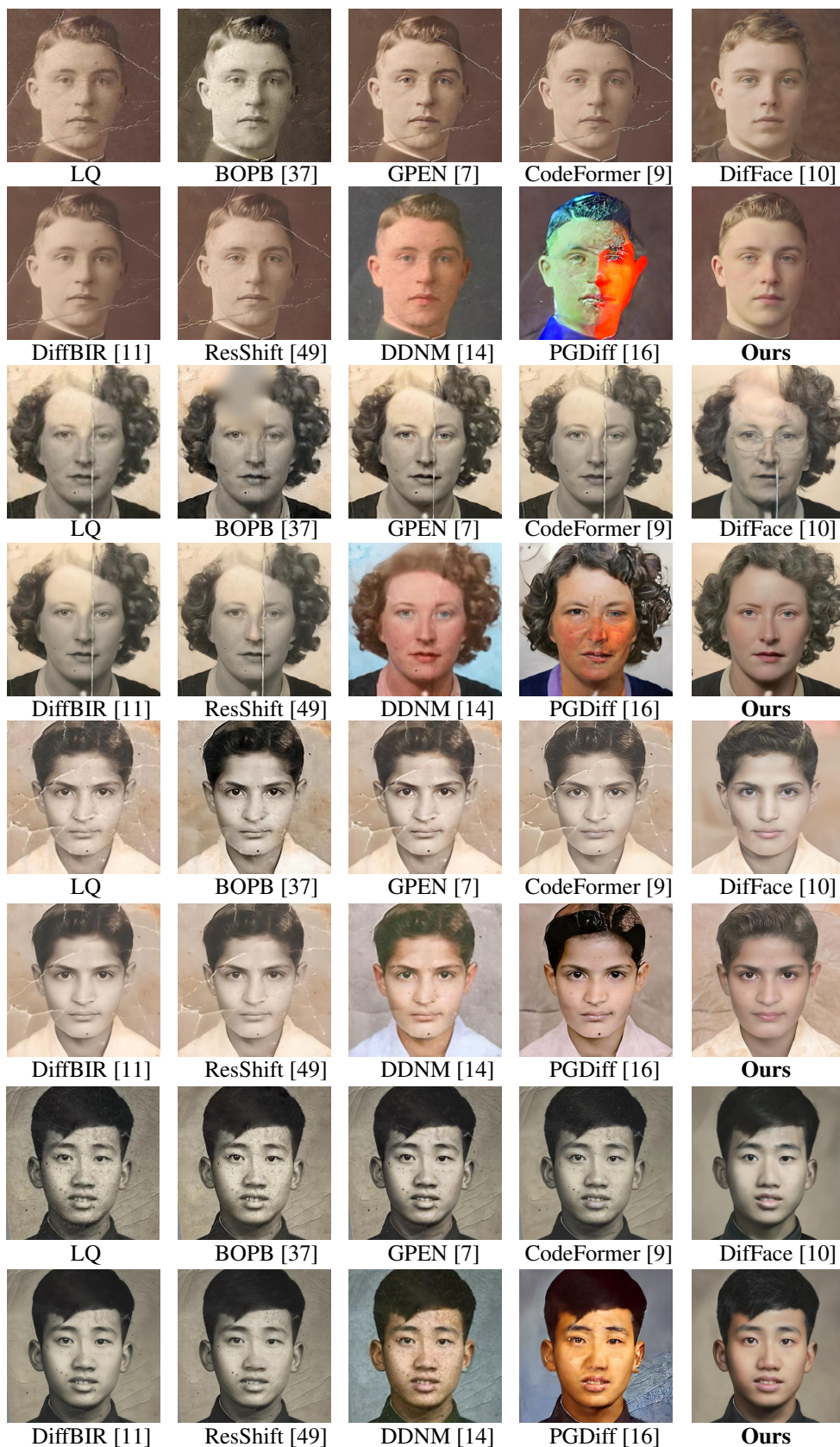


Figure S16: Qualitative comparisons with existing methods on our VintageFace.

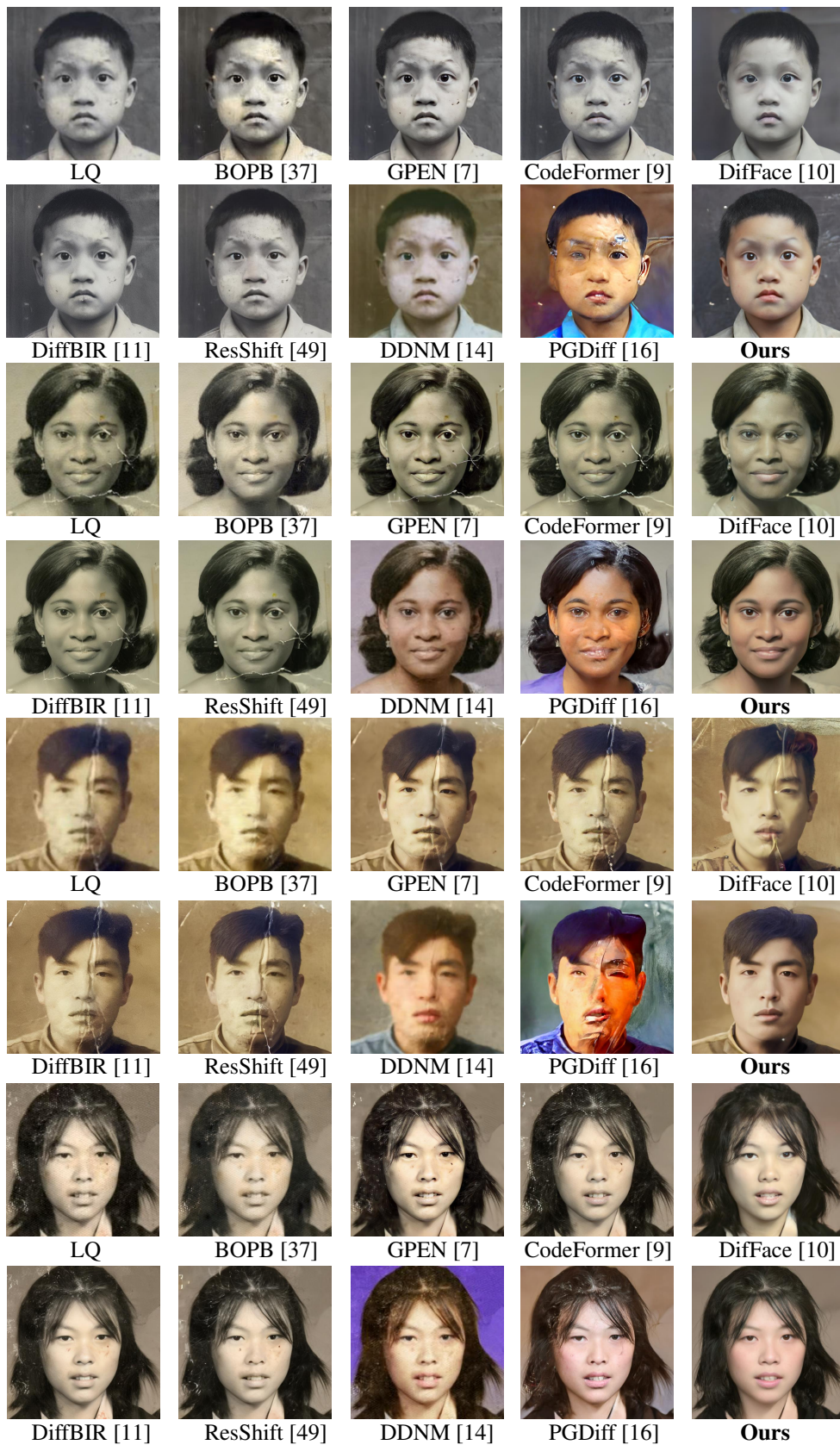


Figure S17: Qualitative comparisons with existing methods on our VintageFace.



Figure S18: Qualitative comparisons with existing methods on our VintageFace.



Figure S19: Qualitative comparisons with existing methods on WebPhoto-Test and CelebA-Child.