CURRICULUM-DRIVEN DEGRADATION-AWARE DIFFUSION TRANSFORMERS FOR REAL-WORLD OLD-PHOTO FACE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world old-photo face restoration is severely hindered by the complexity of unknown physical damages and the synthetic-to-real domain gap. While recent Diffusion Transformer (DiT) architectures exhibit powerful generative priors, their static training paradigms treat all degradation levels equally. This inherently leads to a sub-optimal trade-off: over-smoothing on mild degradations and severe identity hallucination on complex physical damages. In this paper, we propose CurrDiT, a curriculum-driven DiT that dynamically adapts to degradation severity. First, we introduce a feedback-driven curriculum sampling strategy built upon a generative degradation process, which dynamically reweights degradation difficulty according to the evolving training state to progressively shift training from simple degradations to complex physical damages, thereby guiding the model to gradually unlock its untapped restoration potential. Second, CurrDiT adopts a dual-stream architecture that integrates degradation-aware local repair with global appearance guidance, where the Damage Restoration Branch (DRB) leverages damage-specific local cues to restore structurally corrupted regions, while the Global Degradation Restoration Branch (GDRB) exploits reliable non-damaged regions to enhance global appearance consistency. We further develop a generative degradation synthesis pipeline tailored to old-photo face restoration and construct a dedicated VintagePortraits benchmark for real-world evaluation of naturally aged old-photo portraits. Extensive experiments on old-photo face and blind face restoration benchmarks demonstrate that CurrDiT achieves superior perceptual quality, structural fidelity, and identity consistency.

1 INTRODUCTION

Real-world old-photo face restoration is a fundamental yet highly ill-posed inverse problem in computer vision. While early face restoration methods (Liu et al., 2007; Baker & Kanade, 2000) typically assume pre-defined degradation models, such simplifications drastically fail to generalize to real-world scenarios. Consequently to handle these scenarios, Blind Face Restoration (BFR), aiming to recover facial details from unknown corruptions, is promoted. However, most existing BFR paradigms are implicitly biased towards digital artifacts, e.g., Gaussian blur, uniform noise, and JPEG compression. Real-world old-photo face photos present a vastly different problem. While keeping the degradation setting unknown, this task contains more compound degradations that include severe structural physical damages (e.g., scratches, folds, and cracks) and chemical discoloration. Crucially, these damages often spatially entangle with wrinkles and pores of human faces. Blindly removing such complex artifacts easily leads to over-smoothing or hallucinated structures, resulting in severe facial identity distortion. Therefore, these task-specific challenges require the restoration network to not only robustly decouple heterogeneous unknown degradations from intrinsic facial details, but also enforce strict identity preservation during the restoration process.

To tackle these complex compound degradations, early efforts primarily focused on general old photo restoration. Early studies, such as Bringing Old Photos Back to Life (BOPBTL) (Wan et al., 2020), utilize specialized branches to handle structured defects (e.g., scratches) and unstructured ones (e.g., noise). Subsequent studies have further expanded this paradigm to address temporal coherence in old films (Wan et al., 2022) and context-aware scratch removal (Cai et al., 2023). Despite significantly

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

improving overall visual quality, directly transferring these general-purpose restoration methods to face-centric regions frequently falls short. Old photo restoration primarily optimizes perceptual naturalness from general image representations; however, human faces comprise strong semantics that demand strict identity preservation. As highlighted by a line of blind face restoration (BFR) works (Wang et al., 2021a; Yang et al., 2021; Chan et al., 2021; Zhu et al., 2022; Wang et al., 2022b; Gu et al., 2022), there exists an inherent trade-off between perceptual quality and fidelity. Lacking explicit semantic constraints, general-purpose representations easily confuse fine-grained facial textures with physical artifacts, inevitably leading to severe identity drift.

To break this bottleneck and handle unknown severe degradations, researchers naturally turn to strong generative priors. While early BFR methods heavily relied on pre-trained GANs (Karras et al., 2019), the field has recently shifted towards diffusion models (Ho et al., 2020; Rombach et al., 2022) due to their superior generation capabilities in mitigating complex real-world corruptions. Targeting old-photo face restoration where paired supervision is rare, SSDiff (Li et al., 2026) introduced a training-free framework by leveraging pseudo-reference faces and selective region guidance (e.g., scratch masks and parsing maps). However, such training-free methods present critical structural drawbacks, as their identity preservation relies heavily on heuristic guidance scales and test-time randomness, often leading to unstable outputs and inconsistent fidelity. Furthermore, existing training-based methods (Lin et al., 2024) conventionally rely on a static training paradigm that treats highly heterogeneous degradations equally, inevitably forcing a sub-optimal compromise between over-smoothing and structural hallucinations. Therefore, it is highly demanded to develop a curriculum-driven framework that progressively adapts to escalating degradation difficulties, while explicitly handling local structural damages and global appearance degradation in a decoupled manner.

In this paper, we propose **CurrDiT** (Curriculum-driven Diffusion Transformer), an end-to-end training framework for real-world old-photo face restoration. Unlike random sampling or fixed easy-to-hard curricula, CurrDiT introduces a feedback-driven curriculum learning strategy built upon a generative degradation synthesis pipeline, which dynamically reweights degradation difficulty according to the evolving training state and gradually shifts training from simple degradations to complex physical damages. For local structural damage restoration, CurrDiT leverages the **Damage Restoration Branch (DRB)** to exploit damage-specific local cues, enabling more effective repair of severe physical damages such as scratches, folds, and cracks. For global appearance restoration, we introduce the **Global Degradation Restoration Branch (GDRB)**. GDRB suppresses unreliable damaged conditional regions and exploits non-damaged regions as reliable global references, helping maintain coherent facial structure and identity-related appearance. In addition, we develop a generative degradation synthesis pipeline tailored to old-photo face restoration and construct **VintagePortraits**, a real-world benchmark of naturally aged old-photo portraits without prominent structural damage, complementing the mask-centric VintageFace benchmark (Li et al., 2026). Extensive experiments on old-photo face and blind face restoration benchmarks demonstrate that CurrDiT achieves superior perceptual quality, structural fidelity, and identity consistency.

2 Related work

2.1 Real-World Old-Photo Face Restoration

Real-world old-photo face restoration originates from the task of *old-photo restoration*. Pioneering work (Wan et al., 2020) formulates this as a latent-space translation problem to handle structured scratches and unstructured noise, with subsequent efforts addressing specific media like scanned halftone prints (Wan et al., 2022) and localized physical artifacts (Cai et al., 2023). Other efforts explore colorization and controllable modernization of historical photos (Luo et al., 2021; Jin et al., 2021; Liu et al., 2021; Gunawan et al., 2023), as well as temporally coherent restoration of old films (Wan et al., 2022; Iizuka & Simo-Serra, 2019). Recently, with the advent of strong generative models, diffusion priors have been adapted for general blind restoration (Lin et al., 2024). Despite these advances, real-world old-photo face restoration is a domain-specific problem that cannot be trivially solved by general-purpose methods. Unlike natural scenes, human faces are extremely sensitive to identity drift, and structural physical damages (e.g., severe cracks) in old photos easily disrupt crucial facial topologies. Consequently, specialized methods for real-world old-photo face restoration have emerged. Notably, SSDiff (Li et al., 2026) identifies this as an

independent challenge characterized by coupled fading, blurring, and physical damages, proposing a training-free framework driven by pseudo-references and selective region guidance. However, both such training-free paradigms and existing static training-based models fall short in real-world scenarios, necessitating an adaptive end-to-end training framework for old-photo face restoration.

2.2 Blind Face Restoration

To recover high-quality facial details from complex degradations, earlier methods mainly leverage GAN-based prior modeling, discrete codebook lookup. Specifically, GAN-based frameworks (Yang et al., 2021) integrate structural priors via feature modulation and cross-attention mechanisms. Meanwhile, discrete codebook-based approaches (Gu et al., 2022; Zhou et al., 2022; Tsai et al., 2024) quantize high-frequency facial details to bridge the degradation gap through reliable dictionary lookup. Recently, diffusion models have dominated BFR owing to their strong distribution modeling capability. To tackle complex corruptions beyond direct mapping, existing approaches employ diverse progressive strategies, such as two-stage degradation-agnostic diffusion (Wang et al., 2023) and recursive posterior error contraction (Yue & Loy, 2024). Furthermore, structural and semantic constraints—including frequency-domain decomposition (Miao et al., 2024), personalized priors (Varanka et al., 2024), degradation-agnostic learning (Xie et al., 2024), and hybrid 3D-2D prior fusion (Chen et al., 2024c)—are integrated to handle extreme degradations. Despite these advances, the application of DiT (Peebles & Xie, 2023) remains largely unexplored in this task, and existing paradigms fundamentally fail to decouple unstructured physical damages from intrinsic textures in real-world old-photo faces.

2.3 Diffusion Transformer for Low-Level Vision

In the diffusion era, DiTs (Peebles & Xie, 2023) are rapidly replacing U-Nets as the dominant backbone. Recent studies adapt DiTs to restoration tasks by incorporating specialized frequency conditions (Cheng et al., 2025), coarse-to-fine latent pyramids (Yu et al., 2025), or prompt/adapter-based foundation model transfer (Kong et al., 2025; Kim et al., 2025; Wang et al., 2025). However, applying DiT to real-world old-photo face restoration is a non-trivial task. As unstructured physical damages cannot be explicitly represented, networks must discover underlying patterns purely through data training. Yet, existing static training paradigms inevitably bias the network towards simple digital artifacts, leading to severe structural hallucinations on complex damages. To overcome these limitations, we propose a curriculum-driven DiT framework that progressively conquers unstructured corruptions, effectively tailoring the powerful generative capabilities of DiTs for this task.

3 Method

CurrDiT aims to recover a clean face I_{clean} from old photos under unknown, compound degradations. The framework is inherently driven by a curriculum learning strategy using synthetic samples $I_{\text{deg}} = \mathcal{D}(I_{\text{clean}}, M)$ from our generative degradation pipeline (detailed in Sec. 4.1), where M denotes a binary physical damage mask. According to the training state, the curriculum strategy dynamically updates sample weights to achieve smooth easy-to-hard learning. Given the degraded image I_{deg} and noisy target x_t, our dual-stream DiT handles localized physical damage and global appearance via the Damage Restoration Branch (DRB) and Global Degradation Restoration Branch (GDRB), respectively. Together, these components allow CurrDiT to progressively conquer complex corruptions while decoupling local structure repair from global facial consistency.

3.1 Feedback-Driven Adaptive Curriculum Learning

The ideal optimization trajectory for old photo restoration is that the model should first learn stable low-frequency geometry from easy samples, and gradually tackle complex physical damages from hard samples. However, existing methods predominantly rely on a static, randomly mixed training strategy. Particularly in the demanding task of old-photo face restoration, such a paradigm may result in a sub-optimal training strategy, limiting its ability to comprehensively cover the real-world distribution of diverse degradation types. We propose a dynamic sampling strategy that smoothly shifts the sampling probability distribution from easy-biased samples to hard-biased samples.

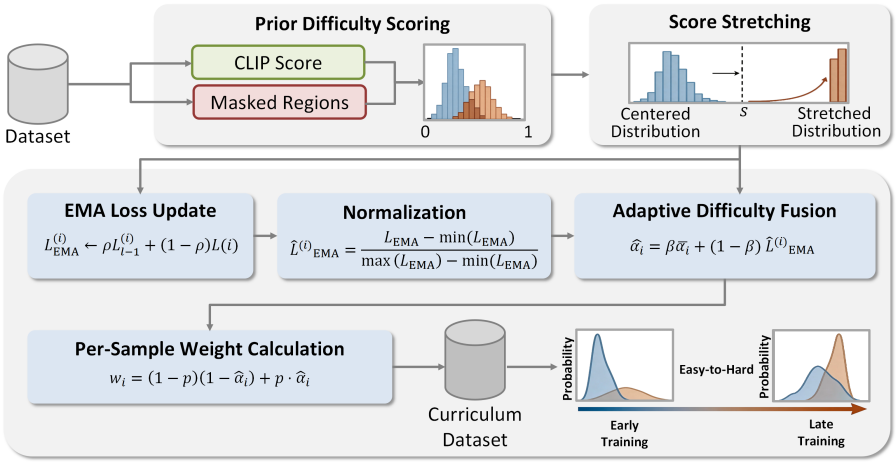

Figure 1: Overview of our feedback-driven curriculum learning process. Score stretching re-balances the difficulty range by spreading mid-range samples and enhancing easy–hard separability for better curriculum controllability. The EMA-based feedback provides a dynamic difficulty signal aligned with the evolving training state, as samples become easier once learned. Finally, progressive sampling smoothly shifts the sampling distribution from easy-biased to hard-biased, increasing the effective training difficulty while maintaining coverage across all difficulty levels for stable optimization.

Figure 1 presents the overall pipeline of this process. We initialize each sample i with a static proxy difficulty score $\alpha_i^{\text{prior}} \in [0,1]$ (e.g., derived from CLIP (Radford et al., 2021) or mask cues) to characterize how difficult it appears a prior. The detailed scoring rules are provided in Appendix C.2. A straightforward strategy is to schedule training or sampling from easy to hard solely according to this α_i^{prior}. However, we find that such a *prior-only curriculum* is unstable for old-photo face restoration and can even degrade performance (see ablation in Table 2). We attribute this to the coarse and uneven distribution of prior difficulty, and more importantly, to the evolving effective difficulty of each sample under the current training state; therefore, a reliable curriculum should progressively increase the proportion of hard samples while still keeping samples across all difficulty levels visible for stable optimization.

To address the aforementioned issues, we apply a contrast-stretching transform: $\tilde{\alpha}_i^{\text{prior}} = \texttt{clip}(0.5 + s(\alpha_i^{\text{prior}} - 0.5), 0, 1)$, where $s = 2$ is a stretching factor. Furthermore, recognizing that a static prior is insufficient because the relative difficulty of a sample decreases as the model training proceeds, we introduce a model-aware dynamic feedback signal. Specifically, we maintain a per-sample Exponentially Moving Averaged (EMA) loss, updated with momentum ρ via:

$$\mathcal{L}^{(i)}_{\text{EMA}} \leftarrow \rho\mathcal{L}^{(i)}_{\text{EMA}} + (1-\rho)\mathcal{L}^{(i)}. \tag{1}$$

We then linearly fuse the stretched prior $\tilde{\alpha}_i^{\text{prior}}$ and the normalized dynamic loss $\hat{\mathcal{L}}^{(i)}_{\text{EMA}}$ using a balancing factor β, and apply min-max normalization to yield the final self-consistent model-aware dynamic difficulty:

$$\alpha_i^{\text{dynamic}} = \text{Norm}\left(\beta\tilde{\alpha}_i^{\text{prior}} + (1-\beta)\hat{\mathcal{L}}^{(i)}_{\text{EMA}}\right). \tag{2}$$

This formulation ensures that the difficulty metric is both prior-guided and adaptively updated by the current learning state. To determine the curriculum stride, we define a global progress clock $p = \min(1, l/T) \in [0,1]$ at training step l. Finally, the sampling probability weight w_i for each sample is determined by:

$$w_i = (1-p)(1-\alpha_i^{\text{dynamic}}) + p\alpha_i^{\text{dynamic}}. \tag{3}$$

Through this mechanism, when the training begins ($p \approx 0$), i.e., $w_i \approx 1 - \alpha_i^{\text{dynamic}}$, firstly utilize easy samples. As the curriculum advances towards the end ($p \approx 1$), the weighting dynamically flips to $w_i \approx \alpha_i^{\text{dynamic}}$, shifting the distribution to prioritize the hardest corruptions, thereby establishing solid easy-to-hard training trajectory for CurrDiT.

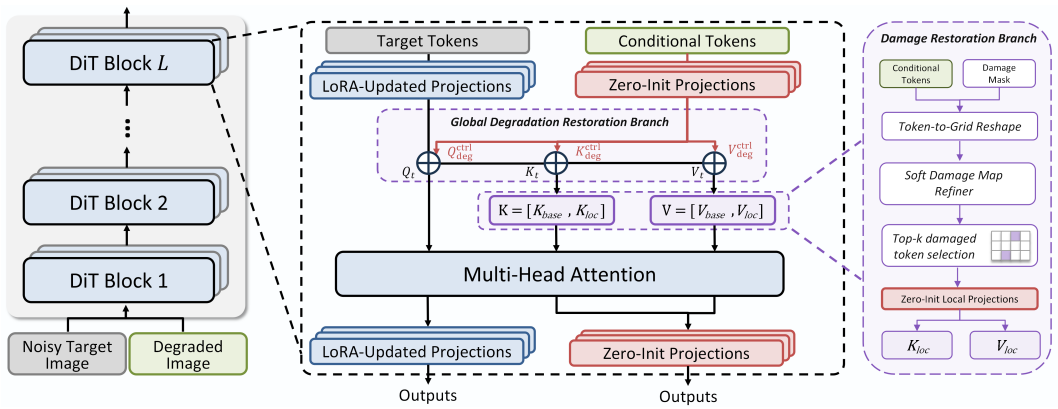

Figure 2: Overview of the proposed dual-branch restoration module. The left panel shows the dual-stream DiT architecture, while the middle and right zoom-ins illustrate GDRB and DRB.

3.2 Dual-Branch Degradation Restoration

Damage Restoration Branch (DRB). Given the conditional degraded tokens $X^c \in \mathbb{R}^{N\times d}$ from the degraded input branch and the binary damage mask $M_h \in \{0,1\}^{H\times W}$ aligned with the conditional token grid ($N = H \times W$), DRB first estimates a soft damage confidence map to explicitly localize restoration-critical regions. Specifically, the damage mask is resized to the same spatial grid as the conditional tokens. We reshape X^c back to a 2D feature map $F_c \in \mathbb{R}^{d\times H\times W}$ and concatenate it with M_h along the channel dimension. A lightweight projection predicts the soft damage confidence map M_s. Compared with the binary mask, M_s provides smoother position-wise confidence scores, better reflecting the relative importance of damaged regions. The soft map is then flattened into token scores and used to select the most damage-relevant conditional tokens. Instead of using all conditional tokens uniformly, we keep only the top-k tokens with the highest soft damage scores, where $k = \mathrm{clip}(\lfloor rN \rfloor,\ N_{\min},\ N_{\max})$. Here, r is the top-k ratio, while $N_{\min}$ and $N_{\max}$ stabilize the number of selected local tokens. Let $\mathcal{I}_{\mathrm{topk}}$ denote the selected indices. We gather the corresponding conditional tokens and weight them by their confidence values, obtaining $T_{\mathrm{loc}} = \{X^c_i \cdot M_s(i) \mid i \in \mathcal{I}_{\mathrm{topk}}\}$. Specifically, T_{loc} is projected into a dedicated local key-value space as $K_{\mathrm{loc}} = \mathrm{Proj}^k_{\mathrm{loc}}(T_{\mathrm{loc}})$ and $V_{\mathrm{loc}} = \mathrm{Proj}^v_{\mathrm{loc}}(T_{\mathrm{loc}})$, where $\mathrm{Proj}^k_{\mathrm{loc}}$ and $\mathrm{Proj}^v_{\mathrm{loc}}$ are local projection layers independent from the global conditional branch. The final key-value set is augmented as $K = [K_{\mathrm{base}}, K_{\mathrm{loc}}]$ and $V = [V_{\mathrm{base}}, V_{\mathrm{loc}}]$, where K_{base} and V_{base} denote the original attention keys and values from the main and conditional streams. DRB injects damage-specific local cues into the key-value set, enabling focused structural restoration while preserving the original DiT attention pathway. Empirically, we observe that focusing only on damaged regions is insufficient, since intact regions may still exhibit color cast or fading, motivating GDRB to exploit reliable non-damaged context.

Global Degradation Restoration Branch (GDRB). Let $X = [X^m; X^c]$ denote the visual token sequence in a DiT block, where $X^m \in \mathbb{R}^{N_m\times d}$ are the main restoration tokens and $X^c \in \mathbb{R}^{N_c\times d}$ are the conditional-image tokens. $M \in \{0,1\}^{N_c}$, where $M_j = 1$ indicates that the j-th conditional token corresponds to a damaged region. Here, Q^c, K^c, and V^c correspond to the degradation-control projections illustrated as $Q^{\mathrm{ctrl}}_{\mathrm{deg}}$, $K^{\mathrm{ctrl}}_{\mathrm{deg}}$, and $V^{\mathrm{ctrl}}_{\mathrm{deg}}$ in Figure 2; they are concatenated with the target projections. The two streams are then concatenated and processed by joint attention. To prevent damaged conditional regions from serving as unreliable global references, we construct an additive visibility mask Δ, where $\Delta_{ij} = -\infty$ if $i \in \Omega_m$, $j \in \Omega_c$, and $M_j = 1$, and $\Delta_{ij} = 0$ otherwise. Here, Ω_m and Ω_c denote the index sets of main and conditional tokens, respectively. The masked joint attention is computed as $\mathrm{Attn}(Q,K,V) = \mathrm{Softmax}\Big(QK^\top/\sqrt{d} + \Delta\Big)V$. Compared with simply zeroing damaged tokens, the additive visibility mask Δ excludes damaged conditional tokens from the attention scope before softmax normalization, avoiding residual attention weights on unreliable regions. As a result, GDRB complements DRB by using trustworthy non-damaged context to correct global degradations and maintain coherent facial structure and identity-related appearance.

Table 1: Quantitative comparison on old-photo face and blind face restoration datasets. Green, blue, and orange headers respectively indicate unsupervised, supervised, and our method. Best and second-best results are shown in **bold** and underline.

Metric		DDNM	PGDiff	SSDiff	BOPBTL	GFP-GAN	CodeFormer	DifFace	DiffBIR	InterLCM	CurrDiT
Old-Photo Face Datasets											
QAlign↑	Simple	2.1537	3.2544	3.6482	2.4788	2.5935	1.9223	3.0477	3.4958	3.8633	**4.3718**
MUSIQ↑		41.90	65.54	66.45	41.68	51.98	29.55	61.74	66.33	72.72	**75.50**
TOPIQ↑		0.4352	0.6140	0.6271	0.3416	0.3885	0.3552	0.5559	0.6126	0.6981	**0.7621**
MANIQA↑		0.3705	0.4430	0.4563	0.2276	0.2817	0.1924	0.3976	0.4822	0.5489	**0.5884**
FID↓		185.49	**123.46**	133.59	177.17	156.88	196.92	128.41	142.47	138.63	132.15
QAlign↑	Medium	2.1243	3.2288	3.5485	2.3424	2.6111	1.8227	2.9790	3.4596	3.7584	**4.4802**
MUSIQ↑		41.33	65.35	64.53	39.72	52.08	27.62	60.48	65.88	73.03	**76.20**
TOPIQ↑		0.4379	0.5968	0.6081	0.3255	0.3819	0.3394	0.5388	0.6033	0.6959	**0.7705**
MANIQA↑		0.3600	0.4274	0.4482	0.2099	0.2742	0.1792	0.3813	0.4730	0.5407	**0.5972**
FID↓		184.47	132.23	129.00	175.20	173.08	208.26	129.38	145.34	144.46	**125.51**
QAlign↑	Hard	1.8364	2.9219	3.1246	2.0552	2.3642	1.6048	2.6580	3.0674	3.5679	**4.2610**
MUSIQ↑		35.98	62.66	60.35	33.31	48.21	24.57	54.85	61.65	72.06	**75.97**
TOPIQ↑		0.3665	0.5689	0.5704	0.2821	0.3648	0.3183	0.4856	0.5666	0.6770	**0.7553**
MANIQA↑		0.2987	0.3950	0.3987	0.1780	0.2495	0.1725	0.3220	0.4101	0.5151	**0.5839**
FID↓		201.48	131.44	125.47	217.89	178.41	244.45	136.43	160.25	136.19	**122.97**
QAlign↑	VintagePortraits	1.6551	2.6595	2.7781	1.9453	2.5772	1.5747	2.5063	3.9994	3.8853	**4.5152**
MUSIQ↑		30.63	56.94	58.20	34.81	55.40	24.98	52.24	74.96	70.55	**76.78**
TOPIQ↑		0.3247	0.4844	0.5217	0.2639	0.4271	0.3525	0.4400	0.7520	0.6559	**0.7666**
MANIQA↑		0.2343	0.3236	0.3430	0.1835	0.3024	0.1831	0.2803	**0.6448**	0.5206	0.6102
FID↓		115.57	65.13	61.67	100.47	56.54	154.30	66.25	52.66	54.31	**43.28**
Blind Face Datasets											
QAlign↑	WebPhoto	1.3986	3.5388	3.2528	1.1820	2.4743	1.3640	3.1057	3.9881	3.7984	**4.1248**
MUSIQ↑		24.30	68.60	61.24	16.73	53.02	19.86	62.04	75.34	72.16	**76.63**
TOPIQ↑		0.2867	0.6047	0.5595	0.1774	0.4389	0.3583	0.5540	0.7586	0.6931	**0.7644**
MANIQA↑		0.2153	0.4460	0.3936	0.1210	0.3122	0.1857	0.4047	**0.6506**	0.5293	0.5968
FID↓		153.58	86.96	94.78	195.64	86.92	165.22	**82.75**	88.87	91.46	89.87
QAlign↑	CelebChild	2.2105	3.5707	3.5353	1.9316	2.7675	1.8335	3.2424	4.1157	4.0925	**4.2626**
MUSIQ↑		44.43	68.04	66.49	34.78	58.68	31.08	63.65	75.25	72.16	**76.20**
TOPIQ↑		0.4645	0.6134	0.5893	0.4033	0.4493	0.3459	0.5581	**0.7592**	0.6668	0.7583
MANIQA↑		0.3952	0.4432	0.4243	0.2738	0.3288	0.1976	0.4037	**0.6415**	0.5152	0.5811
FID↓		155.86	114.49	121.27	160.78	113.23	147.62	**102.36**	120.13	112.83	112.20
QAlign↑	CelebAdult	2.8612	4.3392	4.3819	2.7206	3.2618	2.0964	4.0747	4.7257	4.6875	**4.7476**
MUSIQ↑		54.18	71.71	70.57	50.32	62.07	42.12	69.35	76.33	74.97	**76.89**
TOPIQ↑		0.5715	0.6761	0.6325	0.5729	0.4712	0.3845	0.6280	**0.7806**	0.6938	0.7702
MANIQA↑		0.5078	0.5043	0.4716	0.4228	0.3609	0.2510	0.4642	**0.6903**	0.5414	0.5879
FID↓		154.58	101.03	107.95	156.07	106.15	127.09	**94.30**	108.50	107.11	104.18

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Benchmark and Degradation Pipeline.

We construct **VintagePortraits**, an unpaired real-world benchmark with 1,103 real old-photo faces for evaluation. Unlike VintageFace (Li et al., 2026), which mainly focuses on visibly damaged photos, VintagePortraits covers naturally aged portraits without obvious physical breakage, including fading, blur, low contrast, color shift, noise, and scanning artifacts.

In particular, we observe that existing old-photo restoration works lack a dedicated mask-generation strategy for modeling realistic structural damages in aged portraits. To support training, we design a degradation pipeline tailored for old-photo face restoration, where a generative model, FLUX-1-Kontext-dev (Black Forest Labs et al., 2025), is used to synthesize structural damage masks and combined with task-tailored Real-ESRGAN-style degradation (Wang et al., 2021b). These masks are injected at different stages of the degradation chain to simulate damages with diverse sharpness, aging effects, and com-

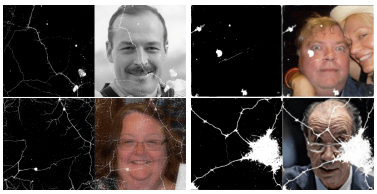

Figure 4: Pipeline-generated damage masks and degraded images.



Figure 3: Real-world old-photo face restoration comparison between our method and existing methods.

pression artifacts, enabling the synthesis of diverse training samples, as shown in Figure 4. More details are provided in Appendix B.

Experimental Protocol. We fine-tune FLUX-1-Kontext-dev (Black Forest Labs et al., 2025) with LoRA (Hu et al., 2021) on FFHQ (Karras et al., 2019) images at 512×512 resolution using our proposed degradation pipeline. We evaluate on VintageFace (Li et al., 2026), VintagePortraits, WebPhoto (Wang et al., 2021a), CelebChild (Wang et al., 2021a), and CelebAdult (Chen et al., 2024b) to assess performance on real-world old-photo and blind face restoration. Since real-world old-photo face restoration is unpaired and lacks ground truth, we follow existing methods (Li et al., 2026) and adopt no-reference metrics. We report four no-reference metrics, namely QAlign (Wu et al., 2023), MUSIQ (Ke et al., 2021), TOPIQ (Chen et al., 2024a), and MANIQA (Yang et al., 2022), along with FID (Heusel et al., 2017), on these unpaired benchmarks. To evaluate identity preservation, we further report ArcFace Deg. (Gu et al., 2022) and LMD (Gu et al., 2022) on paired synthetic data (more details in Appendix C.6). Detailed dataset configurations, implementation settings, and evaluation protocols are provided in Appendix C.

4.2 Performance Comparison

Quantitative Results. Table 1 reports the quantitative results of CurrDiT compared to existing unsupervised and supervised methods. Results compared to unsupervised methods, including DDNM (Wang et al., 2022a), PGDiff (Yang et al., 2023), and SSDiff (Li et al., 2026), show that CurrDiT achieves the best scores across the majority of metrics, indicating its superior capability in balancing perceptual quality. When compared to supervised methods, e.g., BOPBTL (Wan et al., 2020), GFP-GAN (Wang et al., 2021a), CodeFormer (Zhou et al., 2022), DifFace (Yue & Loy, 2024), DiffBIR (Lin et al., 2024), and InterLCM (Li et al., 2025), where these methods do not have degradation-specific modeling, our method obtains substantial improvements in overall restoration performance. Notably, compared to the most relevant work, SSDiff (Li et al., 2026), which fundamentally functions in a training-free, unsupervised manner, CurrDiT consistently outperforms it across nearly all evaluation metrics. This result validates that our curriculum-driven approach offers better robustness than existing methods. Furthermore, as the difficulty increases from the *simple* to the *hard* subset, CurrDiT maintains strong performance across the evaluated metrics, demonstrating its effectiveness in handling complex real-world degradations. In addition, on the newly introduced VintagePortraits dataset, CurrDiT achieves the best results on most metrics, showing its effectiveness on old-photo portraits without structural damage. Furthermore, on three blind face datasets, CurrDiT consistently achieves the best or second-best results on most perceptual metrics, indicating strong generalization beyond old-photo-specific restoration.

Qualitative Results. Figure 3 compares CurrDiT with representative baselines under real-world degradations, including structurally damaged old photos and naturally aged portraits from VintagePortraits. Unsupervised methods struggle with scratch removal and often produce inconsistent colorization, while specialized old-photo restoration and generic BIR methods remain sensitive to complex physical damages. In contrast, CurrDiT robustly restores severely damaged old photos even when the original identity is barely recognizable. On VintagePortraits, CurrDiT also produces more natural colors, cleaner details, and fewer artifacts. This visual superiority closely aligns with the quantitative observations in Table 1. To assess generalization beyond the degradation types and severity covered by our synthesis pipeline, we further evaluate on annotated OOD subsets and analyze failure cases (Appendix E).

4.3 Ablation Study

Table 2 reports the component-wise ablation on the Medium subset of VintageFace (Li et al., 2026), while Figure 5 provides the corresponding visual comparison. Overall, progressively enabling the adaptive curriculum, DRB, and GDRB consistently improves all perceptual metrics, indicating that dynamic difficulty adaptation, local structural repair, and global appearance guidance provide complementary benefits.

Table 2: Component-wise ablation study.

Backbone	Prior Curr	Adaptive Curr	DRB	GDRB	QAlign↑	MUSIQ↑	TOPIQ↑	MANIQA↑
✓	-	-	-	-	4.2446	74.29	0.7354	0.5538
✓	✓	-	-	-	4.1199	72.98	0.6974	0.5090
✓	-	✓	-	-	4.2554	74.87	0.7439	0.5639
✓	-	✓	✓	-	<u>4.4694</u>	<u>75.75</u>	<u>0.7595</u>	<u>0.5876</u>
✓	-	✓	✓	✓	**4.4802**	**76.20**	**0.7705**	**0.5972**

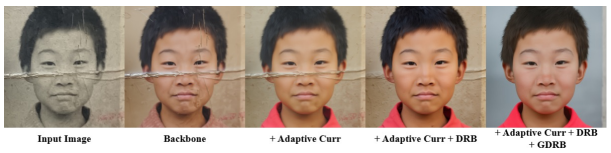

Figure 5: Qualitative ablation of enabled components.

Effect of Curriculum Learning. Compared with the randomly sampled backbone, our adaptive curriculum yields consistent gains across all evaluated metrics. This indicates that random sampling under-exploits the restoration potential of the model under highly heterogeneous old-photo degradations. By contrast, the prior-only curriculum lowers QAlign and MANIQA relative to random sampling, suggesting that a static difficulty order may provide misleading training guidance when the prior difficulty does not match the model's current restoration capability. These results demonstrate that dynamically adjusting the training emphasis better matches the evolving optimization difficulty of old-photo face restoration. As further shown in Figure 6, Easy samples gradually decrease from about 59% to 25%, while Medium/Hard samples increase from about 41% to 75%, verifying a stable easy-to-hard curriculum behavior.

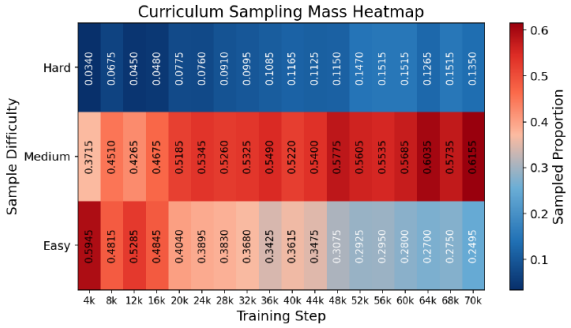

Figure 6: Proportions of Easy, Medium, and Hard samples across training intervals.

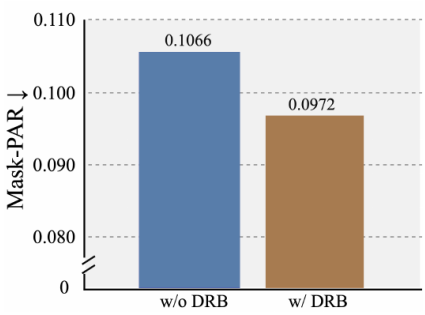

Figure 7: Mask-region artifact comparison with and without DRB.

Local Restoration and Global Consistency. As shown in Table 2, DRB substantially improves local structural damage restoration. This indicates that exploiting damage-specific local cues helps the model distinguish physical scratches and folds from intrinsic facial textures. In Figure 5, the model without DRB still shows incomplete restoration and residual artifacts around damaged facial regions, whereas DRB produces cleaner local repair and more complete facial details. Figure 7 further supports this observation. We use Mask-PAR to measure residual perceptual artifacts inside the original damage masks, averaged over the dataset. DRB reduces the averaged Mask-PAR, with an 8.80% relative reduction, confirming that DRB improves local restoration in damaged regions. Additional ablations examine how token selection and injection depth affect DRB's restoration performance (Appendix D.2). Beyond local repair, adding GDRB after DRB further improves all metrics in Table 2, indicating that local structural repair alone is insufficient for globally coherent restoration. This is because old-photo degradations also involve color fading, illumination shifts, and global appearance inconsistency. As shown in Figure 5, GDRB produces more natural skin tones and more consistent global appearance, confirming that reliable non-damaged regions provide useful references for global restoration. We further evaluate the robustness of both branches to imperfect damage masks in Appendix D.3.

Identity Preservation. Since real-world benchmarks are unpaired, we re-split the training data and construct an independent paired synthetic test set for identity and structure evaluation. See Appendix C.6 for the protocol. We train CurrDiT on the new training split and evaluate all representative methods on the same test set. As shown in Table 3, our method achieves the lowest ArcFace Deg. and LMD, indicating better identity consistency and facial structure preservation. We further report additional perceptual metrics under the same evaluation setting in Appendix C.7.

Table 3: Identity preservation comparison.

Metric	PGDiff	SSDiff	DiffBIR	InterLCM	Ours
ArcFace Deg.↓	54.94	50.67	37.29	<u>33.93</u>	**31.15**
LMD↓	4.37	3.69	2.58	<u>2.41</u>	**1.93**

5 Conclusion

In this paper, we proposed CurrDiT, a feedback-driven curriculum DiT framework for real-world old-photo face restoration, which progressively adapts training from simple degradations to complex physical damages. CurrDiT combines a task-tailored generative degradation synthesis pipeline with adaptive curriculum sampling, and integrates DRB for local structural damage repair and GDRB for global appearance guidance. Training feedback guides the curriculum progression, allowing the learning emphasis to adapt as restoration capabilities develop. The two branches use damage-specific local cues to repair corrupted facial structures and reliable undamaged regions to guide global appearance restoration, respectively. We also construct the dedicated VintagePortraits benchmark for naturally aged old-photo portraits. Extensive experiments on old-photo face and blind face restoration benchmarks demonstrate that CurrDiT achieves superior perceptual quality, structural fidelity, and identity consistency. Despite these advances, CurrDiT may still struggle with extreme out-of-distribution damages beyond the degradation synthesis range, and future work will explore explicit 3D geometric priors for restoring severely corrupted facial topology.

AI USE STATEMENT

We used OpenAI ChatGPT to assist with manuscript translation, language polishing, and organization, and to support literature search and the identification of relevant prior work. The authors reviewed the AI-assisted revisions and checked the cited references against their original sources.

ETHICS STATEMENT

VintagePortraits contains 1,103 publicly accessible historical face images collected through the official Unsplash API and is used exclusively for restoration evaluation, not model training or identity recognition. Its construction involved no participant recruitment or interaction with the depicted individuals. Before release, we will review per-image provenance and applicable licensing and redistribution terms. Images without a clear redistribution basis will not be distributed directly. We will document these checks and provide a removal-request mechanism for depicted individuals and relevant rights holders.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we will publicly release our code and models, and all implementation details, training hyperparameters are provided in the Appendix.

REFERENCES

Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7941–7950, 2021.

Simon Baker and Takeo Kanade. Hallucinating faces. In *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, pp. 83–88. IEEE, 2000.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Weiwei Cai, Huaidong Zhang, Xuemiao Xu, Shengfeng He, Kun Zhang, and Jing Qin. Contextual-assisted scratched photo restoration. *IEEE Transactions on circuits and systems for video technology*, 33(10):5458–5469, 2023.

Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14240–14249. IEEE, 2021.

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*, 33:2404–2418, 2024a.

Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. Toward real-world blind face restoration with generative diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8494–8508, 2024b.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1263–1271, 2024c.

Kun Cheng, Lei Yu, Zhijun Tu, Xiao He, Liyu Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Effective diffusion transformer architecture for image super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 2455–2463, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

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 *European Conference on Computer Vision*, pp. 126–143. Springer, 2022.

Agus Gunawan, Soo Ye Kim, Hyeonjun Sim, Jae-Ho Lee, and Munchurl Kim. Modernizing old photos using multiple references via photorealistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12460–12469, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qilong Sun, and Qingyu Liu. Focusing on persons: Colorizing old images learning from modern historical movies. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1176–1184, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Jin Hyeon Kim, Paul Hyunbin Cho, Claire Kim, Jaewon Min, Jaeeun Lee, Jihye Park, Yeji Choi, and Seungryong Kim. Unified diffusion transformer for high-fidelity text-aware image restoration. *arXiv preprint arXiv:2512.08922*, 2025.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12809–12819. IEEE, 2025.

Senmao Li, Kai Wang, Joost van de Weijer, Fahad Khan, Chun-Le Guo, Shiqi Yang, Yaxing Wang, jian Yang, and Ming-Ming Cheng. InterLCM: Low-quality images as intermediate states of latent consistency models for effective blind face restoration. In *International Conference on Learning Representations*, volume 2025, pp. 59165–59197, 2025.

Wenjie Li, Xiangyi Wang, Heng Guo, Guangwei Gao, and Zhanyu Ma. Self-supervised selective-guided diffusion model for old-photo face restoration. *Advances in Neural Information Processing Systems*, 38:144382–144407, 2026.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pp. 430–448. Springer, 2024.

Ce Liu, Heung-Yeung Shum, and William T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

Jixin Liu, Rui Chen, Shipeng An, and Heng Zhang. Cg-gan: Class-attribute guided generative adversarial network for old photo restoration. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 5391–5399, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M Seitz. Time-travel rephotography. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6583–6592. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. In *International Conference on Learning Representations*, volume 2024, pp. 44867–44890, 2024.

Tuomas Varanka, Tapani Toivonen, Soumya Tripathy, Guoying Zhao, and Erman Acar. Pfstorer: Personalized face restoration and super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2372–2381. IEEE, 2024.

Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2744–2754. IEEE, 2020.

Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17673–17682. IEEE, 2022.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2161–2172. IEEE, 2025.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9164–9174. IEEE, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021b.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022a.

Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1704–1713. IEEE, 2023.

Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17491–17500. IEEE, 2022b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9120–9129, 2024.

Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances in Neural Information Processing Systems*, 36:32194–32214, 2023.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 2022.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 672–681. IEEE, 2021.

Yongsheng Yu, Haitian Zheng, Zhifei Zhang, Jianming Zhang, Yuqian Zhou, Connelly Barnes, Yuchen Liu, Wei Xiong, Zhe Lin, and Jiebo Luo. Zipir: Latent pyramid diffusion transformer for high-resolution image restoration. *arXiv preprint arXiv:2504.08591*, 2025.

Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9991–10004, 2024.

Lingzhi Zhang, Yuqian Zhou, Connelly Barnes, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for inpainting. In *European Conference on Computer Vision*, pp. 146–164. Springer, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7652–7661. IEEE, 2022.