Flow Straight to Reality: Perceptually Consistent Flow Matching for Efficient Image Restoration

Sangwoo Jo[1], Donggeun Ko[2], Jayeon Kang[1], Youngsang Kwak[2], Jaehwa Kwak[2], and Sungjoon Choi[1]

[1] Korea University, Seoul, South Korea
{jasonjo97, nature0213, sungjoonc}@korea.ac.kr
[2] Aim Future, Seoul, South Korea
{sean.ko, youngsang.kwak, jaehwa.kwak}@aimfuture.ai

Abstract. Image restoration is fundamentally constrained by the tradeoff between distortion and perception: minimizing pixel-wise error yields over-smoothed results, whereas optimizing for perceptual realism often introduces structural deviations. Recent approaches attempt to balance this tradeoff via posterior sampling or multi-stage generative pipelines, yet remain computationally expensive and architecturally complex. To overcome these limitations, we propose PCFlow (Perceptually Consistent Flow Matching), a unified framework that directly parameterizes a continuous transport from degraded observations to clean targets, jointly optimizing distortion and perceptual quality. While its latent consistency flow objective drives stable and efficient few-step inference, a Latent Consistency Perceptual Loss (LCPL) imposes semantic constraints directly on the guiding velocity field, steering the dynamics toward visually sharp data manifolds. Furthermore, recognizing the inherent conflict between structural and perceptual consistencies, we integrate a conflict-free gradient projection strategy to stabilize the multi-objective optimization landscape. Combined with lightweight, convolution-only backbone, PCFlow achieves competitive performance across diverse restoration tasks at a fraction of traditional computational costs.

1 Introduction

Image restoration aims to recover a clean image from its degraded observation, constituting a fundamental class of inverse problems in computer vision. Fundamentally, this process is plagued by the inherent ambiguity of degradation: multiple plausible high-quality reconstructions can map to a single corrupted input. As established in classical literature [4, 10], this ambiguity forces a strict *distortion-perception tradeoff*. Methods that minimize distortion error (*e.g.*, MSE) mathematically collapse to the conditional expectation, yielding over-smoothed results. Conversely, approaches optimizing for perceptual realism (*e.g.*, FID) hallucinate high-frequency details, improving visual quality but incurring severe structural deviations and high reconstruction error.

A principled way to achieve perceptually optimal solutions is posterior sampling, which generates samples from the underlying posterior distribution and

Fig. 1: Visual Results of PCFlow on the Blind Face Restoration (BFR). PCFlow consistently produces visually faithful and high-quality reconstructions while requiring only a few inference steps ($K = 5$).

thus attains the optimal perceptual index [4, 10]. However, such methods do not minimize distortion and, in expectation, yield an MSE that can be up to twice the minimum achievable error. Recent diffusion and score based restoration approaches leverage generative priors to approximate posterior sampling [5, 13, 21, 23, 26, 33]. While these approaches significantly improve perceptual quality, they rely on iterative stochastic sampling and typically require multiple sampling steps, resulting in substantial computational cost at inference time. To circumvent this, recent multi-stage frameworks decompose the problem: they first predict a fidelity-oriented minimum mean-squared error (MMSE) estimate, and subsequently apply heavy generative refinement to recover details [6, 16, 19, 29]. While effective, such two-stage pipelines inherit architectural complexity and still rely on costly generative steps during inference.

In this work, we revisit the problem from a different perspective. Instead of performing stochastic posterior sampling or decomposing restoration into multiple stages, we propose **PCFlow** (Perceptually Consistent Flow Matching), a unified direct transport framework that effectively balances distortion and perception. Rather than decomposing the restoration process, we tackle the multi-objective optimization directly within a continuous latent space.

At its core, PCFlow parameterizes a direct vector field from the degraded input to the clean target. By employing a latent consistency flow matching (LCFM) objective, we enforce geometric consistency along the trajectory, effectively straightening the integration path and enabling image restoration in as few as three inference steps. However, learning a few-step transport with an L_2-based objective inherently risks regression to the posterior mean. To mitigate this effect, we introduce a Latent Consistency Perceptual Loss (LCPL), which encourages the trajectory endpoints to align with perceptually sharp, high-density data

manifolds. Nevertheless, simply combining these objectives introduces gradient conflicts, particularly at early transport timesteps (low SNR regimes). To stabilize this multi-objective optimization, we adopt a *conflict-free gradient projection* strategy along with an SNR-adaptive schedule, using the perceptual objective as a steering signal while removing structural gradient components that conflict with perceptual optimization.

Furthermore, PCFlow employs a lightweight convolution-only model architecture without attention modules, significantly reducing model size and computational cost. Combined with our conflict-free consistency training, PCFlow operates at a fraction of traditional computational costs while achieving effective balance on the distortion-perception curve.

Our main contributions are summarized as follows:

- We propose PCFlow, a unified direct transport framework that integrates perceptual consistency into a latent flow objective, enabling highly efficient few-step image restoration.
- We analyze the destructive gradient interference between structural and perceptual objectives, and introduce a conflict-free, SNR-adaptive gradient alignment method to stabilize training.
- We adopt a convolution-only architecture that rivals computationally heavy multi-stage diffusion pipelines, significantly reducing parameter count and inference time.
- Extensive evaluations across several benchmarks, including blind face restoration, super-resolution, denoising, inpainting, and colorization, demonstrate that PCFlow advances the distortion-perception tradeoff frontier.

2 Related Work

2.1 Generative Models for Image Restoration

Image restoration has undergone a paradigm shift with the advent of generative modeling. Primary works have formulated the problem through the lens of Bayesian posterior sampling [5, 13, 21, 23, 26, 33]. By iteratively drawing samples from the posterior $p(x \mid y) \propto p(x)\, p(y \mid x)$ using diffusion priors, these models attain exceptional perceptual quality. Beyond explicit posterior sampling, conditional generative frameworks have been also explored [16,32]. Instead of deriving a posterior via likelihood models, these approaches introduce various mechanisms to incorporate degraded images as conditional inputs.

Recent works attempt to bridge distortion minimization and perceptual realism through a disjointed, two-stage pipeline [6,19]. Motivated by the distortion-perception tradeoff theory [4, 10], they first anchor the generation to the minimum mean-squared error (MMSE) estimator,

$$x^* = \mathbb{E}[x \mid y], \tag{1}$$

which mathematically guarantees minimal distortion. Subsequently, they must learn an optimal transport from this safe, over-smoothed estimate x^* to the target distribution x. To execute this transport, approaches like DOT [1] attempt

to bypass iterative sampling by directly approximating the transport trajectory with a closed-form solution under Gaussian assumptions. Although effective, such two-stage pipelines introduce additional architectural complexity and computational cost at inference time.

2.2 Consistency Flow Matching

Flow-based generative models [2, 17, 18] have recently emerged as an alternative class of generative models that learn a continuous vector field $v(\mathbf{x}, t)$ transporting samples between two distributions $\mathbf{x}_0 \sim \pi_0$ and $\mathbf{x}_1 \sim \pi_1$ through an ordinary differential equation (ODE):

$$\frac{d\mathbf{x}}{dt} = v(\mathbf{x}(t), t), \tag{2}$$

In other words, flow-based approaches directly parameterize the transport vector field between distributions, enabling explicit trajectory modeling and efficient inference through numerical ODE integration.

To further improve training and inference efficiency, Consistency Flow Matching (CFM) [27] introduces a consistency objective that effectively straightens the flow trajectory by aligning predictions across neighboring timesteps. Instead of explicitly modeling the entire probability path, CFM enforces consistency in both trajectory outputs and velocity field, leading to stable training and supporting few step inference. While CFM has demonstrated strong performance in unconditional generative modeling, it has primarily been studied in pixel space. Its extension to conditional generative tasks, such as image restoration, as well as to latent representation spaces remains relatively underexplored.

2.3 Perceptual Objective

Perceptual objectives have been widely adopted in image restoration to enhance high-frequency details [11, 14, 24, 25, 31]. A common approach is to incorporate feature-based perceptual losses, most notably LPIPS [30], which measures the distance between pretrained VGGNet features [22] extracted from predicted and ground truth images. Such external perceptual losses encourage semantic similarity beyond pixel-wise fidelity and have become standard regularizers in image restoration tasks. E-LatentLPIPS [12] extends LPIPS to latent space, where the training objective remains identical but with additional data augmentation to address the suboptimal loss landscape inherent in latent representations.

More recently, several works have explored leveraging internal features of generative models themselves for perceptual supervision, instead of relying on external pretrained networks. These methods utilize intermediate features, such as U-Net midblock layer or decoder features, to define self-perceptual losses [3, 15]. By aligning intermediate features, such approaches aim to better preserve generative priors and semantic coherence. However, such methods have been developed primarily for image generation, and their extension to restoration settings remains limited.

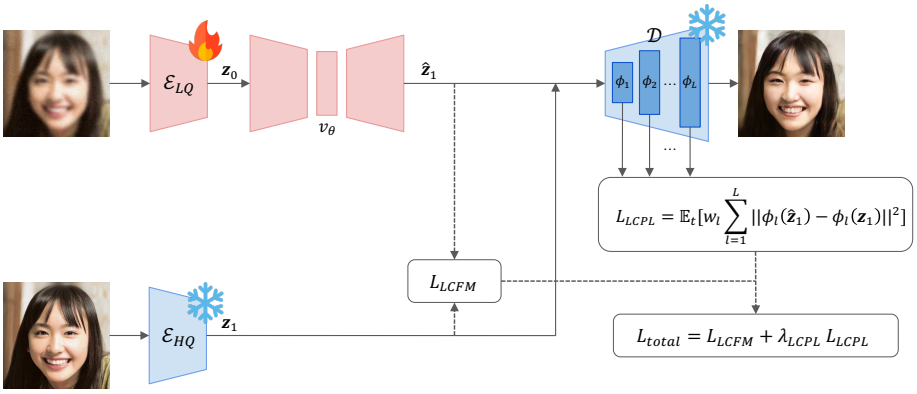

Fig. 2: Overview of PCFlow. Given a degraded image, our proposed PCFlow encodes it into latent $\mathbf{z}_0$ using the LQ encoder, and learns the latent transport with a flow model v_θ. The restored latent prediction $\hat{\mathbf{z}}_1$ is then decoded by $\mathcal{D}$ to obtain the final output image $\hat{\mathbf{x}}$. The training objective combines a latent consistency flow matching loss L_{LCFM} with a latent consistency perceptual loss L_{LCPL}, computed from multi-level decoder features $\{\phi_l\}_{l=1}^{L}$, yielding the final objective as $L_{\text{total}} = L_{\text{LCFM}} + \lambda_{\text{LCPL}} L_{\text{LCPL}}$.

3 Method

3.1 Latent Consistency Flow Matching

Formulating Latent Transport. Following ELIR [6], we begin by defining a latent consistency flow matching (LCFM) objective that integrates latent flow matching [7] with consistency training [27]. Given a low-quality (LQ) latent $\mathbf{z}_0$ and its corresponding high-quality (HQ) latent $\mathbf{z}_1$, we parameterize a vector field v_θ that governs the latent transport over time $t \in [0, 1]$:

$$\frac{dz}{dt} = v_\theta(\mathbf{z}(t), t), \tag{3}$$

where we define a linear interpolation path in the latent space:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1 - (1 - \sigma_{\min})\, t)\, \mathbf{z}_0, \quad t \in [0, 1]. \tag{4}$$

Here, $\sigma_{\min} > 0$ prevents degeneration at early timesteps and stabilizes transport learning.

Given the number of segments K, we partition the time interval $[0, 1]$ into K subintervals $\left\{\left[\frac{i}{K}, \frac{i+1}{K}\right]\right\}_{i=0}^{K-1}$, and let $i = \lfloor Kt \rfloor$ denote the segment index. The LCFM objective then penalizes discrepancies in both the predicted trajectory endpoints and the velocity fields:

$$L_{\text{LCFM}}(\theta) = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t}\left[\Delta f_\theta^i(\mathbf{z}_t, \mathbf{z}_{t+\Delta t}, t) + \alpha \Delta v_\theta^i(\mathbf{z}_t, \mathbf{z}_{t+\Delta t}, t)\right], \tag{5}$$

where $t \sim \mathcal{U}(0, 1 - \Delta t)$ and the penalty terms are defined as:

$$\Delta f_\theta^i(\mathbf{z}_t, \mathbf{z}_{t+\Delta t}, t) = \left\| f_\theta^i(\mathbf{z}_t, t) - \mathrm{sg}\left(f_\theta^i(\mathbf{z}_{t+\Delta t}, t + \Delta t)\right)\right\|^2, \tag{6}$$

$$\Delta v_\theta^i(\mathbf{z}_t, \mathbf{z}_{t+\Delta t}, t) = \left\| v_\theta^i(\mathbf{z}_t, t) - \mathrm{sg}\left(v_\theta^i(\mathbf{z}_{t+\Delta t}, t + \Delta t)\right)\right\|^2, \tag{7}$$

$$f_\theta^i(\mathbf{z}_t, t) = \mathbf{z}_t + \left(\frac{i+1}{K} - t\right) v_\theta^i(\mathbf{z}_t, t). \tag{8}$$

Here, $v_\theta^i(\mathbf{z}_t, t)$ denotes the predicted vector field within the i-th segment, and $f_\theta^i(\mathbf{z}_t, t)$ represents the predicted trajectory output obtained by a single Euler step toward the end of the segment. $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator, and $\alpha, \Delta t$ are set as hyperparameters.

Unconditional and Conditional Training. Following PMRF [19], we consider both conditional and unconditional flow models for image restoration. For the unconditional flow model, the flow is initialized from the degraded observation, and the transport is defined as:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1 - (1 - \sigma_{\min})\, t)\mathbf{z}_0^*, \quad \mathbf{z}_0^* = \mathbf{z}_0 + \sigma_s \epsilon, \tag{9}$$

where the vector field $v_\theta(z_t, t)$ learns to transport samples from this initialization toward the clean image distribution. Here, σ_s controls the standard deviation of the Gaussian noise that alleviates singular mapping between low and high dimensional manifolds.

For comparison, we also consider conditional flow formulation where the vector field learns from the noise distribution conditioned on the degraded input $\mathbf{z}_0$, formulated as the following:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1 - (1 - \sigma_{\min})\, t)\mathbf{z}_0^*, \quad \mathbf{z}_0^* \sim \mathcal{N}(0, I), \tag{10}$$

where the corresponding velocity field $v_\theta(\mathbf{z}_t, t, \mathbf{z}_0)$ learns to transport samples from the noise distribution toward the target clean image conditioned on $\mathbf{z}_0$.

3.2 Latent Consistency Perceptual Loss

While LCFM ensures geometric consistency in latent space, such a method does not explicitly enforce perceptual realism. To incorporate semantic alignment, we introduce a Latent Consistency Perceptual Loss (LCPL).

External Perceptual Network. Given a predicted latent $\hat{\mathbf{z}}_1$ and the ground-truth latent $\mathbf{z}_1$, we adopt E-LatentLPIPS [12] as an external perceptual objective, where the perceptual distance is computed between the corresponding latent representations by applying differentiable augmentation and measuring feature discrepancies in a pretrained network:

$$L_{\text{external}}(\mathbf{z}_1, \hat{\mathbf{z}}_1) = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t, \mathcal{T}}\left[\left\| f_{\text{VGG}}\left(\mathcal{T}(\mathbf{z}_1)\right) - f_{\text{VGG}}\left(\mathcal{T}(\hat{\mathbf{z}}_1)\right)\right\|^2\right], \tag{11}$$

where f_{VGG} represents a VGG network pretrained on ImageNet [8] and BAPPS dataset [30], and $\mathcal{T}$ denotes random differentiable augmentations implemented

in E-LatentLPIPS. We separately train VGGNet for 256×256 resolution for our experiments, as the publicly available pretrained model from E-LatentLPIPS is trained on 512×512 images.

Internal Perceptual Network. Although E-LatentLPIPS [12] provides strong perceptual supervision, it relies on an externally pretrained network and may require additional training for dataset-specific tasks. To reduce dependency on external modules, we additionally adopt an internal perceptual objective function defined by the model itself [3].

Specifically, let $\{\phi_l(\mathbf{z})\}_{l=1}^{L}$ denote intermediate decoder features extracted from latent feature z. The internal perceptual loss is then defined as:

$$L_{\text{internal}}(\mathbf{z}_1, \hat{\mathbf{z}}_1) = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t}\left[w_l \sum_{l=1}^{L} \left\| \hat{\phi}_l(\mathbf{z}_1) - \hat{\phi}_l(\hat{\mathbf{z}}_1) \right\|^2 \right], \tag{12}$$

where $\hat{\phi}_l(\cdot)$ denotes per-channel normalized features and $\{w_l\}_{l=1}^{L}$ are weight values assigned to each feature layer. Note that the formulation is similar to LPL loss [3] except that binary map masking is omitted for simplicity.

Perceptual Objective for Latent Consistency Training. To incorporate perceptual alignment into consistency learning, we define a Latent Consistency Perceptual Loss (LCPL) that enforces perceptual similarity between adjacent trajectory predictions:

$$L_{\text{LCPL}}(\theta) = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t}\left[L_{\text{percep}}\left(f_\theta^i(\mathbf{z}_t, t), f_\theta^i(\mathbf{z}_{t+\Delta t}, t+\Delta t) \right) \right], \tag{13}$$

where $t \sim \mathcal{U}(0, 1-\Delta t)$ and L_{percep} can be instantiated as either L_{external} or L_{internal}. This objective encourages perceptual consistency of the predicted trajectory across neighboring timesteps in latent space.

Overall Objective. Hence, our final training objective combines flow consistency and perceptual consistency as the following:

$$L_{\text{total}}(\theta) = L_{\text{LCFM}}(\theta) + \lambda_{\text{LCPL}}\, L_{\text{LCPL}}(\theta), \tag{14}$$

where λ_{LCPL} controls the strength of perceptual steering relative to the structural transport objective.

3.3 Improving Objective with Conflict-Free Gradient Alignment

Diagnosing Gradient Conflict. We compute the gradients of the latent consistency flow matching objective $\nabla_\theta L_{\text{LCFM}}$ and perceptual objective $\nabla_\theta L_{\text{LCPL}}$ defined in Eq. (14). We observe that the gradients between the structural and the perceptual objective are highly noise-level dependent: the two gradients conflict in low log-SNR regimes but become increasingly aligned in high log-SNR regimes. This indicates that the two objectives induce conflicting descent directions in the parameter space, particularly in early transport stages. However, naively summing the gradients implicitly assumes

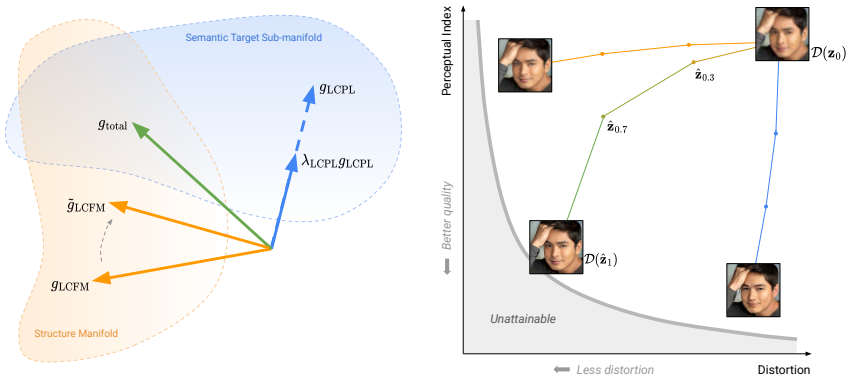

Fig. 3: Gradient alignment between reconstruction and perceptual objectives. (Left) The reconstruction objective L_{LCFM} promotes structural fidelity, while the perceptual objective L_{LCPL} encourages perceptual realism toward a semantic target sub-manifold. Our gradient alignment method mitigates conflicts between the gradients and produces a balanced update g_{total}. (Right) In the distortion-perception plane, relying solely on LCFM (orange) leads to blurry outputs, whereas over-optimizing LCPL (blue) severely deviates from the input structure and introduces artifacts. In contrast, our aligned trajectory (green) incorporates perceptual guidance while mitigating conflicting structural updates, allowing the model to move closer to the optimal PD limit.

$$\langle \nabla_\theta L_{\text{LCFM}}, \nabla_\theta L_{\text{LCPL}} \rangle \geq 0, \tag{15}$$

which does not hold in practice. When the inner product is negative, the two objectives produce conflicting optimization directions, leading to destructive gradient interference and unstable optimization. This motivates our *conflict-free* gradient update method combined with noise-aware weighting strategy.

λ-scheduling. Recognizing that the gradient alignment naturally improves in high SNR regimes (near the clean target), we modulate the perceptual influence dynamically. Instead of a static hyperparameter, we define $\lambda_{\text{LCPL}}(t)$ as a monotonically increasing function of the timestep t. Conceptually, this ensures that the model establishes a robust, conflict-free structural foundation during the noisy early stages, and progressively unleashes the full power of perceptual steering as the generation refines toward reality.

Conflict-Free Gradient Update. To resolve this destructive interference, we draw inspiration from multi-task gradient surgery [28] but adapt it as an *asymmetric orthogonal projection*. While the LCFM objective provides the structural transport signal, the LCPL objective serves as perceptual guidance that steers the transport trajectory toward perceptually realistic regions of the target manifold. Therefore, when the two vectors conflict (i.e., $\langle g_{\text{LCFM}}, g_{\text{LCPL}} \rangle < 0$), we preserve the perceptual gradient and remove only the component of the structural gradient that conflicts with it.

Denoting the gradients of the two objectives as $g_{\text{LCFM}} = \nabla_\theta L_{\text{LCFM}}$ and $g_{\text{LCPL}} = \nabla_\theta L_{\text{LCPL}}$, we update the parameters as the following procedure:

$$\theta \leftarrow \theta - \eta(\tilde{g}_{\text{LCFM}}(t) + \lambda_{\text{LCPL}}(t) g_{\text{LCPL}}(t)), \tag{16}$$

where $\tilde{g}_{\text{LCFM}}$ is defined as the orthogonal component of g_{LCFM} with respect to g_{LCPL}:

$$\tilde{g}_{\text{LCFM}}(t) = g_{\text{LCFM}}(t) - \mathbf{1}_{\{\langle g_{\text{LCFM}}, g_{\text{LCPL}} \rangle < 0\}} \frac{\langle g_{\text{LCFM}}, g_{\text{LCPL}} \rangle}{\|g_{\text{LCPL}}\|^2} g_{\text{LCPL}}. \tag{17}$$

As a result, the perceptual objective remains an effective steering signal throughout training, whereas reconstruction updates are incorporated only when they do not interfere with perceptual optimization. This asymmetric design enables conflict-free multi-objective optimization and leads to improved perceptual restoration performance. We further provide an ablation study comparing alternative projection strategies in our supplementary material.

4 Experiments

We evaluate PCFlow on the following tasks: blind face restoration (BFR), super-resolution, denoising, inpainting, and colorization. Following previous baselines [6, 19], for BFR, models are trained on FFHQ 512×512 dataset and evaluated on CelebA-Test, LFW-Test and CelebAdult. For the remaining tasks, models are trained on FFHQ 256×256 dataset and evaluated on CelebA-Test dataset.

4.1 Implementation Details

Training. We train PCFlow using a two-stage procedure. Initially, the model focuses solely on reconstruction quality by setting $\lambda_{\text{LCPL}} = 0$ for the first 250 training epochs. In the second stage, we enable the perceptual objective and continue training for an additional 250 epochs with the proposed SNR-adaptive weighting, encouraging fine-grained details and perceptual realism. We experiment with different λ-scheduling strategies and report the best-performing configuration. For consistency training, we follow ELIR setting the number of consistency steps $K = 5$ for BFR and $K = 3$ for the remaining tasks, with a fixed time interval $\Delta t = 0.05$. We use a batch size of 128 using AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with weight decay 0.02. We use exponential moving average (EMA) with decay rate 0.999 throughout training.

Model Architecture. We employ a pretrained Tiny AutoEncoder [20], a lightweight version of Stable Diffusion VAE [9]. The encoder-decoder operates in a 16-channel latent space and contains approximately 2.4M parameters, providing efficient latent representation while maintaining high reconstruction fidelity. We implement convolution-only U-Net architecture from ELIR. The model takes 32 input channels for conditional setting and 16 input channels for unconditional

Table 1: Quantitative results on BFR. Quantitative comparison of PCFlow with baselines on blind face restoration (BFR) task. The best and second-best results are highlighted in **bold** and underlined, respectively. PCFlow achieves a favorable restoration quality with efficient model architecture and fast inference, obtaining state-of-the-art FID and NIQE scores for CelebA-Test, and FID for CelebAdult dataset.

Model	Efficiency		CelebA-Test						LFW	CelebAdult
			Perceptual Quality			Distortion			Perceptual Quality	
	#Params[M]↓	FPS↑	FID↓	NIQE↓	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID↓
CodeFormer	94	27.13	52.23	4.65	**75.55**	24.77	0.6732	**0.3432**	54.28	114.34
GFPGAN(v1.3)	87.14	**59.73**	45.95	4.42	75.29	24.60	0.6802	0.3643	49.58	112.31
VQFRv2	83.49	16.97	46.01	4.17	74.40	22.85	0.6446	0.3624	52.50	106.83
Diffface ($K = 100$)	182.07	0.78	37.43	4.37	68.31	24.44	0.6579	0.4172	47.23	101.11
DiffBIR ($K = 50$)	1666.93	0.38	46.23	4.21	75.27	23.27	0.6379	0.3876	**46.03**	111.84
ResShift ($K = 4$)	196.70	6.85	43.60	4.37	72.16	25.32	0.6965	0.3435	54.23	109.04
PMRF ($K = 25$)	182.75	0.57	37.22	4.12	70.36	**25.85**	**0.7098**	0.3470	49.98	104.44
ELIR ($K = 5$)	37.52	33.11	44.64	5.26	67.24	25.56	0.7030	0.3735	53.19	105.55
PCFlow (Ours)	**32.02**	42.62	**35.89**	**3.95**	70.35	24.44	0.6680	0.3850	50.68	**98.85**

setting. Note that the overall architecture is identical to ELIR, but without the MMSE estimator module, resulting in reduction of 5.5M parameters.

Latent Perceptual Network. For the internal perceptual network, we extract features from the decoder's intermediate representations, including the mid-block, three upsampling stages, and the final output layer. The contribution of each feature level is weighted according to its spatial resolution, except for the final output layer, *i.e.*, $w_l \propto 2^{-r_l}$, where r_l denotes the resolution level of feature l. The weights are subsequently normalized to sum to one. For the external perceptual network, to compute the perceptual objective within our customized latent space, we separately train the model following the procedure from E-LatentLPIPS. In addition, we adopt batch normalization which empirically demonstrates more stable training compared to group normalization.

4.2 Quantitative Results

Blind Face Restoration. Quantitative results on blind face restoration (BFR) are reported in Tab. 1. PCFlow achieves state-of-the-art perceptual quality, obtaining the best FID and NIQE on CelebA-Test and the best FID on CelebAdult, while requiring only 32M parameters and five sampling steps ($K = 5$). Compared to ELIR, PCFlow uses fewer parameters (32M vs. 37.5M) and achieves 1.29× higher inference speed (42.62 vs. 33.11 FPS), while substantially improving FID (35.89 vs. 44.64) on CelebA-Test. Furthermore, despite being significantly smaller and faster than diffusion-based baselines such as PMRF (183M parameters, 0.57 FPS), PCFlow outperforms in perceptual quality metrics while delivering over 75× higher throughput. These results demonstrate that PCFlow provides a favorable trade-off between restoration quality and computational efficiency, achieving strong perceptual restoration performance under a practical few-step inference regime.

Table 2: Quantitative results on the remaining restoration tasks. Quantitative comparison of PCFlow with baselines ELIR and PMRF. The best and second-best results are highlighted in **bold** and underlined, respectively. PCFlow consistently outperforms ELIR in FID, despite using only 21M parameters compared to ELIR (27M) and PMRF (176M), demonstrating favorable perceptual quality with efficient model architecture and fast inference.

Task	Model	Efficiency	Perceptual Quality	Distortion		
		#Params[M]↓	FID↓	PSNR↑	SSIM↑	LPIPS↓
Super Resolution	PMRF ($K = 25$)	176	**44.64**	**24.40**	**0.6708**	**0.2991**
	ELIR	27	49.25	23.57	0.6439	0.3299
	PCFlow (Ours)	21	45.50	23.38	0.6512	0.3328
Denoising	PMRF ($K = 25$)	176	**44.35**	**27.92**	**0.7782**	**0.2401**
	ELIR	27	47.70	26.67	0.7521	0.2619
	PCFlow (Ours)	21	45.42	26.26	0.7480	0.2800
Inpainting	PMRF ($K = 25$)	176	**42.88**	**26.19**	**0.7383**	**0.2626**
	ELIR	27	47.82	24.85	0.7105	0.2840
	PCFlow (Ours)	21	45.50	24.92	0.7195	0.2936
Colorization	PMRF ($K = 25$)	176	**42.61**	**23.47**	0.7122	**0.3463**
	ELIR	27	51.72	23.15	0.7113	0.3587
	PCFlow (Ours)	21	45.21	22.18	**0.7499**	0.3596

Other Restoration Tasks. We further provide quantitative results for the remaining restoration tasks, including super-resolution, denoising, inpainting, and colorization, in Tab. 2. PCFlow consistently improves over ELIR in FID across all four image restoration tasks. These results challenge the prevailing paradigm that prioritizes explicit posterior mean estimation followed by transport refinement. Instead, we show that directly learning the conditional transport from degraded observations, combined with perceptual supervision as a steering signal, is sufficient to achieve superior distortion-perception trade-offs under a few-step regime. Notably, the improvements are consistent across diverse degradation types, suggesting that the proposed formulation generalizes well beyond a specific restoration setting.

4.3 Qualitative Results

Blind Face Restoration. Fig. 1 presents qualitative comparisons on blind face restoration (BFR). Compared with ELIR, PCFlow consistently restores sharper facial structures, including hair, beard, and facial contours, while maintaining more natural textures and local contrast. These improvements lead to visually more faithful reconstructions without introducing noticeable artifacts. Compared with diffusion-based methods, PCFlow produces visually balanced reconstructions without the tendency toward over-sharpened textures observed in DiffBIR, while achieving visual quality competitive with DifFace and PMRF. Notably, these results are obtained using only five inference steps, demonstrating that the

Fig. 4: Qualitative comparison across image restoration tasks. From left to right: Input (LQ), ELIR, PMRF($K = 3$), PMRF($K = 25$), PCFlow (Ours), and ground truth image (HQ). PCFlow produces visually sharp and coherent reconstructions even with a small number of integration steps $K = 3$, establishing an effective training framework that provides a favorable distortion-perception trade-off.

proposed framework effectively achieves high perceptual restoration quality with highly efficient inference.

Other Restoration Tasks. Fig. 4 presents qualitative comparisons across four image restoration tasks. Compared to ELIR, PCFlow consistently generates visually more coherent and realistic reconstructions. In particular, ELIR tends to generate slightly over-smoothed facial structures, while PCFlow restores sharper facial boundaries and more consistent skin textures, leading to perceptually more faithful reconstructions. Note that PCFlow can occasionally generate subtle artifacts around high-frequency regions such as the eyes. Compared to PMRF with a larger number of integration steps $K = 25$, PCFlow attains strong perceptual quality while operating in a significantly more efficient few-step regime. These observations align with the quantitative trends in Tab. 2, where PCFlow improves perceptual metrics over ELIR and achieves a favorable distortion-perception trade-off, despite requiring substantially fewer transport steps and model parameters than PMRF($K = 25$).

4.4 Ablation Studies

We conduct several ablation studies and analyze the contribution of each component in our proposed PCFlow training objective.

Table 3: Validation of the Preheating Period. The combination of preheating period, gradient surgery, and linear warmup schedule demonstrates the best overall performance in both distortion and perceptual metrics on super-resolution task.

Setting	L_{LCPL}	Preheating	λ-scheduling	FID↓	PSNR↑	SSIM↑	LPIPS↓
L_{LCFM} only	–	–	–	46.10	23.35	0.6476	0.3373
w/o Gradient Alignment	epoch 0	×	linear warmup	46.40	23.32	0.6484	0.3363
w/ Gradient Alignment	epoch 0	×	linear warmup	46.26	23.31	0.6480	0.3364
PCFlow	epoch 250	✓	linear warmup	**45.50**	**23.38**	**0.6512**	**0.3328**

Table 4: Ablation on PCFlow components. Quantitative results for adding each PCFlow component on colorization task. The best performance under conditional flow formulation (B–E) is highlighted in **bold**, while the second-best are underlined.

	FID↓	PSNR↑	SSIM↑	LPIPS↓
Base (A)	56.27	23.11	0.7651	0.3675
+ Conditional (B)	47.21	21.88	0.7220	0.3845
+ Encoder Fine-Tuning (C)	45.97	<u>22.19</u>	0.7465	0.3654
+ Perceptual Objective (L_{LCPL}) (D)	<u>45.78</u>	**22.27**	**0.7530**	**0.3551**
+ Conflict-Free Gradient Alignment (E)	**45.21**	22.18	<u>0.7499</u>	<u>0.3596</u>

Preheating Period. Tab. 3 evaluates the contributions of the proposed preheating period. The preheating period corresponds to the first 250 epochs, during which only the reconstruction objective is optimized while perceptual supervision is disabled. Jointly training both distortion and perception objectives from the beginning yields inferior results compared to PCFlow, suggesting that the initial preheating stage is essential for establishing a stable coarse-to-fine transport trajectory before introducing perceptual supervision. In addition, applying conflict-free gradient alignment from epoch 0 consistently improves FID over naive joint optimization (46.40 → 46.26), indicating that mitigating optimization conflicts between reconstruction and perceptual objectives is beneficial. Combining both components yields the best performance across all metrics.

Individual Training Components. Tab. 4 shows quantitative results for progressively adding each of the PCFlow components. For the baseline unconditional model (config A), we set $\sigma_{\min}$ to 0.05 for super-resolution, 0.01 for denoising and colorization, and 0.025 for inpainting task. Training with conditional flow matching formulation (config B) improves FID by a substantial margin from 56.27 to 47.21. Fine-tuning the encoder along with the vector field (config C) further enhances model performance in both distortion and perceptual metrics. Adding the perceptual objective (config D) improves perceptual quality while maintaining reconstruction fidelity, effectively steering the model towards perceptually sharp data manifolds. Finally, incorporating the proposed conflict-free gradient alignment (config E) achieves the lowest FID of 45.21, pushing the model closer to the optimal distortion–perception tradeoff. We observe that similar trends consistently hold across the remaining image restoration tasks.

Table 5: Ablation on λ-scheduling. Comparison of different scheduling methods for λ_{LCPL}. Increasing the perceptual weight via linear warmup schedule consistently achieves the best model performance in FID across four image restoration tasks.

Task	λ-scheduling	FID↓	LPIPS↓
Super-Resolution	constant	46.18	**0.3300**
	linear	<u>45.74</u>	<u>0.3326</u>
	linear warmup	**45.50**	0.3328
Denoising	constant	46.12	**0.2790**
	linear	<u>45.48</u>	0.2814
	linear warmup	**45.42**	<u>0.2800</u>
Inpainting	constant	46.25	**0.2918**
	linear	<u>45.56</u>	0.2944
	linear warmup	**45.50**	<u>0.2936</u>
Colorization	constant	45.78	**0.3551**
	linear	<u>45.45</u>	0.3611
	linear warmup	**45.21**	<u>0.3596</u>

λ-scheduling. We further investigate the role of the perceptual weight λ_{LCPL} as a function of the diffusion timestep t. Tab. 5 shows quantitative comparison of different scheduling methods, including constant, linear, and linear warmup schedules. We observe that linear warmup schedule consistently achieves the best FID across all four image restoration tasks, demonstrating the benefit of gradually introducing perceptual supervision during transport. We hypothesize that early transport steps primarily require stable global alignment between degraded observations and the target distribution, where excessive perceptual supervision may introduce unnecessary optimization conflicts. In contrast, as the transport progresses closer to the true image manifold, perceptual supervision becomes more beneficial for recovering fine structures and visually realistic details.

4.5 Analysis on Gradient Conflict

We analyze the interaction between reconstruction objective $\mathcal{L}_{\text{LCFM}}$ and perceptual objective $\mathcal{L}_{\text{LCPL}}$ by comparing the results from standard joint optimization with our proposed gradient alignment method. As illustrated in Fig. 5, the cosine similarity heatmap reveals substantial regions of negative alignment (blue) under the standard update, indicating that the two objectives frequently produce conflicting gradients. Such conflicts can hinder stable optimization and slow down convergence. In contrast, after applying the proposed conflict-free update, the gradients exhibit significantly improved alignment, with fewer negatively correlated regions and more consistent positive interactions.

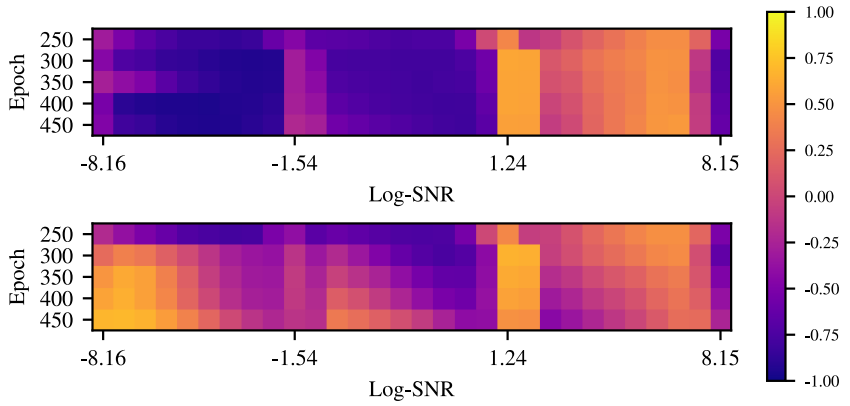

Fig. 5: Analysis of gradient interaction between flow matching and perceptual objectives. Cosine similarity maps between the LCFM and LCPL gradients as a function of training epochs and log-SNR, shown without (top) and with (bottom) the proposed conflict-free update. Our proposed method significantly mitigates gradient conflict, particularly in low-SNR regions, and promotes consistent optimization throughout the sampling process.

5 Conclusion

In this paper, we propose PCFlow, a unified latent transport framework for image restoration that jointly optimizes distortion and perceptual quality. By formulating the problem with latent consistency flow matching objective, our method directly learns the transport between degraded and clean images, hence enabling efficient few-step inference. To enhance perceptual realism, we additionally introduce a latent consistency perceptual objective that enforces semantic alignment along the transport trajectory. Furthermore, motivated by our analysis of the interaction between flow consistency and perceptual objectives, we propose a conflict-free gradient update method that enables the perceptual objective to steer optimization while removing conflicting structural gradient components. Extensive experiments demonstrate that PCFlow establishes a more favorable distortion–perception trade-off frontier compared to previous baselines across various image restoration tasks, while requiring only a few inference steps and maintaining an efficient model architecture.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00436857, 50%) and by IITP grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), 50%).

References

1. Adrai, T., Ohayon, G., Elad, M., Michaeli, T.: Deep optimal transport: A practical algorithm for photo-realistic image restoration. Advances in Neural Information Processing Systems **36**, 61777–61791 (2023)
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023)
3. Berrada, T., Astolfi, P., Hall, M., Havasi, M., Benchetrit, Y., Romero-Soriano, A., Alahari, K., Drozdzal, M., Verbeek, J.: Boosting latent diffusion with perceptual objectives. In: The Thirteenth International Conference on Learning Representations (2025)
4. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
5. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: The Eleventh International Conference on Learning Representations (2023)
6. Cohen, E., Achituve, I., Diamant, I., Netzer, A., Habi, H.V.: Efficient image restoration via latent consistency flow matching. In: 36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025. BMVA (2025)
7. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
10. Freirich, D., Michaeli, T., Meir, R.: A theory of the distortion-perception tradeoff in wasserstein space. Advances in Neural Information Processing Systems **34**, 25661–25672 (2021)
11. Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., Cheng, M.M.: Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In: European Conference on Computer Vision. pp. 126–143. Springer (2022)
12. Kang, M., Zhang, R., Barnes, C., Paris, S., Kwak, S., Park, J., Shechtman, E., Zhu, J.Y., Park, T.: Distilling diffusion models into conditional gans. In: European Conference on Computer Vision. pp. 428–447. Springer (2024)
13. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in neural information processing systems **35**, 23593–23606 (2022)
14. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
15. Lin, S., Yang, X.: Diffusion model with perceptual loss. arXiv preprint arXiv:2401.00110 (2023)
16. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European conference on computer vision. pp. 430–448. Springer (2024)

17. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
18. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
19. Ohayon, G., Michaeli, T., Elad, M.: Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. In: The Thirteenth International Conference on Learning Representations (2025)
20. Platen, P.V., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models (2022)
21. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4713–4726 (2022)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
23. Song, J., Vahdat, A., Mardani, M., Kautz, J.: Pseudoinverse-guided diffusion models for inverse problems. In: International Conference on Learning Representations (2023)
24. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9168–9178 (2021)
25. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
26. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. In: The Eleventh International Conference on Learning Representations (2023)
27. Yang, L., Zhang, Z., Zhang, Z., Liu, X., Liu, J., Xu, M., Meng, C., Ermon, S., Zhang, W., CUI, B.: Consistency flow matching: Defining straight flows with velocity consistency (2025)
28. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)
29. Yue, Z., Loy, C.C.: Diface: Blind face restoration with diffused error contraction. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9991–10004 (2024)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
31. Zhou, S., Chan, K., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. Advances in Neural Information Processing Systems **35**, 30599–30611 (2022)
32. Zhu, Y., Zhao, W., Li, A., Tang, Y., Zhou, J., Lu, J.: Flowie: Efficient image enhancement via rectified flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13–22 (2024)
33. Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., Van Gool, L.: Denoising diffusion models for plug-and-play image restoration. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 1219–1229 (2023)