HonestFace: Towards Honest Face Restoration with One-Step Diffusion Model

Jingkai Wang[1], **Wu Miao**[1], **Jue Gong**[1], **Zheng Chen**[1],
Xing Liu[2], **Hong Gu**[2], **Yutong Liu**[1,*] **Yulun Zhang**[1,*]
[1]Shanghai Jiao Tong University, [2]vivo Mobile Communication Co., Ltd

Abstract

Face restoration has achieved remarkable advancements through the years of development. However, ensuring that restored facial images exhibit high fidelity, preserve authentic features, and avoid introducing artifacts or biases remains a significant challenge. This highlights the need for models that are more "honest" in their reconstruction from low-quality inputs, accurately reflecting original characteristics. In this work, we propose HonestFace, a novel approach designed to restore faces with a strong emphasis on such honesty, particularly concerning identity consistency and texture realism. To achieve this, HonestFace incorporates several key components. First, we propose an identity embedder to effectively capture and preserve crucial identity features from both the low-quality input and multiple reference faces. Second, a masked face alignment method is presented to enhance fine-grained details and textural authenticity, thereby preventing the generation of patterned or overly synthetic textures and improving overall clarity. Furthermore, we present a new landmark-based evaluation metric. Based on affine transformation principles, this metric improves the accuracy compared to conventional L2 distance calculations for facial feature alignment. Leveraging these contributions within a one-step diffusion model framework, HonestFace delivers exceptional restoration results in terms of facial fidelity and realism. Extensive experiments demonstrate that our approach surpasses existing state-of-the-art methods, achieving superior performance in both visual quality and quantitative assessments. The code and pre-trained models will be made publicly available at https://github.com/jkwang28/HonestFace.

1 Introduction

Face restoration is an ill-posed problem that aims to recover a high-quality (HQ) face image from a given low-quality (LQ) input. In practice, LQ face images often suffer from severe degradations (*e.g.*, blur, noise, compression artifacts) that result in significant loss of detail. This loss of information makes accurate restoration extremely challenging, particularly when recovering fine details and preserving the subject's identity. Over the years, numerous approaches have been explored to improve face restoration performance, including models based on Transformers [78, 62, 66, 52], GANs [3, 58, 70, 2], and diffusion models [61, 37, 5, 40, 47, 63, 30, 73, 57]. These methods have significantly improved the quality of restored faces, marking notable milestones in low-level vision.

However, a key requirement for a "high-quality" restored face is a natural appearance, *i.e.*, it should look like a genuine photograph with authentic textures. In practice, many generative restoration models produce results that, while sharp, appear unnatural. A common issue is over-smoothing: skin may look overly airbrushed, lacking pores and fine wrinkles, and hair often appears too uniform, missing the irregular details of real hair. For example, the diffusion-based model DiffBIR [30], despite its powerful generative prior, often yields overly smooth, even portrait-like faces. Similarly,

*Corresponding authors: Yutong Liu <isabelleliu@sjtu.edu.cn> and Yulun Zhang <yulun100@gmail.com>.

recent transformer-based models like CodeFormer [78] and DAEFR [52] sometimes produce artifacts in certain regions. In particular, their restored hair textures can appear artificial or repetitive, lacking the natural randomness of real hair strands. Another common issue is color shift, where the restored image color deviates significantly from the input. Some methods [31, 63] attempt to address this using AdaIN or wavelet-based techniques, but the problem remains largely unresolved. These limitations highlight the ongoing challenge of generating truly natural-looking results without sacrificing detail.

Moreover, a practical face restoration model intended for real-world applications should handle a wide range of degradation levels. One of the toughest challenges with severely degraded inputs is keeping identity fidelity. When the identity information in the LQ image is largely lost, relying solely on the degraded input often causes the restored face to deviate from the person's true appearance. To tackle this issue, reference-based face restoration techniques have been proposed [25, 11, 28, 27, 51, 53]. These methods incorporate an additional high-quality reference image of the same person to guide the restoration process, providing essential identity cues that are missing in the LQ input. Furthermore, having multiple reference images could provide even more information. Some approaches attempt to leverage multiple references by selecting the single most reliable one [28] or simply averaging information from all of them. However, such strategies do not fully exploit all the available input data. In summary, recent efforts in face restoration have primarily focused on two key objectives: perceptual naturalness and identity fidelity. And the "high-quality" restored face needs to both maintain identity consistency while producing realistic, natural textures.

In this paper, we propose a novel face restoration model, *HonestFace*, designed to address the challenges outlined above. **Firstly**, we address the critical issue of identity consistency, where existing methods often fail by altering key facial features, such as eye color and skin texture. To overcome this, we present the identity embedder (IDE), which comprises two parallel components: a facial feature extractor that captures fine details like wrinkles and eye color, and a face identity encoder that preserves the overall identity. This dual design ensures comprehensive identity retention while emphasizing subtle, person-specific textures. **Secondly**, we propose the masked face alignment (MFA) method to enhance the naturalness. Recognizing that details like wrinkles and skin structure are crucial for realism, MFA extracts a heatmap of facial landmarks and applies it as a mask on the image's alpha channel. This approach focuses restoration efforts on areas most relevant to human perception, resulting in outputs with authentic textures. **Thirdly**, we introduce a new affine transformation-based landmark distance metric for evaluation. Unlike traditional L2-based distances, this metric better reflects the geometric properties of facial landmarks, offering a more precise assessment of identity consistency. **Finally**, with our proposed techniques, we establish HonestFace, a one-step diffusion model that accepts multiple reference images. In particular, our method avoids the over-smoothed, "plastic" appearance and the repetitive artifacts commonly seen in prior methods. The faces restored by HonestFace not only remain faithful to the individual's identity but also exhibit realistic skin texture and hair detail. Our methods achieve state-of-the-art (SOTA) results in both visual quality and quantitative metrics. Our contributions are summarized as follows:

- We introduce the identity embedded (IDE), consisting of a facial feature extractor and a face identity encoder, to ensure that key facial features are preserved.
- We present the masked face alignment (MFA) method to enhance the naturalness of the restored images. By applying a heatmap-based mask on the alpha channel, MFA ensures that restoration efforts are concentrated on perceptually important areas.
- We develop a novel affine transformation-based landmark distance metric that provides a more accurate evaluation of identity consistency compared to traditional L2-based metrics.
- We propose HonestFace, a one-step diffusion model that uses multiple reference images. Our method prevents common issues such as over-smoothing and repetitive artifacts, delivering "honest" face restoration that preserves both identity and realistic textures.

2 Related Work

2.1 Reference-based Face Restoration

Faces are the most detail-sensitive part of the human body and have been widely studied in image restoration [72, 6, 19, 45, 36, 14, 54, 67, 58]. However, traditional methods often struggle with challenging tasks. To address the loss of identity information in low-quality inputs, reference-based methods utilize high-quality images of the same individual [26, 11]. For instance, ASFFNet [28] focuses on multi-exemplar tasks by selecting the optimal reference and adaptively fusing features.

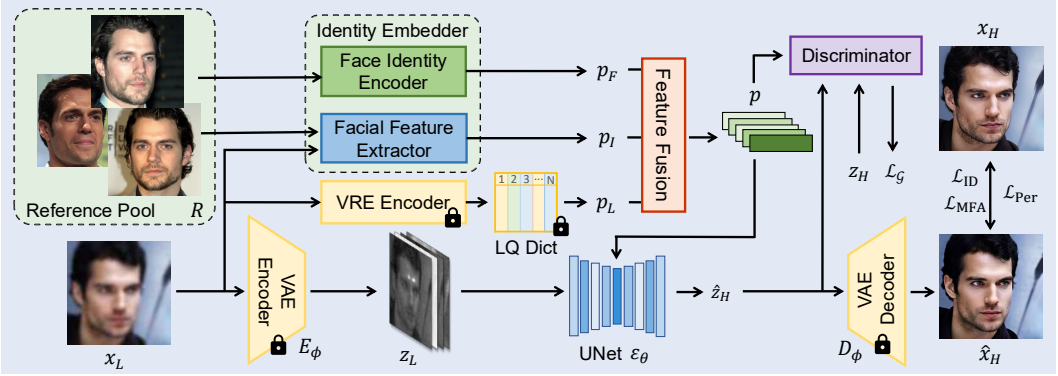

Figure 1: Overall training pipeline of HonestFace. First, the LQ input x_L is encoded into z_L by the VAE encoder E_ϕ. Meanwhile, x_L and HQ references $R = \{r_i\}$ pass through IDE and VRE, then fused to form the prompt embedding p. Next, the UNet predicts ε_θ to estimate $\hat{z}_H$. Finally, the VAE decoder D_ϕ reconstructs the output $\hat{x}_H$. Generator and discriminator are trained alternately.

DMDNet [29] introduces dual memory dictionaries to store general facial priors and identity-specific features. Nitzan et al. [38] proposed MyStyle, which learns a subject's unique identity by fine-tuning a pre-trained face generator, creating a personalized prior. Zeng et al. [74] improved MyStyle with a new optimization system to better organize the latent space. PFStorer [53] used reference faces to fine-tune and get a new model, generating identity-specific restoration. These methods enable high-fidelity reconstruction and even allow attribute editing of specific facial features.

2.2 Diffusion Models

Latent diffusion models [42] and related works [41, 4, 21, 65] have driven progress in high-resolution image synthesis. At the same time, pre-trained diffusion models have been widely applied in low-level vision tasks [56, 64, 71, 49, 12, 24]. Additionally, many methods use diffusion models for face image restoration. Yang et al. [68] proposed partial guidance for diffusion models, focusing on modeling high-quality image properties instead of degradation. Ding et al. [10] fine-tuned a pre-trained diffusion model with high-quality image sets to constrain the generative space. FaceMe [31] used an identity encoder to extract identity features as guidance. Tao et al. [51] combined attribute text prompts, high-quality reference images, and identity information using a dual-control adapter and two-stage training, showing the effectiveness of multi-modal inputs in reference-based face restoration. Zhang et al. [75] applied a one-step diffusion model with an attention-sharing mechanism to achieve fast, personalized restoration without per-identity fine-tuning. Wang et al. [57] introduced a one-step diffusion model with a visual representation embedder to capture prior information and input features better. These methods significantly improve the quality of face restoration.

3 Methods

We develop the model to enhance the realism, naturalness, and identity consistency in restored faces. The overall architecture is described in Sec. 3.1. The features are extracted using VRE and IDE, as described in Sec. 3.2. In Sec. 3.3, we introduce face alignment guidance to help the model converge faster and better leverage input and prior knowledge. Finally, in Sec. 3.4, we propose a new affine landmark distance metric to overcome the limitations of previous L2 metrics.

3.1 Model Formulation

Reference-based face restoration aims to reconstruct a high-quality (HQ) face $x_H \in \mathbb{R}^{H\times W\times 3}$ from a low-quality (LQ) input $x_L \in \mathbb{R}^{H\times W\times 3}$. This process is augmented by a set of N high-quality reference images $R = \{r_1, r_2, \ldots, r_N\}$, where each $r_i \in \mathbb{R}^{H\times W\times 3}$.

Our approach operates within a latent diffusion framework [42]. LQ image x_L is encoded to its latent representation $z_L = E_\phi(x_L)$, and similarly, HQ image x_H is encoded to $z_H = E_\phi(x_H)$ for training.

The forward diffusion process progressively introduces Gaussian noise to a clean latent vector z_0 (*e.g.*, z_H in our context) over a sequence of T timesteps:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ is the noise. The parameters $\alpha_t = 1 - \beta_t$ are determined by a predefined noise schedule $\beta_t \in (0, 1)$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denotes their cumulative product.

The reverse diffusion process endeavors to denoise z_t and recover an estimate of z_0. This is typically accomplished by training a noise prediction network $\varepsilon_\theta(z_t, p, t)$ to approximate the noise ε, conditioned on the noisy latent z_t, a prompt embedding $p = p(x_L, R)$, and the timestep t. The estimated clean latent $\hat{z}_0$ can then be derived from z_t and the predicted noise $\hat{\varepsilon} = \varepsilon_\theta(z_t, p, t)$ as follows:

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\varepsilon}}{\sqrt{\bar{\alpha}_t}}. \tag{2}$$

Once $\hat{z}_0$ (which corresponds to the estimated HQ latent $\hat{z}_H$) is obtained, the final restored image is synthesized by the VAE decoder: $\hat{x}_H = D_\phi(\hat{z}_H)$.

During inference, the LQ input x_L is encoded to its latent form $z_L = E_\phi(x_L)$. The one-step diffusion model [63, 49] will be achieved by a predetermined fixed timestep T_L (where $0 \leq T_L < T$). The target high-quality (HQ) latent vector $\hat{z}_H$ is subsequently computed in a single step:

$$\hat{z}_H = \frac{z_L - \sqrt{1-\bar{\alpha}_{T_L}}\,\varepsilon_\theta(z_L, p, T_L)}{\sqrt{\bar{\alpha}_{T_L}}}. \tag{3}$$

This formulation enables an end-to-end, single-step restoration from the LQ latent z_L to the HQ latent $\hat{z}_H$. Then, denoting the entire model as a generator $\mathcal{G}_\psi$, the overall procedure can be represented as

$$\hat{x}_H = \mathcal{G}_\psi(x_L, p). \tag{4}$$

To optimize the one-step diffusion model ε_θ for high-fidelity generation, we employ an adversarial distillation strategy. The generative adversarial network (GAN) discriminator [44, 43, 23], and the variational score distillation (VSD) [33, 60, 63] are widely recognized for significantly enhancing the performance of one-step diffusion models. By introducing the discriminator presented in OSDFace [57], we integrate the GAN discriminator to jointly optimize both the generator $\mathcal{G}_\psi$ and the discriminator $\mathcal{D}_\psi$. The corresponding loss functions are defined as:

$$\begin{aligned} \mathcal{L}_\mathcal{G} &= -\mathbb{E}_t\left[\log \mathcal{D}_\psi\left(F(\hat{z}_H, t)\right)\right], \\ \mathcal{L}_\mathcal{D} &= -\mathbb{E}_t\left[\log\left(1 - \mathcal{D}_\psi\left(F(\hat{z}_H, t)\right)\right)\right] - \mathbb{E}_t\left[\log \mathcal{D}_\psi\left(F(E_\phi(x_H), t)\right)\right], \end{aligned} \tag{5}$$

where $F(\cdot, t)$ denotes the forward diffusion process applied to the input at timestep $t \in [0, T]$.

3.2 Conditional Embeddings

The prompt embedding p utilized in Eq. (3) provides crucial guidance for the face restoration process. Its primary objective is to provide targeted semantic information, guiding the generation of high-quality latent representation $\hat{z}_H$. To avoid semantic ambiguities that arise from direct analysis of LQ input [64, 63], and to eliminate the need for additional textual prompts [7, 51], our framework integrates two specialized prompt embedding modules: our proposed identity embedder (IDE) and visual representation embedder (VRE) [57]. These modules collaboratively extract conditioning information directly from the input LQ image x_L and reference images R. For convenience, we assume the final prompt embedding p is a sequence of n tokens, each of dimension k, *i.e.*, $p \in \mathbb{R}^{n \times k}$.

Identity Embedder (IDE). The identity embedder (IDE) is principally focused on enhancing the identity fidelity of the restored face. As shown in Fig. 1, the IDE comprises two parallel components: a facial feature extractor and a face identity encoder.

Facial Feature Extractor. Key facial characteristics, such as the eyes, nose, and mouth, are important for human identity perception. Variations in eye color, skin texture, and edge structure significantly affect an individual's distinct appearance. Given that the input LQ image often suffers from considerable degradation, extracting such detailed identity solely from x_L is unreliable. Consequently, leveraging information from high-quality reference images becomes essential.

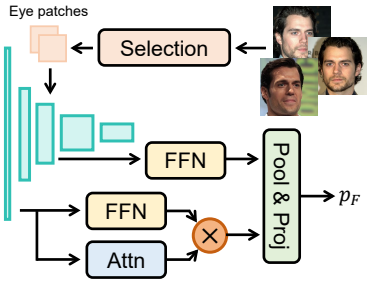

Figure 2: Facial feature extractor.

We observe that certain facial attributes, particularly ocular features, exhibit strong consistency for the same individual across different images. This allows us to select an optimal reference image that closely matches the facial characteristics of the LQ input. Let $R = \{r_1, r_2, \ldots, r_N\}$ be the set of available reference images. The optimal reference image r_0 could be identified by

$$r_0 = \arg\min_{r_i \in R} d(r_i, x_L), \tag{6}$$

where $d(\cdot, \cdot)$ denotes the affine landmark distance, which is formally defined in Sec. 3.4.

From the selected reference r_0, we crop patches corresponding to the left eye $e_l \in \mathbb{R}^{H_e \times W_e \times 3}$ and the right eye $e_r \in \mathbb{R}^{H_e \times W_e \times 3}$. These patches are processed using a pre-trained VGG16 network [46] to extract distinct color and structural features.

Drawing inspiration from studies on VGG representations [13, 34], we utilize features from different depths. Early layers could capture low-level details like color and fine textures, while mid-level layers capture more complex patterns and structural information. As shown in Fig. 2, for an eye patch e_p (where $p \in \{l, r\}$), its layer 0 feature $f_{\text{vgg},0}(e_p)$ is processed to yield a color embedding $f_{\text{color},p}$, and its layer 2 feature $f_{\text{vgg},2}(e_p)$ is processed to derive a structural embedding $f_{\text{struct},p}$. The resulting four embeddings are then concatenated and projected to form the extractor output $p_F \in \mathbb{R}^{n_1 \times k}$.

Face Identity Encoder. Extracting robust global identity information is a central challenge in face recognition. Many face recognition models work hard to make embeddings from the same identity clustered closely, while keeping embeddings from different identities apart. We process the set of aligned reference images $R = \{r_1, r_2, \ldots, r_N\}$ using such a pre-trained face recognition model X_{FR}. This yields a set of identity embeddings $E_{\text{id}} = \{e_{\text{id},1}, e_{\text{id},2}, \ldots, e_{\text{id},N}\}$, where each $e_{\text{id},i} = X_{\text{FR}}(r_i) \in \mathbb{R}^{d_{\text{id}}}$. This set of embeddings E_{id} is then aggregated by an attention-based module, shown in Fig. 3, to select the most important identity cues selectively. It could effectively pool identity information without distorting the learned feature space, thus producing the final output $p_I \in \mathbb{R}^{n_2 \times k}$.

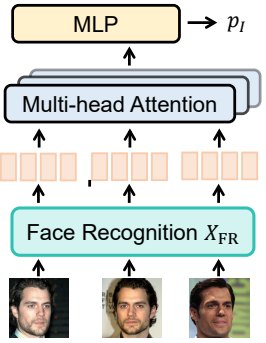

Figure 3: Illustration of face identity encoder.

Visual Representation Embedder (VRE). Extracting pertinent information directly from the LQ input x_L, despite its degradation, is also vital. Visual representation embedder (VRE) [57] typically consists of a pre-trained VQ-VAE architecture, featuring an encoder E_{VRE} specialized for LQ inputs. This encoder tokenizes x_L, then puts the tokens to query a learned codebook Dict_{LQ} via nearest-neighbor lookup to produce the visual prompt p_L:

$$p_L = \text{Dict}_{\text{LQ}}(E_{\text{VRE}}(x_L)). \tag{7}$$

The resulting $p_L \in \mathbb{R}^{n_3 \times k}$ captures distilled representations from x_L.

Finally, the IDE components and VRE run in parallel to produce feature sequences $p_F \in \mathbb{R}^{n_1 \times k}$, $p_I \in \mathbb{R}^{n_2 \times k}$, and $p_L \in \mathbb{R}^{n_3 \times k}$. These embeddings are integrated by feature fusion, to yield the final prompt embedding $p \in \mathbb{R}^{n \times k}$ (where $n \leq n_1 + n_2 + n_3$), which is input to the diffusion model.

3.3 Face Alignment Guidance

To restore faces that faithfully reflect the original subject's features, it is vital to design effective guidance mechanisms during training. Although the adversarial loss introduced in Sec. 3.1 helps align the overall distribution of restored and real images, it does not focus on tiny properties inherent to real-world faces. For example, the faithful reproduction of natural-looking textures, such as those found in hair and skin, including fine wrinkles and unique blemishes, is crucial for "honest". Therefore, our model uses several loss functions to encourage both global harmoniousness and local coherence. These include an identity loss, a global perceptual loss, and our proposed masked face alignment (MFA) method for enhancing localized perceptual fidelity.

Identity Preservation Loss. Accurately preserving facial identity is essential for truthful face restoration. Recent advances in face recognition [8, 20] show that aligned faces can be mapped to highly distinctive identity embeddings using compact neural networks. Based on this, we define the identity loss $\mathcal{L}_{\text{ID}}$. To reduce the bias that may arise from relying on a single recognition model, we employ an ensemble of M distinct pre-trained facial identity extractors, denoted as $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$. Each model $f_i \in \mathcal{F}$ maps an input face to a respective identity embedding. The identity loss is then computed as the sum of cosine distances between the embeddings of the ground truth image x_H and the restored image $\hat{x}_H$:

$$\mathcal{L}_{\text{ID}} = \sum_{i=1}^{M} \left(1 - \frac{f_i(x_H) \cdot f_i(\hat{x}_H)}{\|f_i(x_H)\|_2 \cdot \|f_i(\hat{x}_H)\|_2}\right). \tag{8}$$

This objective encourages $\hat{x}_H$ to share the same identity features as x_H from the collective perspective of multiple sophisticated recognition systems.

Global Perceptual Loss. To ensure that the overall structural and textural attributes of the restored image $\hat{x}_H$ align with those of the ground truth x_H from a human perceptual standpoint, we incorporate a global perceptual loss $\mathcal{L}_{\text{Per}}$. We utilize the DISTS metric [9], which is engineered to assess

image similarity in a manner that correlates well with human subjective judgment of image quality. Consistent with prior works [23, 57] that have demonstrated the efficacy of an edge-aware DISTS (EA-DISTS) variant, we adopt a similar formulation. The loss is given by:

$$\mathcal{L}_{\text{Per}} = \mathcal{L}_{\text{dists}}(x_H, \hat{x}_H) + \mathcal{L}_{\text{dists}}(\mathcal{S}(x_H), \mathcal{S}(\hat{x}_H)), \tag{9}$$

where $\mathcal{L}_{\text{dists}}$ represents the DISTS score. The Sobel operator $\mathcal{S}(\cdot)$ could extract edge maps, thus extracting prominent structural components, placing additional emphasis on structural consistency.

Local Perceptual Loss via Masked Face Alignment. Human attention focuses mainly on distinctive facial details, such as moles, scars, or wrinkles near the eyes and mouth. While other areas, like the cheeks or hair, may only need natural-looking textures rather than exact replication. Previous approaches [57] often treated all regions equally, potentially expending model capacity on less critical areas while missing fine details. To address this, our proposed masked face alignment (MFA), shown in Fig. 4, introduces a localized perceptual loss $\mathcal{L}_{\text{MFA}}$ that strategically focuses on perceptually significant facial regions.

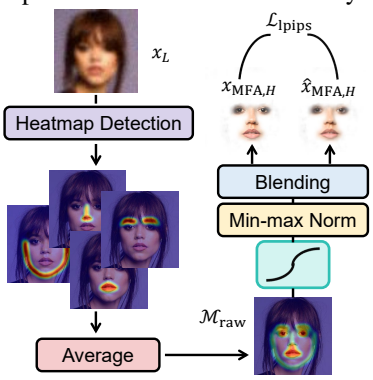

Figure 4: Masked face alignment.

Firstly, MFA creates an attention mask derived from heatmap predictions on the input image. Heatmap generation is a well-established technique in tasks such as facial landmark detection [39, 1, 59, 22, 35]. These methods typically perform spatial feature matching to capture local details. Each heatmap channel often corresponds to a specific landmark or facial region. By aggregating these channel-wise heatmaps, we can construct a composite mask highlighting perceptually important areas. Additionally, heatmap generation techniques are robust to severe image degradation [1], thereby allowing us to obtain reliable attention masks even from low-quality inputs.

Specifically, we use a pre-trained heatmap detection model X_{HM} to process the LQ input x_L. The resulting multi-channel heatmap is averaged to produce a single-channel attention map $\mathcal{M}_{\text{raw}}$:

$$\mathcal{M}_{\text{raw}} = \text{avg}_{\text{channel}}(X_{\text{HM}}(x_L)) \in \mathbb{R}^{H \times W}. \tag{10}$$

To enhance the contrast of the attention map and render it suitable as a blending mask, we apply an exponential transformation followed by min-max normalization. The final normalized attention mask $\mathcal{M} \in [0, 1]^{H \times W}$ is then obtained as:

$$\mathcal{M} = \frac{1 - e^{-k \cdot \mathcal{M}_{\text{raw}}} - \min(1 - e^{-k \cdot \mathcal{M}_{\text{raw}}})}{\max(1 - e^{-k \cdot \mathcal{M}_{\text{raw}}}) - \min(1 - e^{-k \cdot \mathcal{M}_{\text{raw}}}) + \epsilon}, \tag{11}$$

where $k > 0$ controls the sharpening effect, and $\epsilon > 0$ is a small constant introduced for numerical stability. This transformation improves salient regions while preserving the overall spatial structure.

The normalized attention mask $\mathcal{M}$ is subsequently used to selectively blend the ground truth image x_H and the restored image $\hat{x}_H$ with a neutral background $B \in \mathbb{R}^{H \times W \times 3}$. This procedure focuses the subsequent perceptual loss calculation on the identified salient regions:

$$x_{\text{MFA},H} = \mathcal{M} \odot x_H + (1 - \mathcal{M}) \odot B, \quad \hat{x}_{\text{MFA},H} = \mathcal{M} \odot \hat{x}_H + (1 - \mathcal{M}) \odot B, \tag{12}$$

where $\odot$ denotes element-wise multiplication. This masking procedure emphasizes perceptually important areas for targeted color and texture fidelity, thereby reducing problems such as global color shifts or texture inconsistencies in critical facial features.

The local perceptual loss, $\mathcal{L}_{\text{MFA}}$, is then computed using the LPIPS metric [77], known for its correlation with human perception of image similarity, on these masked images:

$$\mathcal{L}_{\text{MFA}} = \mathcal{L}_{\text{lpips}}(x_{\text{MFA},H}, \hat{x}_{\text{MFA},H}). \tag{13}$$

The targeted loss directs the model to focus on the regions identified by the heatmap, ensuring that subtle textures and natural colors are reconstructed accurately.

3.4 Affine Landmark Distance

Facial landmarks are a mature cue for face detection and recognition [76, 48, 50]. Because the geometric structure of a human face is fixed, the landmark contour is smooth and seldom exhibits large deformations or outliers. Consequently, landmark-detection networks can be trained to be highly robust, and landmark-based metrics remain reliable even when the image is severely degraded.

The most common landmark metric is the pixel-wise L2 distance [15]. Since its introduction, many studies [5, 73, 52, 57] have adopted it to quantify the fidelity of identity information. In practical

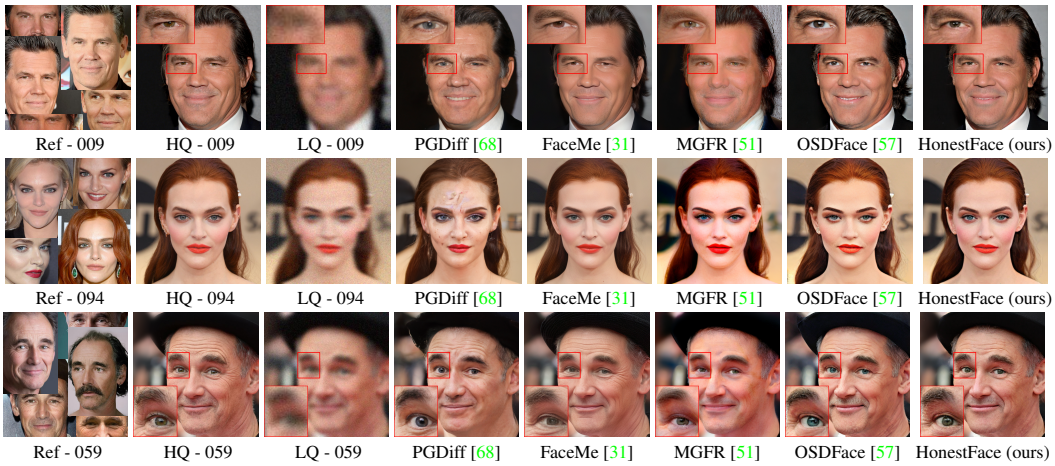

Figure 5: Visual comparison of CelebHQRef-Test. Please zoom in for a better view.

applications, however, a restored image often shows a slight global misalignment with the ground truth. This small offset can be caused by severe degradation in the input, randomness during the restoration, or slight errors made by the landmark detector. These shifts usually appear as small translations, including tiny translations, mild in-plane rotations, small changes of scale, or another affine transform. Because the standard L2 distance is measured in the image plane, it is very sensitive to such shifts. As it is shown in Fig. 6, even when a model recovers fine facial details, a minor global displacement can enlarge the L2 error and lead to an unfairly low score.

Let the landmark of the restored image set be $L = \{l_k\}_{k=1}^N$, and the ground truth HQ landmark set be $H = \{h_k\}_{k=1}^N$, where each landmark $l_k, h_k \in \mathbb{R}^2$. We also define positive weights $W = \{w_k\}_{k=1}^N$ associated with each landmark pair (l_k, h_k), satisfying $w_k > 0$. The central idea is to find an optimal 2D affine transformation matrix $A \in \mathbb{R}^{2\times 3}$ that maps the restored landmarks l_k to their corresponding HQ landmarks h_k. Specifically, we represent each restored landmark using homogeneous coordinates as $l'_k = [l_{k,x}, l_{k,y}, 1]^\intercal \in \mathbb{R}^3$, and define the affine landmark distance $d_{\text{A-LD}}(L, H; W)$ as the minimal sum of weighted squared Euclidean distances, formulated as

$$d_{\text{A-LD}}(L, H; W) = \min_A \sum_{k=1}^{N} w_k \|Al'_k - h_k\|^2. \quad (14)$$

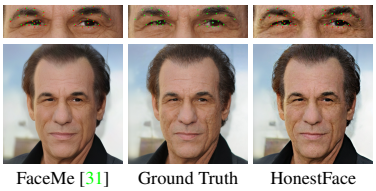

Figure 6: Landmark comparisons for Reface-Test 159. Landmark for ground truth is colored by green. Although FaceMe has a lower L2-LD for the right eyebrow, HonestFace achieves better fidelity, with a smaller A-LD calculated on eyes.

This optimization corresponds to a weighted linear least-squares problem, which has a unique closed-form solution provided that L is non-collinear and $N \geq 3$. Denote the homogeneous coordinate matrix as $L' = [l'_1, l'_2, \ldots, l'_N] \in \mathbb{R}^{3\times N}$, the HQ landmark matrix as $H = [h_1, h_2, \ldots, h_N] \in \mathbb{R}^{2\times N}$, and the diagonal weight matrix as $W = \text{diag}(w_1, w_2, \ldots, w_N)$. The closed-form solution for the optimal affine transformation matrix A is

$$A = (HWL'^\intercal)(L'WL'^\intercal)^{-1}. \quad (15)$$

Although A-LD is not a direct measure of visual quality, it reflects landmark geometry in a way that matches human face perception. People are insensitive to the scale and position of a face in the view and instead attend to the relative layout of facial features. A-LD measures the "remaining" landmark error, thus, it can describe geometric similarity more accurately than the raw pixel-wise L2 distance.

4 Experiments

4.1 Experimental Settings

Datasets. HonestFace is fine-tuned on the Reface-HQ training dataset [51] and the last 905 identities from CelebRef-HQ [29]. Reface-HQ contains 23,500 high-quality facial images from 5,250 identities, with 4,870 used for training and 380 reserved for testing (Reface-Test). CelebRef-HQ includes 1,005 identities and a total of 10,555 images. All images are resized to 512×512 before training. Synthetic low-quality data is generated using the same degradation pipeline as VQFR [15]. For inference, we evaluate our method on Reface-Test and the first 100 identities from the CelebRef-HQ dataset.

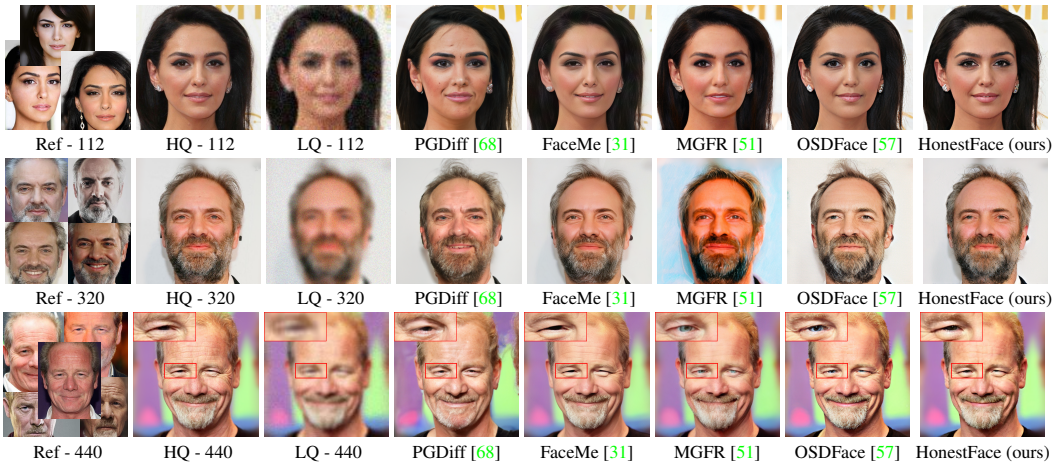

Figure 7: Visual comparison of Reface-Test. Please zoom in for a better view.

Datasets	Methods	LPIPS↓	DISTS↓	C-IQA↑	M-IQA↑	MUSIQ↑	FID↓	Deg.↓	L2-LD↓	ALD-e↓	ALD-m↓
CelebHQRef Test	ASFFNet [28]	0.1969	0.1411	0.6067	0.5346	71.181	45.618	35.636	3.6047	16.422	11.701
	DMDNet [29]	0.2191	0.1499	0.6417	0.5165	71.119	52.406	34.258	2.6437	15.380	10.347
	PGDiff [68]	0.2674	0.1559	0.5735	0.4953	70.762	60.693	46.248	3.6478	18.687	12.032
	FaceMe [31]	0.1945	0.1412	0.6606	0.5758	71.598	46.423	30.578	2.0107	14.673	9.4460
	MGFR [51]	0.2311	0.1520	0.6224	0.6475	70.626	52.103	38.322	2.5274	16.344	10.614
	HonestFace (ours)	**0.1809**	**0.1246**	**0.6673**	**0.6858**	**71.996**	**41.021**	**26.106**	**1.7798**	**13.444**	**8.9793**
Reface Test	ASFFNet [28]	0.2012	0.1422	0.6043	0.5188	70.071	32.046	36.268	3.4611	16.577	11.263
	DMDNet [29]	0.2234	0.1510	0.6394	0.5028	70.278	36.867	35.282	2.6880	15.700	9.9783
	PGDiff [68]	0.2701	0.1575	0.5632	0.4822	70.607	43.946	46.923	3.4423	18.165	11.635
	FaceMe [31]	0.2021	0.1473	0.6595	0.5881	70.426	31.483	32.021	1.9793	14.596	9.3373
	MGFR [51]	0.2307	0.1535	0.6467	0.6436	70.464	34.167	38.762	2.4641	16.043	9.9218
	HonestFace (ours)	**0.1784**	**0.1235**	**0.6818**	**0.6717**	**71.608**	**25.755**	**28.173**	**1.7369**	**13.581**	**8.7287**

Table 1: Comparison with the reference-based methods. C-IQA stands for CLIP-IQA, and M-IQA stands for MANIQA. ALD-e and ALD-m are the affine landmark distances for the eyes and mouth.

Metrics. We evaluate pixel-level restoration fidelity using PSNR and SSIM. Perceptual metrics include LPIPS [77], DISTS [9], and no-reference methods such as CLIP-IQA [55], MANIQA [69], and MUSIQ [18]. FID is calculated between the ground truth and restored faces. For identity consistency, as it is crucial for honest restoration, we assess ArcFace [8] embedding angular differences "Deg.", L2 landmark distance, and our proposed affine landmark distance (A-LD). Specifically, A-LD utilizes landmarks from the eye region (eyes and eyebrows) and the mouth border.

Implementation Details. We adopt OSDFace [57] as our baseline model, which has demonstrated excellent performance in blind face restoration. For fine-tuning ε_θ, *i.e.*, the UNet, we utilize LoRA [17] with both the rank and alpha set to 16. The training process spans 60K iterations across 4 NVIDIA A6000 GPUs, with a batch size of 4. The prompt embedding consists of 77 tokens (n=77), where n_1=4, n_2=7, and n_3=77. For the face identity encoder, we use ArcFace [8] as X_{FR}. In terms of face alignment guidance, we leverage both ArcFace [8] and AdaFace [20] for feature extraction, and k=10 for MFA sharpening. The learning rate is set to 1×10^{-4}, with AdamW [32] optimizer.

4.2 Main Results

To assess the performance in generating realistic facial textures and preserving identity consistency, we compare HonestFace with state-of-the-art reference-based face restoration methods, including ASFFNet [28], DMDNet [29], PGDiff [68], FaceMe [31], and MGFR [51]. Additionally, to emphasize the texture honesty, we further evaluate transformer-based methods, such as CodeFormer [78] and DAEFR [52], as well as one-step diffusion methods like OSEDiff [63] and OSDFace [57].

Quantitative Results. Table 1 shows that HonestFace achieves state-of-the-art results across all metrics. In particular, identity-aware metrics such as Deg., L2-LD, and A-LD demonstrate HonestFace's strong identity fidelity. Our method also performs excellently on perceptual metrics.

A broader comparison with more face restoration methods is presented in Tab. 2. Since these methods do not utilize reference images, we focus the comparison solely on image quality. Pixel-level metrics demonstrate that our method outperforms other diffusion-based approaches, while higher perceptual-level scores indicate better restoration of textures and fine details.

Methods	CelebHQRef-Test						Reface-Test					
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	M-IQA↑	FID↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	M-IQA↑	FID↓
CodeFormer [78]	**25.49**	0.7100	**0.1791**	0.1419	0.5827	48.37	**25.65**	0.7082	0.1872	0.1497	0.5699	33.57
DAEFR [52]	22.85	0.6549	0.2058	0.1439	0.6019	49.35	23.07	0.6574	0.2092	0.1498	0.5842	34.10
OSEDiff [63]	24.07	0.7105	0.2077	0.1533	0.5229	57.06	24.21	0.7110	0.2076	0.1553	0.5122	46.31
OSDFace [57]	23.51	0.6742	0.2050	0.1373	0.6316	46.51	23.54	0.6718	0.2156	0.1481	0.6257	32.59
HonestFace (ours)	25.17	**0.7147**	0.1809	**0.1246**	**0.6858**	**41.02**	25.25	**0.7118**	**0.1784**	**0.1235**	**0.6717**	**25.75**

Table 2: Quantitative comparison with the no-reference methods. M-IQA stands for MANIQA.

Methods	PGDiff [68]	MGFR [51]	OSEDiff [63]	OSDFace [57]	HonestFace
Step	1,000	50	1	1	1
Time (s)	85.81	6.9	0.13	0.10	0.13
Param (M)	176.4	2,029.3	1,302	978.4	1,059
MACs (G)	480,997	≈2,672	2,269	2,132	2,282

(a) Complexity comparison during inference.

w. LPIPS	w. DISTS	w. MSE	C-IQA↑	M-IQA↑	Deg.↓
✗	✗	✗	0.6585	0.6822	28.271
✗	✗	✓	0.6533	0.6532	27.276
✗	✓	✗	0.6582	0.6811	27.324
✓	✗	✗	**0.6673**	**0.6858**	**26.106**

(b) Ablation study of MFA.

FFE	FIE w. Arc	FIE w. Ada	LPIPS↓	DISTS↓	Deg.↓	L2-LD↓	ALD-e↓	ALD-m↓
✗	✗	✗	0.1855	0.1282	30.052	1.9013	14.018	9.1584
✓	✗	✗	0.1835	0.1265	29.322	1.8873	13.661	9.1960
✗	✓	✗	0.1863	0.1288	27.338	1.8790	13.646	9.4423
✓	✗	✓	0.1816	0.1259	27.490	1.8130	13.579	9.1427
✓	✓	✗	**0.1809**	**0.1246**	**26.106**	**1.7798**	**13.444**	**8.9793**

(c) Ablation study of IDE.

Methods	CelebHQRef-Test	Reface-Test
MGFR [51]	0.8000	0.7674
OSEDiff [63]	0.4646	0.4388
OSDFace [57]	0.5634	0.5716
HonestFace (ours)	**0.3461**	**0.3356**

(d) MSSWD [16] metric for color IQA.

Table 3: More experiments for complexity comparison, color shift analysis, and ablation studies. All ablation study experiments are conducted on CelebHQRef-Test.

Qualitative Results. Visual comparisons in Figs. 5, 7 show that HonestFace performs better. It avoids over-smoothing issues seen in FaceMe [31], and does not alter eye or skin color like MGFR [51] and OSDFace [57]. HonestFace effectively preserves both identity and expression, especially in key regions critical for recognition, such as the eyes, mouth, and skin wrinkles.

4.3 Complexity Analysis

Table 3a compares model complexity, including steps, inference time, parameter count, and multiply-accumulate operations (MACs). All tests are conducted on an NVIDIA A6000 GPU with 512×512 input for consistency. Although HonestFace adds new modules to support multi-reference face restoration, it remains competitive with other one-step diffusion face restoration models.

4.4 Color Shift

Perceptual color difference remains a key challenge in image restoration. Due to the strong generative capability, diffusion-based methods tend to produce noticeable color inconsistencies, as clearly illustrated in Figs. 5, 7. We further evaluate color fidelity using MS-SWD [16], a color-aware image quality metric. As shown in Tab. 3d, experimental results demonstrate that our approach outperforms other methods, under both reference-based and no-reference settings.

4.5 Ablation Studies

Identity Embedder (IDE). To select the IDE structure and verify its effectiveness, we conduct an ablation study in Tab. 3c. The results show that adding the facial feature extractor (FFE) improves eye feature consistency, while adding the face identity encoder (FIE) lowers the overall Deg.. Combining FFE and FIE leads to deeper identity understanding and better fidelity consistency.

Masked Face Alignment (MFA). We evaluate the impact of heatmap weighting on texture and identity consistency, as shown in Tab. 3b. We also explore different loss functions for $x_{\text{MFA},H}$ and $\hat{x}_{\text{MFA},H}$, and choose LPIPS as the local perceptual loss finally. It is also worth mentioning that using MSE as the loss causes over-smoothing in facial regions, making it unsuitable here.

5 Conclusion

In this paper, we address the critical challenges in "honest" face restoration: achieving natural, realistic textures while maintaining strict identity fidelity. We propose HonestFace, a novel one-step diffusion model that effectively integrates multiple reference images. HonestFace incorporates the identity embedder for robust identity preservation and masked face alignment to enhance perceptual naturalness. Furthermore, we developed a new affine landmark distance metric for a more precise evaluation of identity consistency. Our experiments demonstrate that HonestFace achieves SOTA performance, delivering faces that are faithful to the subject's identity and exhibit realistic textures.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In *CVPR*, 2018. 6

[2] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. GLEAN: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 1

[3] Chaofeng Chen, Xiaoming Li, Yang Lingbo, Xianhui Lin, Lei Zhang, and Kwan-Yee K. Wong. Progressive semantic-aware style transformation for blind face restoration. In *CVPR*, 2021. 1

[4] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 3

[5] Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaocun Cao. Towards real-world blind face restoration with generative diffusion prior. *arXiv preprint arXiv:2312.15736*, 2023. 1, 6

[6] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. FSRNet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 2

[7] Zheng Chen, Yulun Zhang, Jinjin Gu, Xin Yuan, Linghe Kong, Guihai Chen, and Xiaokang Yang. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2303.06373*, 2023. 4

[8] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5, 8

[9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 2020. 5, 8

[10] Zheng Ding, Xuaner Zhang, Zhuowen Tu, and Zhihao Xia. Restoration by generation with constrained priors. In *CVPR*, 2024. 3

[11] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 2

[12] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *CVPR*, 2025. 3

[13] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *NeurIPS*, 2015. 5

[14] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 2

[15] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*, 2022. 6, 7

[16] Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced Wasserstein distances as perceptual color difference measures. In *ECCV*, 2024. 9

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 8

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer . In *ICCV*, 2021. 8

[19] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. In *BMVC*, 2019. 2

[20] Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *CVPR*, 2022. 5, 8

[21] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3

[22] Xing Lan, Qinghao Hu, and Jian Cheng. Revisting quantization error in face alignment. In *ICCVW*, 2021. 6

[23] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Unleashing the power of one-step diffusion based image super-resolution via a large-scale diffusion discriminator. *arXiv preprint arXiv:2410.04224*, 2024. 4, 6

[24] Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. In *ICML*, 2025. 3

[25] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 2

[26] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 2

[27] 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. 2

[28] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, 2020. 2, 8

[29] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE TPAMI*, 2022. 3, 7, 8

[30] 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 *ECCV*, 2024. 1

[31] Siyu Liu, Zheng-Peng Duan, Jia OuYang, Jiayi Fu, Hyunhee Park, Zikun Liu, Chun-Le Guo, and Chongyi Li. FaceMe: Robust blind face restoration with personal identification. In *AAAI*, 2025. 2, 3, 7, 8, 9

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 8

[33] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In *NeurIPS*, 2023. 4

[34] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them . In *CVPR*, 2015. 5

[35] James McCouat and Irina Voiculescu. Contour-hugging heatmaps for landmark detection. In *CVPR*, 2022. 6

[36] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 2

[37] Yunqi Miao, Jiankang Deng, and Jungong Han. WaveFace: Authentic face restoration with efficient frequency recovery. In *CVPR*, 2024. 1

[38] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM TOG*, 2022. 3

[39] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using CNNs. In *MICCAI*, 2016. 6

[40] Xinmin Qiu, Congying Han, Zicheng Zhang, Bonan Li, Tiande Guo, and Xuecheng Nie. DiffBFR: Bootstrapping diffusion model for blind face restoration. In *ACM MM*, 2023. 1

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *CVPR*, 2022. 3

[43] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *ACM SIGGRAPH Asia*, 2024. 4

[44] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024. 4

[45] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 2

[46] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[47] Maitreya Suin and Rama Chellappa. CLR-Face: Conditional latent refinement for blind face restoration using score-based diffusion models. In *IJCAI*, 2024. 1

[48] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 6

[49] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *CVPR*, 2025. 3, 4

[50] Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. Lafs: Landmark-based facial self-supervised learning for face recognition. In *CVPR*, 2024. 6

[51] Keda Tao, Jinjin Gu, Yulun Zhang, Xiucheng Wang, and Nan Cheng. Overcoming false illusions in real-world face restoration with multi-modal guided diffusion model. In *ICLR*, 2025. 2, 3, 4, 7, 8, 9

[52] Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin CK Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. In *ICLR*, 2024. 1, 2, 6, 8, 9

[53] Tuomas Varanka, Tapani Toivonen, Soumya Tripathy, Guoying Zhao, and Erman Acar. PFStorer: Personalized face restoration and super-resolution. In *CVPR*, 2024. 2, 3

[54] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, 2020. 2

[55] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 8

[56] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 3

[57] Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. OSDFace: One-step diffusion model for face restoration. In *CVPR*, 2025. 1, 3, 4, 5, 6, 7, 8, 9

[58] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2

[59] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, 2019. 6

[60] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 4

[61] Zhixin Wang, Xiaoyun Zhang, Ziying Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *CVPR*, 2023. 1

[62] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE TPAMI*, 2023. 1

[63] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024. 1, 2, 4, 8, 9

[64] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 3, 4

[65] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer. *arXiv preprint arXiv:2410.10629*, 2024. 3

[66] Lianxin Xie, Csbingbing Zheng, Wen Xue, Le Jiang, Cheng Liu, Si Wu, and Hau San Wong. Learning degradation-unaware representation with prior-based latent transformations for blind face restoration. In *CVPR*, 2024. 1

[67] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *ACM MM*, 2020. 2

[68] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023. 3, 7, 8, 9

[69] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *CVPRW*, 2022. 8

[70] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 1

[71] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 3

[72] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *CVPR*, 2018. 2

[73] Zongsheng Yue and Chen Change Loy. DifFace: Blind Face Restoration with Diffused Error Contraction . *IEEE TPAMI*, 2024. 1, 6

[74] Libing Zeng, Lele Chen, Yi Xu, and Nima Khademi Kalantari. Mystyle++: A controllable personalized generative prior. In *ACM SIGGRAPH Asia*, 2023. 3

[75] Howard Zhang, Yuval Alaluf, Sizhuo Ma, Achuta Kadambi, Jian Wang, and Kfir Aberman. Instantrestore: Single-step personalized face restoration with shared-image attention. *arXiv preprint arXiv:2412.06753*, 2024. 3

[76] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 2016. 6

[77] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6, 8

[78] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 1, 2, 8, 9