Extreme Face Super-Resolution through Identity Fitting and Decoupling

Jiarui Yang[1,2], Hang Guo[2], Wen Huang[2], Shutao Xia[2], and Tao Dai[3⋆]

[1] College of Artificial Intelligence, Nankai University, Tianjin, China
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
1120230264@mail.nankai.edu.cn, cshguo@gmail.com, huang-w24@mails.tsinghua.edu.cn, xiast@sz.tsinghua.edu.cn, daitao.edu@gmail.com

Abstract. In recent years, face super-resolution (FSR) methods have achieved remarkable progress, generally maintaining high image fidelity and identity (ID) consistency under standard settings. However, under extreme degradation scenarios (e.g., scale ¿ > 8×), critical facial attributes and ID information are often severely lost in the input image, making it difficult for conventional models to reconstruct realistic and **ID-consistent** faces. Existing methods tend to generate hallucinated faces under such conditions, producing restored images that lack faithful ID constraints. To address this challenge, we propose a novel FSR method with Identity Decoupling and Fitting (IDFSR), designed to enhance ID restoration under large scaling factors while mitigating hallucination effects. Our approach involves three key designs: (1) **masking** the facial region in the low-resolution (LR) image to eliminate unreliable ID cues; (2) **warping** a reference image to align with the LR input, providing style guidance; and (3) leveraging **ID embeddings** extracted from ground-truth (GT) images for fine-grained ID modeling and personalized adaptation. We first pretrain a diffusion-based model to explicitly decouple style and ID by forcing it to reconstruct masked LR facial regions using both style and ID embeddings. Subsequently, we freeze most network parameters and perform lightweight fine-tuning of the ID embeddings using a small set of target ID images. This embedding encodes fine-grained facial attributes and precise ID information, significantly improving both ID consistency and perceptual quality. Extensive quantitative evaluations and visual comparisons demonstrate that the proposed IDFSR substantially outperforms existing approaches under extreme degradation, particularly in terms of ID consistency.

Keywords: Super-resolution · ID-consistent · Fitting · Decoupling

Fig. 1. Visualization of the pretraining and fine-tuning results. Under severe degradation, it is often impossible to reconstruct fine-grained facial details without reference images. General pretraining only preserves coarse ID consistency, whereas personalized fine-tuning achieves significantly stronger consistency in ID-related attributes.

1 Introduction

Face super-resolution (FSR) plays a critical role in downstream applications such as identity (ID) recognition and facial analysis [18,41]. In applications requiring strict ID preservation, FSR models must not only enhance image quality but also faithfully preserve ID consistency. Existing FSR methods can produce high-fidelity, ID-consistent face images under moderate degradation [4,42]. However, under extreme degradation conditions (e.g., scaling factors exceeding 8×), critical ID-related information and fine-grained facial attributes are often severely lost, rendering the reconstruction task highly ill-posed [1]. Consequently, the reconstruction problem becomes severely underdetermined, leading to numerous plausible solutions and making existing methods prone to generating hallucinated facial details that fail to preserve the true ID. As illustrated in Fig. 1, a representative method, CodeFormer [42], incorrectly reconstructs a blurred facial tattoo into an unrealistic dark patch.

Reference-based FSR methods [24,22] provide a promising solution by leveraging high-quality facial information from reference images to compensate for the missing details in low-resolution (LR) inputs. However, single-reference methods are highly sensitive to pose, illumination, and expression variations, and generally perform well only when the reference and LR images are well aligned [23,11]. Although multi-reference approaches [22,25] exploit richer ID information, most

⋆ Corresponding author.

existing methods fail to disentangle the ID information carried by the LR input and the reference images. Such disentanglement is particularly important because FSR is fundamentally a pixel-wise reconstruction task, where the LR image itself provides strong structural priors. During training, the model tends to over-rely on the remaining ID cues in the LR input while underutilizing the complementary ID information provided by the reference images. Consequently, the reconstructed faces may roughly match the target ID but often fail to recover fine-grained ID attributes and personalized facial characteristics.

To address these challenges, we propose a two-stage FSR framework termed **Identity Decoupling and Fitting (IDFSR)**. The proposed framework combines ID-disentangled pretraining with personalized ID embedding fine-tuning, enabling high-fidelity, ID-consistent, and semantically coherent face reconstruction under extreme degradation.

From an information-theoretic perspective, when image degradation becomes sufficiently severe (e.g., ultra-large downsampling factors combined with heavy noise corruption), the LR image no longer contains sufficient information to infer high-level facial semantics, such as gaze direction, subtle facial expressions, or other fine-grained ID-related attributes. Consequently, without reference images or external ID priors, existing methods struggle to recover faithful and controllable identities, frequently generating hallucinated facial details unrelated to the true ID. To overcome this fundamental limitation, our framework introduces three key components:

(1) Corrupted ID Masking. We first localize the facial region in the LR image and mask the detected region, which serves as the conditional input to the diffusion model. This strategy not only preserves consistent modeling of the background regions but also prevents the network from relying on degraded and unreliable ID cues during training. Consequently, the model is encouraged to exploit additional conditional information, including reference images and ID embeddings, thereby improving both ID controllability and generalization.

(2) Style-Condition Embedding. The facial region in the reference image is geometrically aligned with the corresponding region in the LR input. A style encoder then extracts a style embedding, which is injected into the diffusion backbone through feature fusion. This embedding provides coarse appearance and texture priors from the reference image, guiding the generation process to maintain structural plausibility while improving cross-image consistency.

(3) GT ID Embedding and Fitting. During pretraining, a pretrained ID encoder extracts ID embeddings from the ground-truth (GT) high-resolution (HR) images. These embeddings are incorporated into the backbone through an attention mechanism, enabling robust ID-aware representation learning. During the subsequent fine-tuning stage, all backbone parameters are frozen, and only the ID embedding is optimized using a few images of the target ID. This design preserves the generalization capability of the pretrained model while enabling accurate fitting of fine-grained ID attributes, resulting in highly personalized and ID-consistent face reconstruction.

As illustrated in Fig. 1, the pretrained model is capable of capturing coarse ID characteristics. After ID embedding fine-tuning, the model achieves substantially improved restoration of fine-grained facial details and significantly stronger consistency in ID-related attributes. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness of each proposed component and show that the proposed framework effectively suppresses ID hallucination. Quantitative and qualitative comparisons with state-of-the-art methods further confirm that IDFSR consistently achieves superior performance in both reconstruction quality and ID consistency.

2 Related Works

2.1 Single-Image Face Super-Resolution.

Unlike natural image super-resolution (SR) [40,6,13,12,7], face super-resolution (FSR) [31] poses unique challenges because it requires accurate restoration of fine-grained facial details, identity (ID) consistency, and facial structural priors. Early FSR methods predominantly rely on direct regression in the pixel space; however, empirical studies show that such approaches tend to produce over-smoothed results [17,4]. To improve perceptual quality, generative models are introduced to leverage prior distributions or explicit dictionaries for enhancing the naturalness of reconstructed images [14,37]. Nonetheless, owing to the inherent ambiguity of generative modeling, these methods still struggle to maintain pixel-level consistency and ID fidelity. To balance reconstruction fidelity and generation diversity, recent studies incorporate conditional generation mechanisms by introducing facial priors, such as landmarks, segmentation maps, and 3D facial structures, as auxiliary constraints to enforce geometric consistency [36,2]. Although these approaches improve geometric reconstruction, their ability to preserve ID remains limited. To further enhance ID fidelity, ID recognition losses are incorporated into FSR models [3]. However, under large degradation factors, ID cues in LR inputs become severely degraded, making the models prone to generating hallucinated identities.

2.2 Reference-Based Face Super-Resolution.

To alleviate the severe information loss caused by extreme degradation, reference-based FSR methods have been extensively investigated. GFRNet [23] proposes a dual-network architecture, where WarpNet aligns the reference image with the LR target through optical flow and RecNet performs face reconstruction. Landmark loss and total variation regularization are employed to improve alignment accuracy. However, large pose or expression discrepancies still result in inaccurate alignment. GWAINet [11] further introduces an ID loss to enhance ID consistency but remains highly dependent on precise alignment. ReFine [5] improves alignment by employing a fine-tuned landmark detector and further utilizes spatial minimality and cycle consistency losses to facilitate attribute

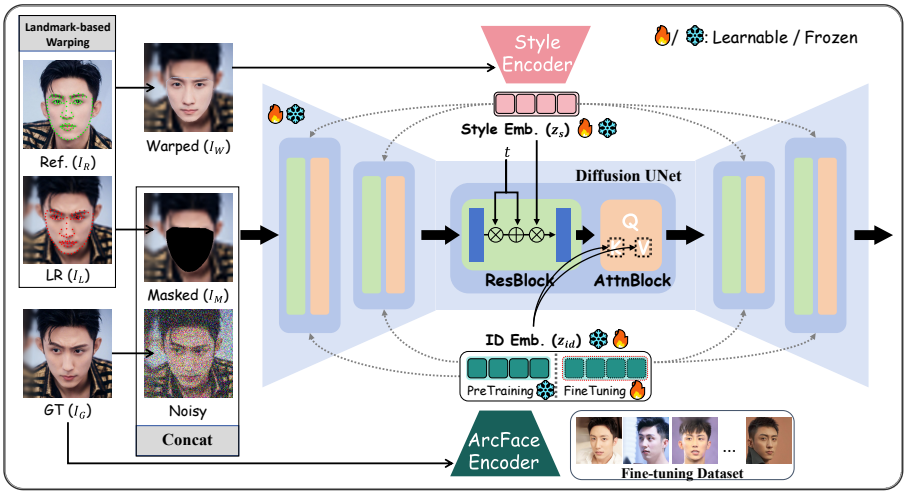

Fig. 2. A schematic diagram of single-step diffusion in IDFSR, including input preprocessing and the overall model architecture.

transfer. ASFFNet [22] selects suitable reference images based on landmark similarity and aligns them in both the spatial and illumination domains using Moving Least Squares (MLS) and Adaptive Instance Normalization (AdaIN), followed by multi-stage feature fusion. DMDNet [24] employs a dual-memory structure to separately store structural and ID features, enabling adaptive fusion across multiple reference images. MyStyle [25] and similar methods construct ID and attribute sets from multiple reference images to provide richer ID guidance. More recently, MGFR [30] utilizes a diffusion model to fuse textual conditions, reference images, and ID cues through dual-control adapters and a two-stage training strategy, achieving more controllable face restoration.

3 Methodology

3.1 Overview

Fig. 2 illustrates the overall architecture of IDFSR. Given a low-resolution (LR) image I_L, its corresponding ground-truth (GT) image I_G, and a same-ID reference image I_R, IDFSR consists of a parameterized diffusion network θ, a style encoder $\mathcal{E}_\text{s}$, and a pretrained ArcFace identity (ID) encoder $\mathcal{E}_\text{id}$ [9]. Prior to training, we perform two preprocessing steps:

- A finetuned facial landmark detector is used to mask the facial region of I_L, producing the masked image I_M.
- Landmark-based warping is applied to map the facial region of I_R onto I_L, generating the warped image I_W.

In the pretraining stage, we adopt a Denoising Diffusion Probabilistic Model (DDPM) [16] to model the image distribution, using the style embedding $z_{\mathrm{s}} = \mathcal{E}_{\mathrm{s}}(I_W)$ and ID embedding $z_{\mathrm{id}} = \mathcal{E}_{\mathrm{id}}(I_G)$ as conditions. The masked image I_M is concatenated with the noisy latent as input. The simplified diffusion objective at timestep t is defined as:

$$\mathcal{L}_{\mathrm{sim}} = \mathbb{E}_{x_0,\epsilon|z_{\mathrm{s}},z_{\mathrm{id}},I_M}\left[\|\epsilon - \epsilon_\theta([x_t, I_M], t, z_{\mathrm{s}}, z_{\mathrm{id}})\|_2^2\right]. \tag{1}$$

where $[\cdot]$ denotes concatenation, and x_t denotes the noisy version of the ground-truth image x_0 at timestep t.

In the finetuning stage, we freeze the diffusion network θ and the style encoder $\mathcal{E}_{\mathrm{s}}$, and replace z_{id} with a learnable embedding vector. This vector is optimized using a small number of same-ID samples under the same training objective, enabling personalized and accurate ID control. Notably, IDFSR maintains strong performance even when landmark detection is imperfect or reference images are unavailable, demonstrating robustness under real-world conditions.

3.2 Landmark Detection and Warping

We use RetinaFace [8] to extract facial landmarks from high-quality (HQ) images and finetune the detector to better handle LR inputs. The facial region of the reference image I_R is then aligned to I_L via affine transformation based on the detected landmarks, producing the warped image I_W. This pseudo-alignment introduces coarse spatial correspondence between reference and LR images. We provide pseudocode of the warping process in Appendix E.1.

Fig. 3. Warping analysis. From top to bottom: reference images, warped images, super-resolved (SR) images, and ground-truth (GT) images.

However, due to pose variations and landmark estimation errors, such warping is often imperfect and may introduce local distortions or mismatches. Interestingly, we find that this imperfection does not hinder performance; instead, it acts as a form of natural data augmentation. The diversity in warped faces encourages the model to avoid over-reliance on precise spatial alignment and instead focus on global ID semantics. This drives the diffusion model to rely more on the ID embedding z_{id} for reconstructing missing facial regions, rather than directly copying textures from the warped reference. As a result, our model becomes robust to variations in reference quality and

alignment. As shown in Fig. 3, IDFSR still generates plausible and ID-consistent faces even under severe misalignment.

3.3 Pretraining and Finetuning

Our diffusion backbone follows the U-Net architecture used in Guided Diffusion [10], consisting of residual blocks, attention modules, and upsampling/downsampling layers. Our conditional injection design enables effective ID decoupling and fitting.

The style encoder adopts the encoder part of the diffusion U-Net. Inspired by DiffAE [26], we introduce adaptive group normalization (AdaGN [35]) to inject style conditions. For a feature map $h \in \mathbb{R}^{c\times h\times w}$, AdaGN is formulated as:

$$\text{AdaGN}(h, t, z_s) = z_f \cdot (t_s \cdot \text{GroupNorm}(h) + t_b)\,, \tag{2}$$

where $z_f \in \mathbb{R}^c$ is obtained via an affine transformation applied to the style embedding z_s:

$$z_f = \text{MLP}_\text{style}(z_\text{s}), \tag{3}$$

and $(t_s, t_b) \in \mathbb{R}^{2\times c}$ are obtained by applying an MLP to the sinusoidal time embedding $\psi(t)$:

$$(t_s, t_b) = \text{MLP}_\text{time}(\psi(t)). \tag{4}$$

In the pretraining stage, we leverage a pretrained ArcFace encoder [9] to extract ID embeddings from GT images. These embeddings are injected into attention modules via cross-attention to provide fine-grained ID priors. The diffusion network and style encoder are trained from scratch, initialized with DiffAE weights pretrained on the FFHQ dataset [19]. In the finetuning stage, we freeze both the U-Net and style encoder, and optimize only a learnable ID embedding initialized from a reference image. Note that the ArcFace encoder is not required during finetuning. Both stages share the same hyperparameters: a maximum diffusion step of $T = 1000$, a linear β schedule, and DDIM sampling with 20 steps during inference.

4 Experiments

4.1 Experimental Setting

Our experiments include both general and personalized evaluations. In the pretraining stage, identity (ID) embeddings from reference images are used to enable general face super-resolution (FSR), while in the fine-tuning stage, we focus on personalized FSR for a specific ID. The pretraining strategy is based on the DiffAE model [26], which is trained for 2 days on four RTX 3090 GPUs, followed by 20 minutes of fine-tuning on a single 3090 GPU using images of the target ID. Detailed implementation details are provided in Appendix E.2.

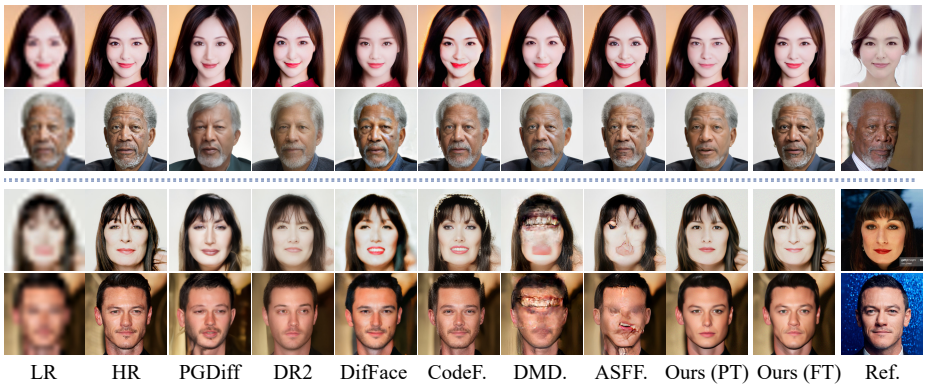

Fig. 4. Visualization results of different methods under upsampling scales of 8 and 16, separated by double dashed lines. For reference-based methods—DMDNet, ASFFNet, and our IDFSR—a single ID is randomly selected as the reference image. FT and PT refer to finetuning and pretraining methods, respectively. Please zoom in for a better viewing.

Datasets. We uniformly sample 10% of identities from CelebRef-HQ, while the remaining 90% is used for pretraining. Each sampled ID is split equally into a fine-tuning set and a test set. We remove identities with too few images and additionally select representative identities to better evaluate ID fitting. As a result, the training set contains approximately 900 identities and 9,000 images, while the test set contains 56 identities and 586 images. In addition, CASIA-WebFace and the CelebText video dataset are used for generalization evaluation after quality filtering. See Appendix D for further details.

Metrics and Baselines. We adopt a comprehensive set of evaluation metrics, including PSNR, SSIM, LPIPS [39], FID [15], CLIPIQA [32], MUSIQ [20], and ID similarity (IDS) measured using DeepFace [29]. We compare our method against state-of-the-art (SOTA) reference-free and reference-based FSR methods, including CodeFormer [42], DR2 [34], DifFace [38], PGDiff [37], DMDNet [24], and ASFFNet [22].

4.2 Qualitative Comparisons

As shown in Fig. 4, we present qualitative comparisons with state-of-the-art (SOTA) FSR methods. Under moderate degradation (8×), non-reference-based methods can achieve visually plausible reconstructions, but still exhibit noticeable inconsistencies in facial ID and fine-grained detail recovery. Unconditional generative methods, such as PGDiff and DR2, tend to produce hallucinated facial features, i.e., unrealistic or incorrect ID-related details. In contrast, conditional generative methods with discriminative constraints, such as DifFace, improve pixel-level consistency but still struggle to preserve fine-grained ID attributes.

This limitation mainly stems from the tendency of these methods to learn averaged representations during training, which weakens their ability to model

Methods	×8 PSNR↑	SSIM↑	LPIPS↓	FID↓	CLIPIQA↑	MUSIQ↑	IDS↓
CodeFormer	24.42	0.7147	0.1489	31.88	0.6873	68.98	0.3865
PGDiff	22.73	0.6892	0.2223	46.05	0.5390	56.10	0.6080
DR2	23.83	0.7067	0.1998	44.46	0.6307	63.35	0.5459
DifFace	24.41	0.7255	0.1694	32.94	0.6132	60.54	0.3815
StableSR	23.95	0.7174	0.1523	28.34	0.6006	67.58	0.3995
DPI	24.04	0.7564	0.1271	27.94	0.6850	70.32	0.3342
ASFFNet	21.82	0.6624	0.1764	26.25	0.6158	65.81	0.3169
DMDNet	22.36	0.6984	0.1733	29.76	0.6570	68.28	0.4169
Ours (PT)	24.29	0.6996	0.1554	27.89	0.6798	69.30	0.3173
Ours (FT)	28.85	0.7604	0.1031	26.69	0.7092	73.73	0.2242
Methods	×16 PSNR↑	SSIM↑	LPIPS↓	FID↓	CLIPIQA↑	MUSIQ↑	IDS↓
CodeFormer	20.72	0.5947	0.2379	43.89	0.6834	67.85	0.6673
PGDiff	20.47	0.6202	0.2727	55.70	0.5008	52.19	0.7355
DR2	22.35	0.6356	0.2481	50.84	0.6102	62.60	0.7082
DifFace	22.23	0.6737	0.2195	38.65	0.5957	57.84	0.5948
StableSR	22.65	0.6680	0.2083	37.48	0.6818	66.82	0.6644
DPI	23.48	0.6849	0.1997	38.35	0.6886	68.26	0.5532
ASFFNet	18.23	0.5732	0.2457	40.64	0.5827	63.72	0.7955
DMDNet	17.38	0.5590	0.2497	46.34	0.6051	64.05	0.8291
Ours (PT)	23.32	0.6863	0.2062	38.58	0.6723	68.56	0.5227
Ours (FT)	24.09	0.7158	0.1845	35.50	0.6992	71.83	0.3625

Table 1. Quantitative comparison on the CelebRef-HQ dataset with upsampling scales of 8× and 16×. FT and PT refer to finetuning and pretraining methods, respectively. **Red** and blue indicate the best and the second best.

ID-specific details. Our pretrained model shows similar behavior to some extent; however, by incorporating reference images, it outperforms non-reference-based methods in mitigating hallucinations and improving ID consistency. Meanwhile, reference-based SOTA methods such as ASFFNet achieve a better trade-off between ID consistency and attribute preservation under 8× degradation. However, their performance degrades significantly under extreme degradation (e.g., 16×), as they heavily rely on accurate alignment between the low-resolution (LR) image and the reference image. In contrast, our proposed personalized fine-tuning strategy substantially enhances the preservation of ID features and facial attributes, demonstrating stronger robustness under extreme low-quality conditions. Under the 16× setting, IDFSR achieves clear advantages in both ID consistency and visual fidelity.

4.3 Quantitative Comparisons

As shown in Table 1, under moderate degradation, the pretrained model does not achieve optimal performance on pixel-level metrics such as PSNR, SSIM, and LPIPS. However, under severe degradation (e.g., 16×), the incorporation of reference-based conditioning significantly improves reconstruction quality, en-

Methods	IDV ↑	ACC_{Gen} ↑	ACC_{Emo} ↑	ACC_{Ra} ↑	$Diff_{Age}$ ↓
CodeFormer	50.3%	95.2%	64.3%	64.8%	4.6 ± 4.3
PGDiff	32.9%	88.7%	51.0%	68.8%	5.2 ± 4.6
DR2	38.9%	91.9%	57.6%	71.5%	5.2 ± 5.1
DiffFace	73.0%	95.4%	67.4%	77.3%	4.6 ± 4.4
ASFFNet	21.4%	72.1%	39.6%	45.2%	7.5 ± 7.2
DMDNet	18.7%	66.3%	36.8%	42.5%	7.6 ± 7.3
Ours (FT)	**89.6%**	**98.7%**	**66.2%**	**96.6%**	**3.3 ± 1.7**

Table 2. The quantitative comparison of face ID verification (IDV) and attribute consistency, including gender accuracy (ACC_{Gen}), emotion accuracy (ACC_{Emo}), race accuracy (ACC_{Ra}), and age difference ($Diff_{Age}$).

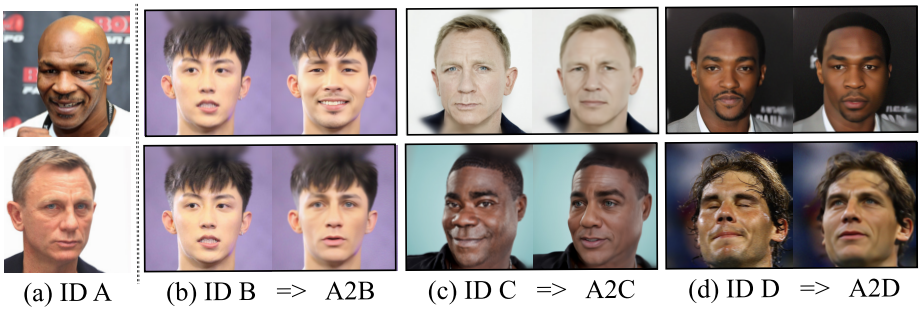

Fig. 5. Visualization of cross-ID attribute transfer. We first fit the embedding on ID A, then perform FSR using the LR and style ID from another ID.

abling the pretrained model to achieve competitive or state-of-the-art performance on pixel-level consistency metrics.

Further analysis shows that perceptual quality metrics, including FID, CLIP-IQA, and MUSIQ, also demonstrate strong performance, validating the effectiveness of the proposed pretraining paradigm. As shown in Table 1, our personalized fine-tuning strategy achieves state-of-the-art performance across pixel-level consistency, perceptual quality, and ID consistency. Specifically, it improves pixel-level metrics by approximately 15%, significantly enhances ID similarity, and consistently outperforms existing methods while maintaining high visual fidelity.

We also observe a discrepancy between embeddings used during pretraining and testing. Specifically, ground-truth (GT) ID embeddings are used during pretraining, whereas reference-based ID embeddings are used during inference. This mismatch may introduce bias from reference-specific fine-grained details, potentially leading to misalignment in attribute representation. We further analyze this effect in the following section.

4.4 Facial Attribute Consistency Comparison

Due to the introduction of the masking mechanism, part of the information in the LR image is inevitably discarded during reconstruction. This raises an important question: under severe degradation, can our method still maintain satisfactory semantic and ID consistency? To answer this, we conduct extensive evaluations on fine-grained attribute consistency.

Specifically, we employ the DeepFace analysis toolkit [18] to perform face verification and attribute prediction between ground-truth (GT) images and their corresponding SR results. Face verification is conducted based on ID similarity, where a default threshold is used to determine whether the GT and SR images belong to the same ID. Attribute prediction is evaluated across four dimensions: age, gender, emotion, and race.

Under the 16× upscaling setting, we quantitatively compare different methods in terms of recognition accuracy, as summarized in Table 2. IDFSR demonstrates a clear advantage in ID verification accuracy. In contrast, many existing methods achieve accuracy below 50%, indicating severe ID distortion under extreme degradation and highlighting the robustness of our approach in such challenging scenarios.

Although most methods exhibit relatively stable performance in gender consistency, emotion and race consistency remain more challenging. Our method achieves a race prediction accuracy of 96%, which can be attributed to the incorporation of personalized ID embeddings that effectively preserve high-level ID-related attributes. However, emotion prediction accuracy is comparatively lower. This is mainly because our framework does not rely on strong priors from the LR image space; instead, it reconstructs facial details based on reference images and customized ID embeddings. While this strategy enhances ID fidelity, it may partially compromise the accurate recovery of expression-related cues. We further analyze this limitation in the following section.

4.5 Progressively Degraded Reference Images

Following the real-world degradation process defined in Real-ESRGAN [33], we investigate the robustness of IDFSR under progressively degraded reference images. Leveraging the controllable degradation process, we gradually apply stronger degradation to the reference image until its ID cues are completely removed, corresponding to an extreme "reference-free" setting.

$$I_L = \left\{ \left[(I_H * g_{s,\sigma}) \downarrow_r + \eta_\delta \right]_{\mathrm{JPEG}(q)} \right\} \uparrow_r, \tag{5}$$

where $g_{s,\sigma}$ denotes a Gaussian smoothing kernel with kernel size s and standard deviation σ, η_δ represents additive Gaussian noise with variance δ, JPEG(q) denotes JPEG compression with quality factor q, and r is the scale factor used for downsampling and subsequent upsampling.

As shown in Fig. 6, we observe that under mild degradation, model performance remains largely unaffected, and consistency with the reference image is

Fig. 6. Robustness analysis under progressively degraded reference images.

well preserved. As degradation becomes more severe and the reference image loses meaningful ID information, the model gradually degenerates into an ID-aware super-resolution model that relies primarily on ID priors. In this extreme case, the model still preserves certain fine-grained ID-related attributes—such as eye color or nose shape inferred from the learned ID embedding—while overall ID consistency inevitably degrades.

5 Ablation Study

5.1 ID Decoupling and Generalization Verification

We design a cross-ID experiment to evaluate the fine-grained modeling capability of the learned ID representation. Specifically, we first fit an ID embedding on images of a source identity (ID A), and then apply it to restore images of a different identity (ID B). During testing, the embedding of ID A is used together with LR inputs and reference images from other identities. As shown in Fig. 5, despite the identity mismatch, the fitted ID A embedding can still effectively modify target facial attributes while preserving background and overall style. This indicates that the learned ID embedding is largely decoupled from both the LR input and the style encoding, suggesting that it is not dependent on pixel-level supervision or explicit alignment via warping. Moreover, this result further demonstrates the strong generalization ability of the learned ID representation.

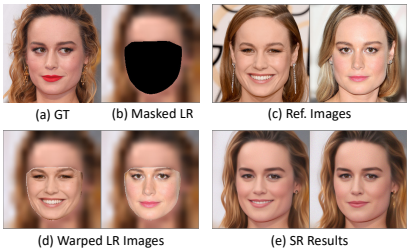

Fig. 7. Latent editing effects of style embedding. Different reference styles mainly affect facial expression, while ID and background remain highly consistent.

5.2 Latent Editing Effects of Style Embedding

The style embedding z_s serves as the only explicit pathway in our framework for capturing spatial facial information, while the warping operation is not always perfectly aligned. In the methodology section, we discussed the impact of imperfect alignment; here, we further analyze the potential editing effect of the style embedding in face super-resolution.

As shown in Fig. 7, we select three images of the same ID with variations in lighting, makeup, and facial expressions. Two images with large motion differences are used as references under the 16× setting. It can be observed that the generated results exhibit clear correspondence with the reference expressions. Despite minor pixel-level misalignment, our method maintains strong ID consistency and visual fidelity. In contrast, existing approaches tend to prioritize pixel-level alignment, which limits their flexibility in modeling ID variations. Notably, aside from facial expression changes, background and fine details remain largely unchanged, further validating the effectiveness of our disentangled design.

5.3 Impact of Reference Image Quantity

To investigate the effect of the number of reference images on fine-tuning performance, we conduct experiments on the CelebHQ-Ref test set, where each ID contains more than 10 images. Specifically, we gradually vary the number of reference images from 2 to 10 and evaluate performance using three normalized metrics: pixel-level consistency, ID consistency, and attribute consistency. As shown in Fig. 8, we report the mean values and confidence intervals across 10 identities.

Overall, increasing the number of reference images improves performance across all metrics. In the early stage, adding a small number of references yields limited gains, likely because the pretrained model already possesses strong single-image generalization capability. Additional references provide only marginal benefits due to variations in quality, pose, and redundancy in ID information. Notably, performance saturates when the number of reference images reaches around five, where both accuracy and stability achieve a good trade-off.

5.4 Component Analysis

Starting from a vanilla diffusion-based super-resolution model, we progressively introduce key components to analyze their contributions, as shown in Fig. 9. The basic formulation concatenates the LR image with noise at each timestep (case1), as in SR3 [28] and SRDiff [21]. This design enforces strong structural constraints, improving consistency but reducing flexibility under severe degradation.

In case2, we mask the LR image before concatenation, reformulating the task as an inpainting problem. This introduces higher generative freedom while maintaining structural priors. Compared with feature-based conditioning methods such as DiffAE and Stable Diffusion [27] (case3), case2 achieves better pixel-level consistency but slightly lower perceptual quality. As shown in Fig. 9, case1

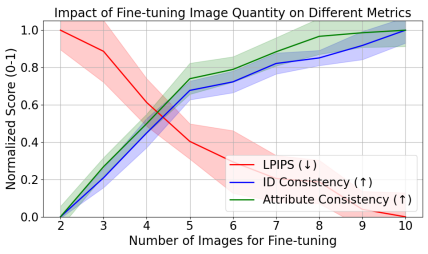

Fig. 8. Effect of reference image quantity. Performance improves with more references and saturates around five images, where stability and accuracy are balanced.

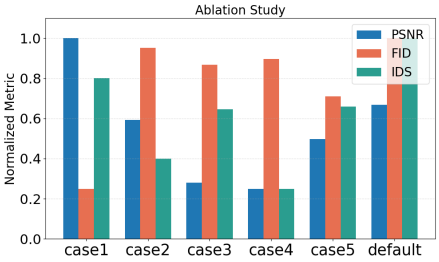

Fig. 9. Ablation on model components. Metrics are normalized to visualize the contribution of each component across different evaluation criteria.

Table 3. Complexity analysis. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Method	PGDiff	DR2	DifFace	StableSR	DPI	IDFSR
Params (M)	159.7	179.31	175.42	918.93	145.53	160.7
FLOPs (G)	185.95	918.85	272.67	2000+	136.57	168.68
NFEs	100	100	100	50	20	20
Time (ms)	71	89	70	362	50	54
Total time (s)	7	9	7	18	1	1

performs best in consistency but worst in perceptual quality, while case2 shows the opposite trend. Case3 achieves a balanced trade-off across metrics.

When removing the ID embedding (case4), the model suffers from degraded ID consistency due to imperfect alignment introduced by warping. Removing the style embedding (case5) leads to a loss of realistic facial attributes, indicating that ID and style embeddings play complementary roles. The full model integrates all components, where masking provides a strong structural prior, while ID and style embeddings enable disentangled and controllable generation.

5.5 Complexity Analysis

As shown in Table 3, IDFSR achieves a favorable trade-off between efficiency and performance. Although it maintains a moderate number of parameters and FLOPs, it requires only 20 sampling steps, resulting in a total inference time of approximately 1 second. In comparison, PGDiff, DR2, and DifFace require 100 steps (7–9 seconds), while StableSR requires up to 18 seconds. Compared with DPI, IDFSR slightly increases per-step computation (54 ms vs. 50 ms) due to the additional style encoder and ID embedding module, but achieves significantly improved ID preservation and reconstruction quality.

6 Conclusion

In this paper, we propose IDFSR, a two-stage diffusion-based face super-resolution framework. In the pretraining stage, we introduce a disentangled conditioning design to separate ID and style information. In the fine-tuning stage, we enable personalized ID fitting through a learnable ID embedding. Extensive experiments demonstrate that IDFSR achieves strong generalization and personalization capabilities, consistently improving both ID and attribute consistency. Ablation studies further validate the effectiveness of each component and the proposed disentanglement design.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309,62571298).

References

1. Alekseev, A.K., Navon, I.M.: The analysis of an ill-posed problem using multi-scale resolution and second-order adjoint techniques. Computer Methods in Applied Mechanics and Engineering **190**(15-17), 1937–1953 (2001)
2. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 109–117 (2018)
3. Chen, J., Chen, J., Wang, Z., Liang, C., Lin, C.W.: Identity-aware face super-resolution for low-resolution face recognition. IEEE Signal Processing Letters **27**, 645–649 (2020)
4. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2492–2501 (2018)
5. Chong, M.J., Xu, D., Zhang, Y., Wang, Z., Forsyth, D., Krishnan, G., Wu, Y., Wang, J.: Copy or not? reference-based face image restoration with fine details. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9660–9669. IEEE (2025)
6. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019)
7. Dai, T., Wang, J., Guo, H., Li, J., Wang, J., Zhu, Z.: Freqformer: Frequency-aware transformer for lightweight image super-resolution. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 731–739 (2024)
8. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5203–5212 (2020)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)

10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
11. Dogan, B., Gu, S., Timofte, R.: Exemplar guided face image super-resolution without facial landmarks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
12. Guo, H., Guo, Y., Zha, Y., Zhang, Y., Li, W., Dai, T., Xia, S.T., Li, Y.: Mambairv2: Attentive state space restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28124–28133 (2025)
13. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. In: European conference on computer vision. pp. 222–241. Springer (2024)
14. He, J., Shi, W., Chen, K., Fu, L., Dong, C.: Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1889–1898 (2022)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: Proceedings of the IEEE international conference on computer vision. pp. 1689–1697 (2017)
18. Jiang, J., Wang, C., Liu, X., Ma, J.: Deep learning-based face super-resolution: A survey. ACM Computing Surveys (CSUR) **55**(1), 1–36 (2021)
19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
20. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
21. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
22. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2706–2715 (2020)
23. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European conference on computer vision (ECCV). pp. 272–289 (2018)
24. Li, X., Zhang, S., Zhou, S., Zhang, L., Zuo, W.: Learning dual memory dictionaries for blind face restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
25. Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mosseri, I., Pritch, Y., Cohen-Or, D.: Mystyle: A personalized generative prior. ACM Transactions on Graphics (TOG) **41**(6), 1–10 (2022)
26. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10619–10629 (2022)

27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. 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)
29. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
30. Tao, K., Gu, J., Zhang, Y., Wang, X., Cheng, N.: Overcoming false illusions in real-world face restoration with multi-modal guided diffusion model. arXiv preprint arXiv:2410.04161 (2024)
31. Tomar, A.S., Arya, K., Rajput, S.S., Rodriguez, C.R.: Comprehensive survey of face super-resolution techniques. Digital Image Enhancement and Reconstruction pp. 213–233 (2023)
32. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
33. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)
34. Wang, Z., Zhang, Z., Zhang, X., Zheng, H., Zhou, M., Zhang, Y., Wang, Y.: Dr2: Diffusion-based robust degradation remover for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1704–1713 (2023)
35. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
36. Yang, J., Dai, T., Zhu, Y., Li, N., Li, J., Xia, S.T.: Diffusion prior interpolation for flexibility real-world face super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9211–9219 (2025)
37. Yang, P., Zhou, S., Tao, Q., Loy, C.C.: Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. Advances in Neural Information Processing Systems **36**, 32194–32214 (2023)
38. Yue, Z., Loy, C.C.: Diffface: Blind face restoration with diffused error contraction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
39. 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)
40. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
41. Zhang, Y., Wu, Y., Chen, L.: Msfsr: A multi-stage face super-resolution with accurate facial representation via enhanced facial boundaries. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 504–505 (2020)
42. 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)