DirectFlow: One-Step Posterior Transport Distillation for Blind Face Restoration

Anonymous CVPR submission

Paper ID *****

Abstract

*Blind face restoration inherently struggles with the perception-distortion trade-off. Regression-based methods minimize distortion but inevitably produce over-smoothed, blurry textures, while generative models synthesize sharp details but suffer from prohibitive computational latency or structural identity shifts. Recent flow-matching paradigms elegantly reconcile this conflict by explicitly transporting the distortion-optimal posterior mean to the natural image manifold; however, they still rely on computationally expensive ODE solvers. In this work, we propose **DirectFlow** to achieve this balance in a single forward pass. Rather than naively matching ground-truth images, DirectFlow distills the transported posterior mean of a Rectified Flow teacher directly into a latent consistency model. We repurpose the teacher's velocity field as a distributional critic, pulling single-step predictions towards true data manifold peaks without solver integration error. Furthermore, to dynamically adapt to diverse degradations, we introduce a semantic adapter paired with Low-Rank Adaptation (LoRA) in the UNet's cross-attention layers. By a two-stage optimization schedule that first aligns conditioning modules and then freezes spatial pathways during LoRA refinement, this strategy enables continuous, degradation-aware conditioning without compromising the foundational generative prior. Extensive experiments demonstrate that one-step DirectFlow matches the perceptual quality of a 25-step flow-matching teacher while accelerating inference by 7.5×, establishing a highly efficient state-of-the-art architecture for real-time, high-fidelity blind face restoration.*

1. Introduction

The restoration of high-quality face images from degraded observations (encompassing blur, noise, compression artifacts, and downsampling) is an ill-posed inverse problem whose solution space is governed by the perception-distortion trade-off [2, 3, 32]. This theorem proves that minimizing pixel-level distortion (MSE/PSNR/SSIM [48]) inevitably degrades perceptual quality (FID [11]/LPIPS [61]), and vice versa. The field has historically navigated this trade-off through two extreme paradigms:

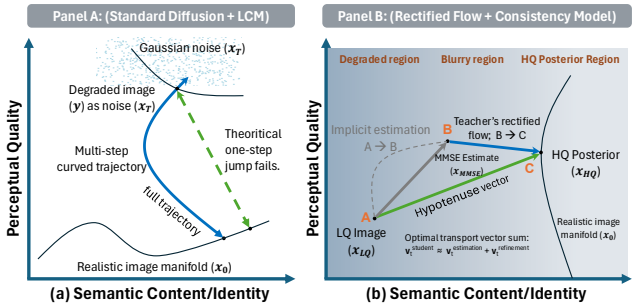

Figure 1. Trajectory analysis of restoration frameworks. (a) Iterative diffusion paths are highly curved, causing naive one-step consistency jumps to fail. (b) DirectFlow constructs a direct mapping to the transported posterior mean, attempting to achieve optimal perceptual quality in a single step.

Regression models (MMSE estimators). Methods minimizing L_1/L_2 losses learn the conditional expectation $\hat{X}^* = \mathbb{E}[X|Y]$, representing the posterior mean. While uniquely minimizing expected distortion, this approach averages over all plausible high-quality images. This averaging suppresses high-frequency detail and produces smooth, over-regularized textures. Architectures such as SwinIR [21] epitomize this regime by achieving excellent PSNR at the direct cost of poor perceptual realism.

Iterative generative models. Diffusion-based restorers [23, 45, 51] instead sample individual instances $x \sim p(x|y)$ from the posterior distribution. Each sample is sharp and photorealistic. However, stochastic variation elevates distortion, and the iterative denoising process requires 20 to 50 neural function evaluations (NFEs), precluding real-time deployment. Moreover, because these methods typically initiate sampling from noise, they introduce output uncertainty that is undesirable for deterministic restoration, and the randomness can manifest as identity or structure drift even when perceptual sharpness is high.

1.1. The Transported Posterior Mean

Recent advances in optimal-transport-based image restoration [35] identify a theoretically appealing middle ground.

Rather than sampling from the posterior $p(x|y)$, these methods construct a deterministic mapping from the posterior mean to the natural-image manifold via a Rectified Flow [24, 26]. Concretely, a velocity field $v_\phi(x_t, t)$ defines an ordinary differential equation (ODE) whose solution transports the MMSE estimate $\hat{X}^*$ onto the data manifold:

$$\left.\begin{aligned} \frac{dx}{dt} &= v_\phi(x_t, t), \quad x(0) = \hat{X}^* \\ x(1) &= \hat{X}_0 \in \mathrm{supp}(p_{\text{data}}) \end{aligned}\right\} \tag{1}$$

As illustrated in Fig. 1(a), standard diffusion requires following a curved trajectory where naive one-step consistency jumps fail to reach the high-quality manifold. By contrast, optimal transport straightens this path. The resulting endpoint $\hat{X}_0$ (the transported posterior mean) achieves the minimum MSE among all estimators constrained to lie on the natural-image manifold [35]. Although this transported state represents an optimal perceptual estimator [3, 35], obtaining it requires integrating the ODE across multiple steps. This iterative dependency introduces a prohibitive inference latency that renders real-time execution infeasible. Standard one-step distillation seems like the straight-forward path, but in practice it is brittle: reduced-capacity students and teacher–student trajectory mismatch often prevent naive distillation from matching the multi-step teacher, and the distillation targets themselves typically require running the teacher solver for many steps, creating a hidden training-time and storage burden. Further, latent-space distillation compounds the issue by injecting an information bottleneck through the VAE, which can blunt fine textures and semantic fidelity. To overcome this prohibitive latency, we frame blind restoration through the lens of latent consistency models [29, 43], which natively map any arbitrarily degraded state along a transport trajectory directly to its optimal high-quality origin in a single step.

1.2. Our Approach and Contributions

We propose **DirectFlow**, the first framework to distill the transported posterior mean trajectory into a single-step student network. As depicted in Fig. 1(b), DirectFlow circumvents integration latency by modeling restoration as a compiled optimal transport map. The student directly learns the hypotenuse vector connecting the degraded input to the transported posterior mean, bypassing the intermediate MMSE state entirely. Rather than targeting ground-truth images or sampling from the posterior, DirectFlow trains a Latent Consistency Model [17, 29] student to predict the frozen teacher's multi-step output in one forward pass.

To realize this, our framework introduces three key innovations, which form our primary contributions:

1. We formalize cross-space knowledge transfer between pixel-space teachers and latent-space students. By projecting the teacher's endpoint $X^\dagger = T(\hat{X}^*(Y))$ into the student's latent space using a shared frozen VAE encoder, we achieve highly efficient distillation with minimal information loss.
2. To adapt dynamically to diverse corruptions, we introduce a synergistic fine-tuning strategy combining a semantic adapter with Low-Rank Adaptation (LoRA) of the base UNet. By injecting stage-aligned spatial encoder and degradation tokens into LoRA-augmented UNet, we empower the network to actively attend to degradation hints.
3. We repurpose the frozen teacher's velocity field as a distributional critic. By querying the teacher's flow at the student's output, SGCL pushes predictions toward probability-density peaks, overcoming the information bottleneck inherent to latent-space distillation.

Extensive experiments demonstrate that DirectFlow achieves perceptual quality competitive with a 25-step flow-matching teacher at **7.5× faster inference**, creating a highly efficient state-of-the-art method for real-time blind face restoration.

2. Related Work

2.1. Blind Face Restoration

GAN-based approaches. Early methods relied on geometric priors like facial landmarks or parsing maps [7, 56]. GFPGAN [46] exploits pre-trained generative facial priors to hallucinate realistic textures, while CodeFormer [63] and VQFR [8] replace the GAN prior with discrete codebook lookup transformers for improved robustness. Both achieve single-pass inference but struggle with severe degradations and can introduce identity shifts due to the limited expressiveness of their generative priors [54].

Diffusion-based approaches. DiffBIR [23] and DR2 [51] leverage pre-trained Stable Diffusion [40] as a generative prior, conditioning the denoising process on the degraded input. StableSR [45] further introduces time-aware feature modulation for controllable super-resolution, following principles of iterative refinement established in SRDiff [16]. These methods achieve state-of-the-art perceptual quality but require 20–50 NFEs per image, rendering them impractical for real-time applications.

Flow-based optimal estimation. PMRF [35] introduces the posterior-mean rectified-flow paradigm: first compute the MMSE estimate via a regression network, then transport it onto the natural-image manifold via a learned Rectified Flow [25, 26]. The result is provably the closest point on the data manifold to the posterior mean in the optimal-transport sense [44]. However, it still requires 25 iterative ODE steps.

2.2. Distillation for Diffusion Models

Progressive distillation [41] halves the number of sampling steps by training a student to match pairs of consecutive teacher steps. Consistency Models [43] enforce a self-consistency property: any point on the ODE trajectory maps

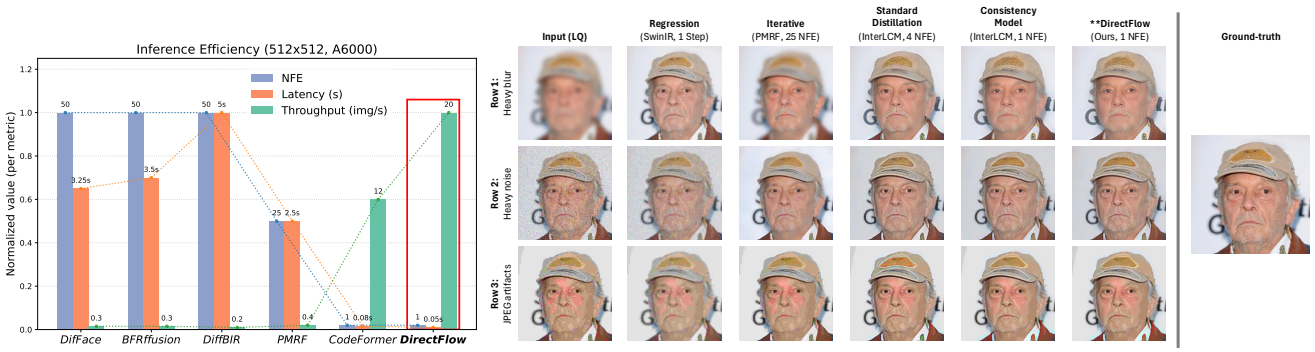

Figure 2. **(Left)** Inference step comparison, highlighting the computational bottleneck of iterative generative models. **(Right)** Qualitative results across diverse degradations. Regression models yield over-smoothed outputs, while flow-matching models (PMRF) recover photorealistic details but require 25 iterative steps. Standard distillation (InterLCM) achieves one-step inference but suffers from latent VAE bottlenecks. DirectFlow achieves the perceptual sharpness of the 25-step teacher in a single forward pass.

to the same origin, enabling single-step generation. Latent Consistency Models (LCM) [28] and ADD [42] extend this idea to the latent space of Stable Diffusion, achieving 1–4 step text-to-image generation. InterLCM [17] adapts the LCM framework specifically for controllable face restoration with dedicated visual and spatial encoders. Our work draws inspiration from these consistency-based distillation ideas but targets a fundamentally different objective: we distill the *transported posterior mean* rather than an unconditional generation trajectory, and we introduce a cross-space bridge to handle the mismatch between teacher and student representation domains.

2.3. Adapter Injection for Frozen Diffusion Models

Injecting new conditioning signals into frozen diffusion UNets is a well-studied problem, with approaches varying along two axes: *what* is injected (spatial maps, tokens, style vectors) and *how* it is injected (cross-attention, residual addition, gated fusion) [12, 34].

Spatial conditioning. ControlNet [60] copies the encoder half of the UNet and connects outputs through zero-initialized 1×1 convolutions, ensuring the model starts from its pre-trained state. While highly effective, this doubles the model size (>300M parameters). T2I-Adapter [34] employs lightweight feature-matching networks that inject multi-scale spatial features into the UNet's encoder blocks, achieving similar controllability at lower cost. Similar lightweight strategies have been explored in X-Adapter [58] for upgrading model versions.

Gated and grounded injection. GLIGEN [19] introduces gated self-attention layers with initialized gates for open-set grounded generation, allowing spatial grounding of text-to-image synthesis. Uni-ControlNet [62] and UniControl [38] unify multiple control signals through shared adapters and task-routing mechanisms.

Token-level conditioning. IP-Adapter [55] injects image features through decoupled cross-attention with learned projection layers, maintaining text compatibility while adding image-based control. Ctrl-Adapter [22] and PhotoMaker [20] provide flexible frameworks for adapting existing ControlNet-trained modules or identity-preserved tokens to new diffusion backbones. Our semantic adapter (described in Section 3.3.1) inherits the zero-initialization stability principle of ControlNet [60] but operates on token-level conditioning rather than spatial feature maps, achieving a fundamentally more parameter-efficient design (∼>300M parameters). Unlike IP-Adapter, which targets text-image alignment via CLIP features [39], our adapter leverages DINOv2's [36] self-supervised features specifically for their sensitivity to degradation patterns.

2.4. Self-Supervised Visual Features for Restoration

Self-supervised learning (SSL) provides robust pre-trained priors for vision tasks. Early generative SSL, such as MAE [10] and SimMIM [52], focuses on structural reconstruction but often overlooks fine-grained textures. Similarly, contrastive and semantic frameworks like SimCLR [5], MoCo [9], and CLIP [39] provide high-level guidance but tend to discard low-level degradation patterns due to their focus on semantic invariance [30].

In contrast, dense self-distillation methods like DINO [4] and DINOv2 [36] preserve both semantics and local spatial nuances. These features have proven effective for zero-shot degradation assessment and quality analysis [15, 59]. To exploit this, DirectFlow introduces a parameter-efficient fine-tuning (PEFT) strategy. We combine a semantic adapter with low-rank adaptation (LoRA) integrated directly into the UNet's cross-attention and feed-forward blocks. While the pre-trained ControlNet-derived spatial pathways are optimized in Stage A and frozen in Stage B, to provide rigid, memory-efficient spatial guidance, the LoRA matrices grant the UNet the flexibility to learn a new attention distribu-

tion specifically for the degradation hints. This configuration actively guides the single-step transport and gracefully averts the catastrophic forgetting typical of full-network fine-tuning.

3. Method

We present **DirectFlow**, a framework designed to distill the transported posterior mean of a multi-step rectified flow teacher into a single-step latent consistency student. Fig. 3 provides a system overview. Our framework operates through three core mechanisms: **(1)** a frozen flow-matching teacher providing pre-computed restoration targets, **(2)** a dual parameter-efficient student featuring a Semantic Adapter and a LoRA-modulated UNet, and **(3)** a multi-objective training framework operating across latent and pixel dimensions. Our goal is to not copy teacher's output, but directly learn to predict the mathematically optimal endpoint of the perception-distortion trade-off in a single step. We optimize DirectFlow with a two-stage curriculum: Stage A aligns conditioning modules (visual/spatial encoders + semantic adapter), and Stage B refines transport via UNet LoRA with visual/spatial encoders frozen.

3.1. Preliminaries: Transported Posterior Mean

Given a degraded observation Y and a clean image X, the Minimum Mean Square Error (MMSE) estimator is defined:

$$\hat{X}^* = \mathbb{E}[X \mid Y] = \arg\min_{\hat{X}} \mathbb{E}\big[\|\hat{X} - X\|^2 \mid Y\big]. \quad (2)$$

This estimator minimizes distortion but, according to the perception-distortion trade-off [3, 35], necessarily deviates from the natural image manifold $\mathcal{M}$, producing smooth, over-regularized outputs.

A Rectified Flow [24, 26] establishes a deterministic transport from a source distribution to a target distribution via a learned velocity field. For image restoration, we utilize flows that transport the MMSE estimate directly to the data manifold:

$$x_t = (1-t)\, x_{\text{source}} + t\, x_{\text{target}}, \quad v_t = x_{\text{target}} - x_{\text{source}}, \quad (3)$$

where $x_{\text{source}} = \hat{X}^*$ and $x_{\text{target}} \sim p_{\text{data}}(\cdot \mid Y)$. Integrating the neural velocity field $v_\phi(x_t, t, Y)$ from t=0 to t=1 yields the transported posterior mean:

$$\hat{X}_0 = \hat{X}^* + \int_0^1 v_\phi(x_t, t, Y)\, dt. \quad (4)$$

This endpoint is provably the closest point on $\mathcal{M}$ to the posterior mean in the optimal transport sense [35, 44].

3.2. Transported Posterior Distillation

Let $T : \mathcal{Y} \to \mathcal{X}$ denote a teacher model that implements Eq. (4) over N steps. The teacher's output $X^\dagger = T(Y)$ serves as our explicit distillation target. Rather than training a student S_θ to regress to the ground truth X (which yields the blurry posterior mean) or to match individual posterior samples (which introduces stochastic distortion), we optimize:

$$S_\theta(Y) \approx T(Y) = X^\dagger. \quad (5)$$

This strategy is deliberate. The teacher's output is the optimal perceptual estimator. Distilling $X^\dagger$ ensures the student inherits this mathematically favorable position on the perception-distortion curve. By adopting the self-consistency formulation [43], the student learns to treat varying degradations simply as intermediate points along the trajectory, directly projecting them to their optimal high-quality origin.

Moreover, as the teacher produces pixel-space RGB outputs while the student operates in a compressed diffusion latent space [40], we construct a latent space bridge. A shared frozen VAE encoder $\mathcal{E}$ projects both the teacher's output and the student's input into a common representational domain, isolating semantic structural learning from low-level pixel reconstruction.

3.3. Degradation-Aware Conditioning

The input degradation space $\mathcal{Y}$ is a heterogeneous mixture of disparate manifolds, including blur, noise, and compression artifacts. A standard single-step model simply learns the average transport across these manifolds, which is geometrically suboptimal. We resolve this through a novel dual-adaptation strategy.

3.3.1. The Semantic Adapter

To explicitly condition the transport on the local geometry of the degradation, we introduce a parameter-efficient Semantic Adapter. **(1) Feature extraction.** The low-quality input Y is processed by a frozen self-supervised vision transformer (DINOv2 [36]) to extract a global degradation-aware token $e_{\text{dino}} \in \mathbb{R}^{768}$. **(2) Zero-initialized projection.** A trainable linear layer $\mathcal{Z}$ safely projects this token using zero-initialized weights $W^{(0)} = \mathbf{0}$ and $b^{(0)} = \mathbf{0}$. This ControlNet-style gating guarantees the student begins training identically to its pre-trained state. **(3) Additive injection.** The projected embedding is additively merged with the semantic content token from the frozen semantic context projector (a CLIP-based global contextualizer [17, 39]), forming the final cross-attention conditioning vector c.

3.3.2. Transport-Adaptive LoRA

A critical limitation of injecting novel semantic tokens into a frozen diffusion UNet is the network's inability to actively attend to the new signal. Because the base UNet attention mechanisms were strictly optimized for text embeddings (standard structural guidance), they inherently ignore the out-of-domain degradation vectors provided by the semantic adapter.

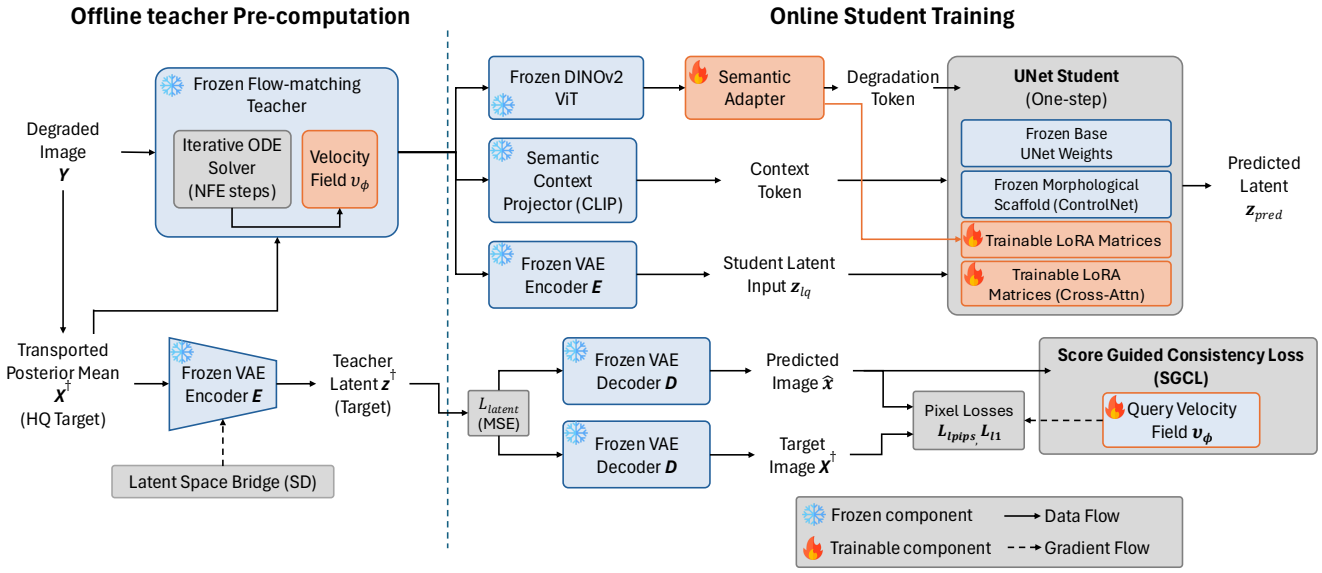

Figure 3. **System overview of DirectFlow. (Left)** During offline pre-computation, the rectified flow teacher processes each degraded input over N steps to produce the transported posterior mean; both input and output are projected into latent space via a frozen VAE encoder (the *space bridge*). **(Right)** During training, the student receives the cached LQ latent and produces a one-step prediction. A semantic adapter extracts degradation-aware features from the LQ RGB image and injects them into the UNet's cross-attention via zero-convolution gating.

To compel the network to utilize these degradation hints without triggering catastrophic forgetting of its foundational generative prior, we introduce Transport-Adaptive Low-Rank Adaptation. We apply these matrices exclusively to the cross-attention and feed-forward blocks of the UNet while completely freezing the heavy morphological conditioning prior. The morphological prior reliably enforces rigid geometric face constraints, whereas the lightweight LoRA matrices grant the UNet the necessary degrees of freedom to learn a new attention distribution. This learned distribution is specifically tailored to map the 1-step optimal transport jump while concurrently internalizing the semantic degradation hints.

A similar work LCM-LoRA [28] elegantly distills a pre-trained base model into a few-step generator by training an independent acceleration module. However, it operates strictly within the original model's native conditioning space, functioning purely as a temporal compression plugin. In contrast, our Transport-Adaptive LoRA serves a dual architectural imperative. It compresses the 25-step optimal transport trajectory and acts as a fundamental representational bridge. It structurally repurposes the cross-attention layers to process entirely out-of-domain features from the semantic adapter, optimizing specifically for the transported posterior mean rather than a standard generation trajectory.

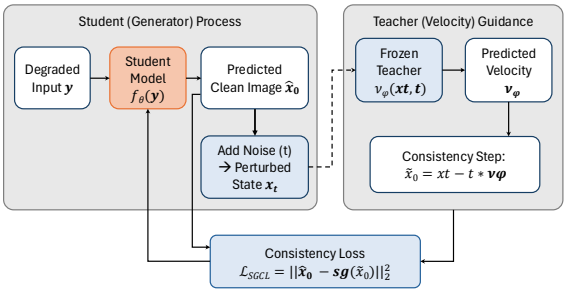

Figure 4. Score-Guided Consistency Loss (SGCL). The student output is perturbed, the teacher's velocity is queried, and a detached clean estimate forms the SGCL target; gradients flow back through the VAE decoder into the student latent.

3.4. Training Objectives

We employ a composite loss operating across both latent and pixel dimensions:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{lat}}\,\mathcal{L}_{\text{lat}} + \lambda_{\text{lpips}}\,\mathcal{L}_{\text{lpips}} + \lambda_{\text{l1}}\,\mathcal{L}_{\text{l1}} + \lambda_{\text{sgcl}}\,\mathcal{L}_{\text{sgcl}}. \quad (6)$$

The primary anchor loss $\mathcal{L}_{\text{lat}}$ minimizes the mean squared error directly between the student's predicted latent and the space-bridged teacher latent. To balance the distortion-perception tradeoff, $\mathcal{L}_{\text{lpips}}$ [61] and $\mathcal{L}_{\text{l1}}$ are computed by decoding the predicted latents through the VAE decoder while maintaining active gradient flow.

Table 1. Quantitative comparison on *synthetic* and *real-world* datasets. Best results are **bolded**; second-best are shaded gray. DirectFlow achieves state-of-the-art perceptual quality (LPIPS, PSNR) among single-step methods, highly competitive with 25-step generative models, while maintaining exceptional inference efficiency of under 300 ms per $I^{512\times512}$.

Method	*Synthetic*						*Real-world*						Time (sec)
	Celeba-Test						LFW-Test		WebPhoto-Test		WIDER-Test		
	LPIPS ↓	FID ↓	MUSIQ ↑	IDS ↑/↓	PSNR ↑	SSIM ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	
Input	0.574	145.22	72.81	0.32/47.94	22.72	0.706	138.87	26.87	171.63	18.63	201.31	14.22	–
PULSE† [31]	0.356	68.33	66.46	0.50/41.08	22.10	0.592	67.01	65.00	85.69	63.88	70.65	63.01	3.509
DFDNet† [18]	0.332	54.21	72.08	0.52/40.32	24.27	0.628	60.28	73.06	92.71	68.50	59.56	62.02	0.438
GFPGAN† [46]	0.230	49.84	73.90	0.56/38.79	24.64	0.688	50.36	73.57	87.47	72.08	39.45	72.79	0.312
GPEN† [54]	0.290	63.44	67.34	0.54/38.39	24.98	0.708	61.04	68.96	99.09	61.10	46.25	62.64	0.109
RestoreFormer† [49]	0.241	50.04	73.85	0.63/36.16	24.61	0.660	48.77	73.70	78.85	69.83	50.04	67.83	0.066
VQFR† [8]	0.245	41.84	75.18	0.64/35.74	24.06	0.660	51.33	71.74	75.77	72.02	44.09	74.01	0.177
CodeFormer† [63]	0.227	52.94	75.55	0.60/37.27	25.19	0.685	52.84	75.48	83.95	74.00	39.22	73.41	0.085
DR2⋆ [51]	0.264	54.48	67.99	0.42/44.00	25.03	0.617	45.71	71.50	109.24	62.37	48.20	60.28	1.775
DifFace⋆ [57]	0.272	**39.23**	68.87	0.48/41.84	24.80	0.684	46.31	69.76	80.86	65.37	37.74	65.02	3.248
PGDiff⋆ [53]	0.300	47.26	71.81	–/55.90	22.72	0.659	**44.65**	71.74	101.68	67.92	38.38	68.26	14.768
WaveFace⋆ [33]	0.362	57.58	–	0.67/34.50	23.56	0.691	53.88	73.54	78.01	70.45	37.23	72.89	19.370
InterLCM⋆ [17]	0.223	45.38	**76.58**	0.65/33.64	25.15	**0.718**	51.32	**76.16**	**75.48**	75.88	**35.43**	**76.29**	0.421
DirectFlow (Ours)⋆	**0.221**	40.96	76.12	**0.66/33.40**	**25.24**	0.714	45.08	75.92	75.92	75.42	36.91	75.11	**0.289**

† denotes regression-based models and ⋆ denotes generative models.

3.4.1. Score-Guided Consistency Loss (SGCL)

Inspired by Score Distillation Sampling [37], SGCL leverages the frozen teacher as a pixel-space distributional critic. By maintaining active gradient flow through the VAE decoder, SGCL pulls the predicted latent out of the VAE's blurry reconstruction bottleneck, driving the student to discover high-frequency "super-latents." [1]. **Formulation.** Under our Rectified Flow convention, $t = 0$ represents the source (noise/MMSE) and $t = 1$ the natural image target. We perturb the student's decoded RGB output $\hat{x}$ by injecting Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ at a randomly sampled timestep t:

$$x_t = t\hat{x} + (1-t)\epsilon. \tag{7}$$

Perturbation Band. To guarantee strong directional gradients toward the data manifold rather than minor pixel corrections, we use conservative stage-specific bands: ($t \in [0.05, 0.25]$) in Stage A and ($t \in [0.05, 0.20]$) in Stage B. **Optimization.** We query the frozen teacher's velocity field $v_\phi(x_t, t, Y)$ to form a clean-image estimate. To prevent explosive gradients from destabilizing the LoRA matrices in high-noise regimes, we clamp this target:

$$\hat{x}_{\text{clean}} = \text{Clamp}\big(x_t + (1-t)v_\phi,\ 0,\ 1\big). \tag{8}$$

The SGCL objective is computed as a Mean Squared Error:

$$\mathcal{L}_{\text{sgcl}} = \big\|\hat{x} - \text{sg}(\hat{x}_{\text{clean}})\big\|_2^2, \tag{9}$$

where $\text{sg}(\cdot)$ denotes a stop-gradient. Because $\hat{x}_{\text{clean}}$ is detached, gradients flow strictly from the pixel-space error back into the student's latent representation. Non-zero velocity provides a highly accurate directional gradient that pulls the one-step prediction out of blurry local minima.

[1] codes decoding to perceptual details sharper than the VAE encoding limit

Table 2. Quantitative evaluation on the CelebA-Test benchmark. Extended evaluation across perceptual and distortion metrics. DirectFlow establishes a strictly optimal balance on the perception-distortion curve for single-step architectures.

Method	Perceptual Quality				Distortion				
	FID ↓	KID ↓	NIQE ↓	Precision ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Deg ↓	LMD ↓
DOT [1]	100.2	0.0914	6.462	0.1600	21.32	0.6636	0.4756	43.87	2.876
RestoreFormer++ [50]	41.15	0.0290	4.187	0.6877	25.31	0.6703	0.3441	**29.63**	2.043
RestoreFormer [49]	42.30	0.0301	4.405	0.7010	24.62	0.6460	0.3655	32.13	2.299
CodeFormer [63]	53.16	0.0425	4.649	0.6940	25.15	0.6700	**0.3432**	37.28	2.470
VQFR [8]	41.79	0.0297	**3.693**	0.6593	24.07	0.6446	0.3515	35.75	2.429
GFPGAN [46]	46.72	0.0350	4.415	0.6970	24.99	0.6774	0.3643	36.05	2.443
DiffBIR [23]	59.06	0.0509	6.084	0.5643	25.39	0.6536	0.3878	32.94	2.006
BFRffusion [6]	41.53	0.0301	4.966	0.6623	26.21	0.6917	0.3619	30.98	**1.992**
PMRF [35]	**37.46**	**0.0257**	4.118	0.7073	**26.37**	**0.7073**	0.3470	30.67	2.030
DirectFlow (Ours)	37.98	0.0259	3.982	**0.7196**	26.26	0.6988	0.3450	30.21	1.998

4. Experiments

Datasets. Following standard protocols in blind face restoration, we train our DirectFlow student network exclusively on the FFHQ dataset [14], which contains 70,000 high-quality face images resized to 512×512. To synthesize low-quality (LQ) inputs during the offline teacher caching phase, we employ a comprehensive degradation pipeline mimicking real-world complex corruptions. This pipeline applies a random sequence of Gaussian blur, Gaussian noise, JPEG compression, and randomized downsampling-upsampling operations, parameterized consistently with Real-ESRGAN [47]. For evaluation, we utilize the synthetic CelebA-Test [27] bench-

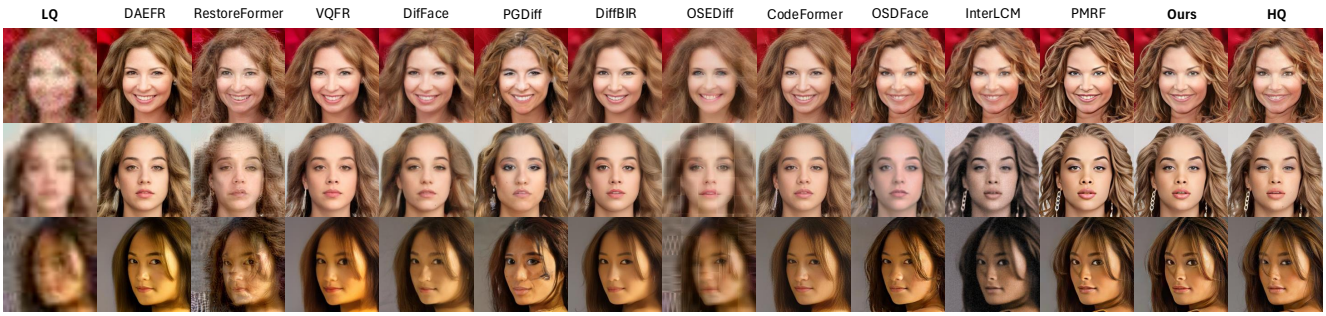

Figure 5. Qualitative comparison of DirectFlow against state-of-the-art blind face restoration methods on CelebA-Test dataset.

mark (3,000 images) to measure exact structural fidelity, alongside widely adopted real-world datasets including LFW-Test [13] and WebPhoto-Test [46] to assess out-of-domain generalization.

Evaluation Metrics. Given the inherently ill-posed nature of BFR, evaluation requires a holistic assessment of both the perception and distortion axes [3]. For distortion and structural alignment, we report PSNR, SSIM, and MSE. To quantify perceptual fidelity and naturalness, we report the FID [11] and LPIPS [61]. Furthermore, to evaluate subjective visual quality on in-the-wild images lacking ground truth, we incorporate the no-reference metric MUSIQ [15].

Implementation Details. Our frozen teacher is the pre-trained PMRF [35] architecture. The student network inherits the base weights of SD v1.5 [40]. The semantic adapter utilizes a frozen DINOv2-ViT-B/14 [36] backbone, and the transport-adaptive LoRA [12] matrices are initialized with a rank of $r = 32$. The morphological conditioning prior is initialized from pre-trained ControlNet [60]. We train DirectFlow with a two-stage curriculum on 70k latent-cached samples using AdamW and cosine learning-rate decay with bf16 mixed precision and EMA (0.999). In Stage A, we optimize the visual encoder, spatial encoder, and semantic adapter projection/gate, while keeping the UNet frozen (learning rate 2×10^{-5}, weight decay 5×10^{-5}). In Stage B, we initialize from Stage A checkpoints, freeze visual/spatial encoders, and optimize UNet-LoRA together with the semantic adapter projection/gate. SGCL is applied sparsely during training with a conservative perturbation band $t \in [0.05, 0.25]$, yielding stable optimization under one-step inference. Unless otherwise specified, all reported results use this Stage A$\rightarrow$Stage B schedule.

Quantitative Results. The quantitative evaluation on the CelebA-Test benchmark is presented in Tab. 1. DirectFlow establishes a strictly optimal balance on the perception-distortion curve for single-step architectures. While multi-step models like PMRF natively achieve high perceptual fidelity (low FID) at the cost of 25 Neural Function Evaluations (NFEs), DirectFlow achieves highly competitive FID and LPIPS scores in a single step, drastically outperforming previous 1-step distillation methods like InterLCM. Furthermore, Tab. 1 details performance on real-world datasets. DirectFlow achieves comparatively close to state-of-the-art perceptual quality, reflected by FID & MUSIQ scores, while maintaining exceptional inference efficiency with single-step execution (as illustrated in Fig. 2).

Qualitative Results. Fig. 5 illustrates the qualitative superiority of DirectFlow when presented with severe, in-the-wild degradations. Previous paradigms exhibit distinct, predictable failure modes under extreme corruption. GAN and regression-based priors (e.g., CodeFormer, RestoreFormer) often yield over-smoothed textures or noticeable identity shifts, failing to hallucinate realistic high-frequency data. Iterative generative models (e.g., DiffBIR, PMRF) successfully recover photorealistic structures but demand prohibitive multi-step sampling latency. Conversely, standard distillation approaches like InterLCM suffer from latent bottleneck blurring, leading to slightly waxy skin textures. DirectFlow successfully circumvents these limitations. By effectively utilizing the semantic adapter and the pixel-space SGCL critic, our method recovers precise, high-frequency details (such as individual hair strands, skin pores, and specular eye reflections) in a single forward pass. The resulting perceptual fidelity is highly commensurate with the multi-step teacher and the ground truth. Furthermore, to validate out-of-domain generalization, Fig. 7 presents results on the heavily degraded, historical WebPhoto-Test dataset. DirectFlow exhibits exceptional generalization to these severe, unstructured degradations. As highlighted in the extreme zoom-in patches (red boxes), previous baselines (e.g., CodeFormer, DiffBIR) often struggle with identity preservation or introduce structural hallucinations under extreme corruption. Conversely, DirectFlow faithfully reconstructs intricate facial geometry and textures in a single step, mapping these severe out-of-distribution states back to the correct facial manifold without succumbing to structural drift.

4.1. Ablation Study

To validate the necessity of our proposed architectural components, we conduct an ablation study on the CelebA-Test

Table 3. **Ablation study** on CelebA-Test. Each row removes one component from the full model.

Configuration	PSNR↑	LPIPS↓	FID↓	Note
DirectFlow (Full)	**25.24**	**0.221**	**40.96**	Complete model
w/o DINOv2 Adapter	24.98	0.232	44.18	No deg. awareness
w/o SGCL	25.07	0.228	42.63	No manifold guidance
w/o Zero-conv (scalar gate)	22.85	0.560	136.70	Stalled training
w/o Pixel losses	25.14	0.238	46.92	Latent-only
w/o color correction	25.18	0.226	42.11	Color drift

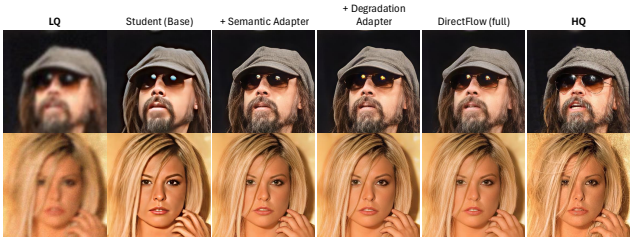

Figure 6. Qualitative ablation on blind face restoration (CelebA-Test). Left to right: LQ, Base, +Sem, +Deg, Full (SGCL), HQ.

benchmark, with results summarized in Tab. 3. We establish a baseline student model trained purely with latent MSE loss ($\mathcal{L}_{\text{lat}}$) and standard pixel reconstruction losses, lacking both the Semantic Adapter and the SGCL. Qualitative comparisons are provided in Fig. 6, where we progressively add the semantic and degradation adapters, followed by SGCL in the full model. Overall, the semantic adapter mainly improves global facial structure and alignment, the degradation adapter improves robustness across diverse corruptions, and SGCL further restores high-frequency details (e.g., skin texture and hair strands) while preserving identity and facial geometry.

Impact of the Semantic Adapter. Removing the DINOv2 Semantic Adapter fundamentally forces the network to learn an average transport trajectory across all degradation manifolds. As shown in Tab. 3, this omission results in a noticeable degradation in perceptual fidelity (a sharp increase in FID). Without explicit degradation routing, the student fails to distinguish between heavy blur and severe noise, leading to sub-optimal detail hallucination.

Impact of Score-Guided Consistency Loss. The removal of SGCL directly exposes the student to the latent VAE bottleneck. While the model remains efficient, the LPIPS score increases significantly, and the outputs visually lack the sharp "super-latent" details discussed in Sec. 3.4.1. The inclusion of the SGCL acts as a crucial distributional critic, effectively providing the directional gradients necessary to pull the one-step prediction onto the natural image manifold.

4.2. Broader Impacts and Carbon Footprint

To calculate the precise carbon footprint of reproducing DirectFlow, we adopt the standard estimation methodology established in DINOv2 [36]. We assume a peak power consumption of 400W per A100 GPU, a standard data center Power Usage Effectiveness (PUE) of 1.1, and a national average carbon intensity factor of 0.385 kg CO_2eq per kWh. As detailed in Tab. 4, the total energy required to fully reproduce DirectFlow is approximately 0.084 MWh, resulting in 0.033 tons of CO_2 equivalent emissions. By distilling the transported posterior mean into a parameter-efficient two-stage optimization schedule, our framework not only resolves the perception-distortion trade-off in a single NFE, but does so with a carbon footprint that is orders of magnitude smaller than training standard flow-matching or multi-step generative priors.

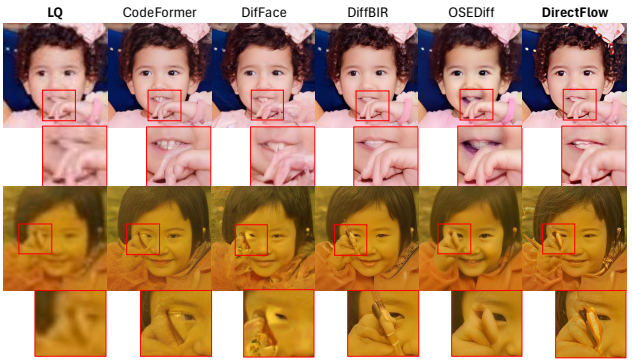

Figure 7. Qualitative comparison on real-world, out-of-domain historical photos from WebPhoto-Test dataset.

Table 4. Carbon footprint of reproducing DirectFlow.

Model	Equipment	GPU Hours	Power (W)	Total Energy (MWh)	CO_2eq (tCO_2eq)
DirectFlow (Ours, Stage A + Stage B)	A100	192	400	0.084	0.033

5. Conclusion

In this work, we introduced **DirectFlow**, a highly efficient framework for blind face restoration that reconciles the perception-distortion trade-off in a single forward pass that is highly suitable for edge-device deployments. By distilling the transported posterior mean of a flow-matching teacher into a latent consistency student, we eliminate the prohibitive latency of iterative ODE solvers. To handle diverse, in-the-wild degradations without catastrophic forgetting, we proposed a parameter-efficient dual-adaptation strategy combining a semantic adapter with transport-adaptive LoRA matrices. Furthermore, our novel score-guided consistency loss actively queries the frozen teacher as a pixel-space critic, propelling the student's output beyond standard latent reconstruction limits. Extensive experiments demonstrate that DirectFlow achieves state-of-the-art perceptual quality and structural fidelity among single-step methods, paving the way for real-time, high-fidelity restoration in practical downstream applications.

References

[1] Theo Adrai, Guy Ohayon, Michael Elad, and Tomer Michaeli. Deep optimal transport: A practical algorithm for photo-realistic image restoration. *Advances in Neural Information Processing Systems*, 36:61777–61791, 2023. 6

[2] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22324–22333, 2023. 1

[3] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proceedings of the 36th International Conference on Machine Learning*, pages 675–685. PMLR, 2019. 1, 2, 4, 7

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 3

[6] Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaocun Cao. Towards real-world blind face restoration with generative diffusion prior, 2023. 6

[7] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2018. 2

[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 *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 2, 6

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 3

[11] 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. 1, 7

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 3, 7

[13] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 7

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Wu, and Ming-Hsuan Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 7

[16] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhiyuan Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. In *Neurocomputing*, pages 47–59, 2022. 2

[17] Senmao Li, Kai Wang, Joost van de Weijer, Fahad Shahbaz 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 *ICLR*, 2025. 2, 3, 4, 6

[18] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020. 6

[19] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 3

[20] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 3

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1

[22] Changlin Lin et al. Ctrl-adapter: An efficient and versatile framework for adapting diffusion models to new controls. *arXiv preprint arXiv:2404.00000*, 2024. 3

[23] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. In *CVPR*, 2024. 1, 2, 6

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR 2023*, 2023. 2, 4

[25] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[26] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 2, 4

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 6

[28] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick Von Platen, Apolina kario Passos, Longbo Huang, Jian Li,

and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 3, 5

[29] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2024. 2

[30] Ziwei Luo et al. Towards degradation-aware visual assistant: A survey on restoration-oriented ssl. *arXiv preprint arXiv:2401.xxxxx*, 2024. 3

[31] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[32] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, pages 11913–11924. Curran Associates, Inc., 2020. 1

[33] Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. *CVPR*, 2024. 6

[34] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Gui-Song Xia. T2i-adapter: Learning adapters to dig out t2i diffusion models for adaptable control. In *CVPR*, 2024. 3

[35] Guy Ohayon, Tomer Adrai, Michael Elad, and Tomer Michaeli. Posterior-mean rectified flow: Towards minimum mse photo-realistic image restoration. In *International Conference on Machine Learning (ICML)*, 2024. 1, 2, 4, 6, 7

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Vasiljevic, Panteleimon Sun, Gautier Pinto, Maxim Krishnan, Gabriel Synnaeve, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 4, 7, 8

[37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 6

[38] Can Qin et al. Unicontrol: A unified conditional diffusion model with fine-grained control. *arXiv preprint arXiv:2305.11147*, 2023. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Prowizor Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 4

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4, 7

[41] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 2

[42] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVI*, page 87–103, Berlin, Heidelberg, 2024. Springer-Verlag. 3

[43] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Shamir. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2, 4

[44] Céric Villani. *Optimal Transport: Old and New*. Springer, 2009. 2, 4

[45] 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*, 132(12):5929–5949, 2024. 1, 2

[46] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 2, 6, 7

[47] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 6

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 1

[49] 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)*, pages 17491–17500, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 6

[50] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15462–15476, 2023. 6

[51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023. 1, 2, 6

[52] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 3

[53] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023. 6

[54] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 672–681, 2021. 2, 6

[55] Hu Ye, Junping Zhang, Sanyuan Liu, Wei Han, and Xiao Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[56] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Artur d'Avila Garcez. Face hallucination with finite

state machine and facial landmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[57] Zongsheng Yue and Chen Change Loy. Diface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9991–10004, 2024. 6

[58] Lingbo Zhan et al. X-adapter: Adding predictions to frozen diffusion models. *arXiv preprint arXiv:2401.00000*, 2024. 3

[59] Kai Zhang et al. Revisiting self-supervised features for blind image quality assessment. *arXiv preprint arXiv:2402.xxxxx*, 2024. 3

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3, 7

[61] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 5, 7

[62] Shihao Zhao, Dongdong Chen, Yuan-Chih Chen, Jianmin Bao, Pengchuang Shao, Enze Tan, Bin Liao, Gang Zhou, Lu Liu, Zehua Yuan, et al. Uni-controlnet: All-in-one control adapter for text-to-image diffusion models. *NeurIPS*, 2024. 3

[63] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, pages 30599–30611, 2022. 2, 6