RefSTAR: Blind Face Image Restoration with Reference Selection, Transfer, and Reconstruction

Zhicun Yin[1, 4], **Junjie Chen**[1], **Ming Liu**[1(✉)], **Zhixin Wang**[2], **Fan Li**[2], **Renjing Pei**[2], **Xiaoming Li**[3], **Rynson W.H. Lau**,[4], **Wangmeng Zuo**[1]
[1]Harbin Institute of Technology,
[2]Huawei Noah's Ark Lab,
[3]Nanyang Technological University,
[4]City University of Hong Kong
cszcyin@outlook.com, csmliu@outlook.com, csxmli@gmail.com, wmzuo@hit.edu.cn

Abstract

Introducing high-quality references can largely alleviate the uncertainty in blind face image restoration tasks, yet the equivocal utilization of reference priors makes it still a struggle to well preserve the human identity. We attribute the identity inconsistency to two deficiencies of existing reference-based face restoration methods, namely the inability to effectively determine which features need to be transferred, and the failure to preserve the structure and details of the selected features. This work mainly focuses on these two issues, and we present a novel blind face image restoration method that considers **ref**erence **s**election, **t**ransfer, **a**nd **r**econstruction (RefSTAR) to introduce proper features from reference images. Specifically, we construct a reference selection (RefSel) module, which can generate accurate masks to select reference features. For training the RefSel module, we construct a RefSel-HQ dataset through a mask generation pipeline, which contains annotated masks for 10,000 ground truth-reference pairs. To guarantee the exact introduction of selected reference features, a feature fusion paradigm is designed for reference feature transferring, and a Mask-Compatible Cycle-Consistency Loss is redesigned based on reference reconstruction to further ensure the presence of selected reference image features in the output image. Experiments on various backbone models demonstrate superior performance, showing better identity preservation ability and reference feature transfer quality.

Code — https://github.com/yinzhicun/RefSTAR

Introduction

Blind face image restoration aims to recover high-quality face images from degraded inputs, without knowing the degradation type and parameters. While current methods (Li et al. 2020a; Yang et al. 2021; Wang et al. 2021a; Zhou et al. 2022; Wang et al. 2022; Yue and Loy 2024; Lin et al. 2024; Wang et al. 2024a) have made significant strides in generating plausible face structures, they still face challenges in preserving identity and faithfully restoring personalized textures, particularly when dealing with severely degraded inputs. As a result, reference-based face restoration methods (Li et al. 2018, 2020b, 2022; Hsiao et al. 2024; Ying et al. 2024; Liu et al. 2025; Zhang et al. 2024) have emerged as a promising solution, using high-quality images of the same person to provide identity-related features.

Ideally, when the reference and ground-truth images have perfectly identical textures, these methods can effectively transfer textures from the reference. However, in practice, images of the same person often exhibit various differences, such as poses, face expressions, and textures corresponding to different age stages. These discrepancies make the network uncertain about whether certain textures should be transferred. As a result, these methods struggle to effectively utilize reference images for more faithful restoration. An alternative approach in the literature is to constrain the model using one of the GT and the reference image that more closely resembles the output. However, the optimization trajectory tends to lean heavily towards the reference image, as the model struggles to recover details similar to the GT, especially when the input is severely degraded.

Therefore, how to effectively leverage references for faithful restoration is an important yet underexplored issue. To address this challenge, as shown in Fig. 1, we propose a new framework, RefSTAR, which methodically addresses reference feature selection, transfer, and reconstruction to improve restoration fidelity. Specifically, to achieve effective texture selection, we introduce a RefSel module to explicitly predict which regions of the reference should be transferred. For instance, if the reference image shows an open mouth and the input LR image features a closed mouth, the model needs to exclude the teeth from the reference image. For training the RefSel module, we manually annotated 10,000 pairs of ground truth and reference images, marking the regions where the textures are consistent between them. By randomly degrading the ground-truth images, we use the low-quality images along with the reference images as input of RefSel to predict the texture consistency regions. Unlike the implicit feature fusion scheme in existing methods, which conflates the features from the input and reference images under the constraint of the final outputs, our method identifies relevant features from the reference images and can be optimized in an explicit and thus more efficient way.

After identifying consistent texture regions, the next challenge is how to align the selected reference textures with the degraded images to facilitate effective feature transfer.

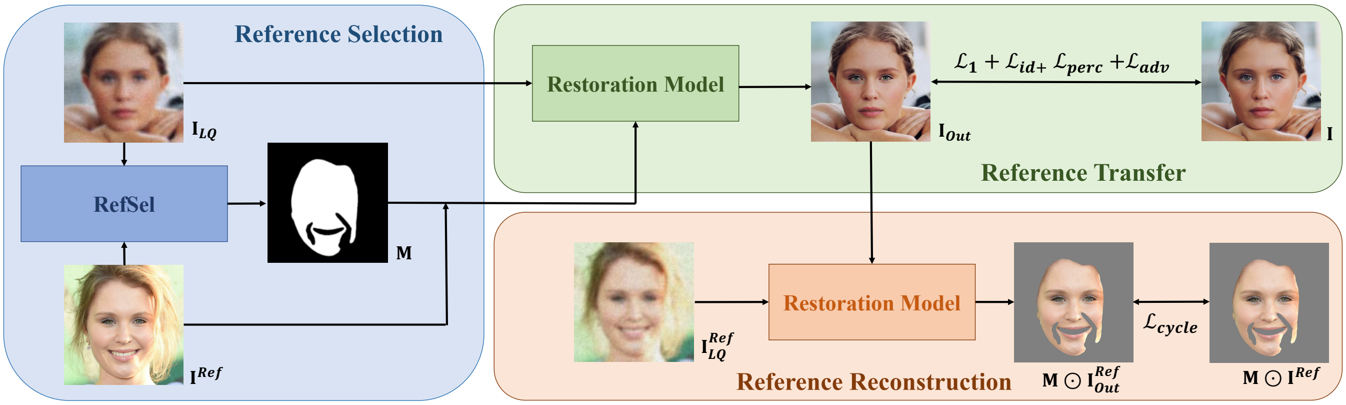

Figure 1: Pipeline of our proposed RefSTAR framework, including reference selection, transfer, and reconstruction. With the RefSel selecting proper reference features, the restoration model conducts the reference feature alignment under the supervision of our Mask-Compatible Cycle-Consistency Loss for reference reconstruction.

In the diffusion-based models (Rombach et al. 2022; Hu 2024), a straightforward and widely adopted approach is to use cross-attention mechanisms. However, standard cross-attention struggles with misalignment and can lead to suboptimal feature fusion, where the model either ignores useful reference details or naively copies textures without proper adaptation. To address this issue, we use a decoupled cross-attention mechanism (Ye et al. 2023) for our restoration task that preserves feature hierarchy independence while enabling proper cross-stream interaction. By carefully guiding the fusion process, our approach prevents the model from indiscriminately copying reference features and instead facilitates a more context-aware and structurally coherent texture transfer, leading to higher-fidelity restoration.

Finally, to further enhance the transfer of selected reference textures, we introduce a Mask-Compatible Cycle-Consistency Loss that aids in reconstructing the reference image from the restored results. Specifically, the restored image is used as a new reference, while the original reference is degraded to serve as the low-quality input. This cycle consistency encourages the model to effectively preserve and transfer relevant textures from the reference, thereby improving the faithful restoration of the personalized details.

Experimental results demonstrate the effectiveness of our RefSel module in identifying texture-consistent regions, which in turn facilitates more precise and fine-grained texture transfer. As a result, our method achieves promising performance on both synthetic and real-world datasets, showcasing its robustness in handling diverse degradation scenarios and improving the fidelity of restored images. The main contributions are summarized as follows.

- We propose RefSTAR, a new framework for reference-guided face restoration, which explicitly addresses inconsistencies between degraded input and reference images.
- We introduce RefSel-HQ, the first high-quality dataset of 10K manually annotated Image-Ref-Mask triplets for exploring effective reference-guided texture transfer.
- We introduce a new reference reconstruction process, enhanced by a redesigned Mask-Compatible Cycle-Consistency Loss, to ensure the effective transfer of relevant reference textures.
- Extensive experiments are conducted on both synthetic and real-world images, demonstrating the effectiveness of the presented RefSTAR method in terms of identity preservation ability and reference feature transfer quality.

Related Works

Blind Face Image Restoration

Blind face image restoration is particularly challenging due to unknown degradation and the difficulty of simulating real-world degradations (Sharipov, Nutfullin, and Maloyan 2023; Li et al. 2023). Therefore, existing methods leverage structured priors to guide the restoration process. Early approaches primarily rely on geometric priors, such as face landmarks (Bulat and Tzimiropoulos 2018; Chen et al. 2018; Kim et al. 2019), parsing maps (Chen et al. 2021; Shen et al. 2018; Yang et al. 2020), and component heatmaps (Yu et al. 2018), to enforce semantic consistency in the reconstructed face structures. However, these methods struggle to generalize under complex real-world degradations. With the development of generative models, generative priors have been widely adopted, particularly those from GANs (Karras, Laine, and Aila 2019; Karras et al. 2020; Esser, Rombach, and Ommer 2021) and diffusion models (Rombach et al. 2022), which capture richer structural and textural details. GAN-based approaches exploit latent space representations to synthesize photorealistic textures. For instance, GFP-GAN (Wang et al. 2021a) and GPEN (Yang et al. 2021) leverage the latent space of StyleGAN (Karras, Laine, and Aila 2019) to restore face details, while CodeFormer (Zhou et al. 2022) utilizes a codebook-based generative prior (Esser, Rombach, and Ommer 2021) to achieve a balance between fidelity and realism. Diffusion-based methods take a different approach by iteratively refining the restoration results through a denoising process. Models such as DifFace (Yue and Loy 2024), DiffBIR (Lin et al. 2024), and PGDiff (Yang et al. 2023) demonstrate great restoration performance by progressively reconstructing missing details. To improve inference efficiency, OSDFace (Wang et al. 2024a) employs a one-step diffusion strategy with a visual representation embedder, significantly reducing computational overhead. While these methods achieve impressive results with powerful generative priors, their fidelity remains fundamentally constrained by the amount of useful information present in the degraded input.

Reference-based Face Restoration

For faithful face restoration, Li (Li et al. 2018, 2020b) first introduce the use of high-quality images of the same identity to guide identity recovery and preserve personalized details. Early reference-based methods enhanced restoration by directly incorporating residual features (Li et al. 2018, 2020b) or constructing reference dictionaries (Li et al. 2022). While these approaches outperform methods without reference images, they struggle to preserve fine-grained textures due to their limited ability to selectively transfer reliable details from the reference images. To address this limitation, recent methods incorporate personalized priors from diffusion-based customized image generation models (Ruiz et al. 2023; Yuan et al. 2023; Wang et al. 2024b). By integrating reference features into the iterative denoising process (Hsiao et al. 2024; Ying et al. 2024; Tao et al. 2024; Liu et al. 2025), these models effectively exploit identity information while benefiting from generative priors to handle extreme degradation. Moreover, InstantRestore (Zhang et al. 2024) employs a one-step inference strategy, achieving near real-time restoration with strong identity preservation. Despite these advancements, most reference-based methods rely heavily on direct ground-truth supervision. Variations in lighting, pose, and face expressions often introduce inconsistencies between the ground truth and reference images, causing models to disregard valuable reference details. Consequently, these approaches struggle to synthesize high-fidelity, personalized textures. Recently, RefineIR (Chong et al. 2025) proposed a loss formulation, which constrains the model with either the ground truth or the reference, based on which one is more similar to the output. However, recovering the details resembling the ground truth is difficult, especially when the input quality is poor, making the optimization bias towards the reference and resulting in excessive dependency that limits the generalization ability. Additionally, reliance on face landmark detection and image warping imposes further constraints, reducing robustness in cases where landmarks cannot be accurately extracted. Therefore, identifying which regions of the reference image to leverage and how much texture to transfer remains a fundamental challenge. In this work, we introduce a new framework that explicitly selects and transfers reference features for effective guidance and robust reconstruction.

Methods

As shown in Fig. 1, to incorporate appropriate features from high-quality reference images, our presented **RefSTAR** considers reference-guided face restoration from three aspects, namely **ref**erence **s**election, **t**ransfer, **a**nd **r**econstruction. The reference selection (RefSel) module predicts a mask to point out the regions whose features can be migrated to the final output with prediction probability, mainly based on the conflict level between the input and the reference image. To ensure the utilization of selected reference feature, we focus on both network structure and loss functions. The transfer component of our network provides a pathway for features from the reference image to be infused into the restoration model, where a decoupled cross-attention mechanism is deployed to force the infusion while avoiding explicit image/feature alignment. Finally, the reconstruction process redesigns the cycle consistency loss as Mask-Compatible Cycle-Consistency Loss, which compels the model output to contain detailed textures from the reference and is consistent with the predicted mask. More details about our RefSTAR are given in the following.

Reference Feature Selection

To ensure an effective reference selection in our RefSel module, we first construct a data engine to obtain the input, reference, and binary mask triplets, *i.e.*, $\{\mathbf{I}_{LQ}, \mathbf{I}^{Ref}, \mathbf{M}\}$, based on which RefSel can be effectively optimized.

Data Engine. To begin with, we collect over 10,000 high-quality image pairs from existing face image datasets (*e.g.*, CelebRef-HQ (Li et al. 2022)), serving as the ground truth $\mathbf{I}$ and reference $\mathbf{I}^{Ref}$, respectively. For effectively obtaining the mask $\mathbf{M}$ for sufficient $\{\mathbf{I}, \mathbf{I}^{Ref}\}$ pairs, a small portion (*i.e.*, 800) of the image pairs are manually annotated, based on which a binary segmentation U-Net (Ronneberger, Fischer, and Brox 2015) is trained. Subsequently, we generate masks using the U-Net and perform a manual filtering/adjusting process until we obtain 10,000 high-quality $\{\mathbf{I}, \mathbf{I}^{Ref}, \mathbf{M}\}$ triplets (at this point, we have generated approximately 12,000 masks). For generating low-quality input $\mathbf{I}_{LR}$, we adopt the degradation model of Real-ESRGAN (Wang et al. 2021b), with motion and defocus deblurring embedded for a more realistic degradation. The mask annotation protocol adhered to a dual-axis taxonomy, *i.e.*, the regions with the following conflicts are discarded.

- Dynamic conflicts capturing expression-induced mismatches (*e.g.*, mouth open/close, eye open/close, and perioral wrinkle discrepancies during the variation of different expressions like smile and frown).
- Static conflicts with feature misalignments arising from physiological traits (*e.g.*, pigmentation disparities like freckles) or some artificial accessories (*e.g.*, eyewear occlusion and texture clashes from heavy cosmetics).

Some examples of our dataset are shown in Fig. 2. Note that all the backgrounds and hair are masked to reduce the complexity of the problem. Note that when degradations are added to $\mathbf{I}$, the visibility of the face details may vary. Therefore, we adaptively replace the annotated mask with an all 1 mask (except for the hair and background regions) when the degradation is severe.

Reference Selection Module. Given the $\{\mathbf{I}_{LQ}, \mathbf{I}^{Ref}, \mathbf{M}\}$ triplets, we first train our RefSel module, which is initialized with the U-Net in the data engine, *i.e.*,

$$\mathbf{p} = f_{\mathbf{M}}([\mathbf{I}_{LQ}, \mathbf{I}^{Ref}]; \Theta_{\mathbf{M}}), \quad (1)$$

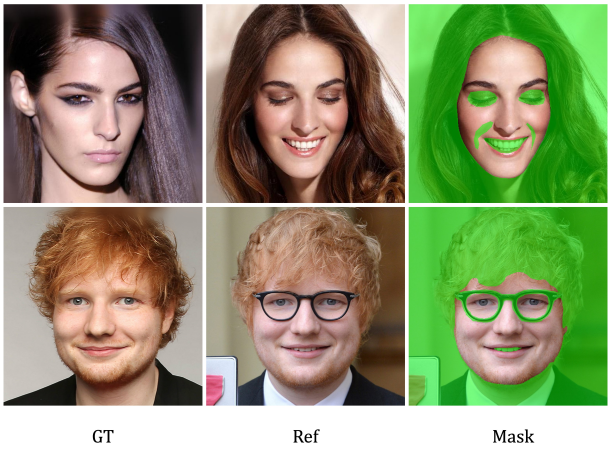

Figure 2: Examples of Annotation Pairs in our RefSel-HQ.

where $\mathbf{p}$ is the masking probability obtained through softmax, and $f_{\mathbf{M}}(\cdot)$ and $\Theta_{\mathbf{M}}$ respectively denote the network structure and parameters of RefSel. We employ the online hard example mining cross-entropy loss (OHEM-CE Loss) to measure the discrepancy between network-predicted masks and manually annotated ground truths, thereby driving the model to focus on challenging regions during the learning process. The loss function is defined by,

$$\mathcal{L}_{\text{OHEM-CE}} = -\tfrac{1}{|\mathbf{M}'|}\Sigma_{i\in\mathbf{M}'} m_i \log(p_i), \tag{2}$$

where p_i and m_i are output and label for the position i, and $\mathbf{M}'$ denotes the regions whose loss is larger than a threshold τ (*i.e.*, these regions are more difficult). Moreover, the result is also automatically fine-tuned with the restoration losses Eq. (6) calculated with ground truth.

Reference Transfer in Restoration Model

We build the restoration model upon a one-step diffusion framework based on Arc2Face (Papantoniou et al. 2024). Note that the Arc2Face framework already introduced ArcFace embeddings for the reference for identity preservation, which is insufficient to preserve detailed textures of the reference image. Therefore, we consider referring to ReferenceNet (Hu 2024) to better introduce reference features. However, due to the similarity between different face regions, the cross-attention deployed in ReferenceNet (Hu 2024) cannot effectively match input and reference features, resulting in a trivial solution in the cross-attention module. As shown in Fig. 3, denote the features from the input by Q_{in}, K_{in}, and V_{in}, and features from the reference by K_{ref} and V_{ref}, the vanilla cross-attnetion can be defined by,

$$\sigma(\tfrac{Q_{in}[K_{in},K_{ref}]^\top}{\sqrt{d_k}})[V_{in}, V_{ref}]. \tag{3}$$

It can be seen that when Q_{in} and K_{ref} are difficult to match, $Q_{in}K_{ref}^\top$ can be easily optimized towards $\mathbf{0}$, making the model ignore V_{ref} in the final output.

Decoupled Cross-Attention. To remedy the dilemma, inspired by IP-Adapter (Ye et al. 2023), we leverage the decoupled cross-attention (DCA). Unlike Eq. (3), we process the input and reference features separately, then put them together to force the infusion, *i.e.*,

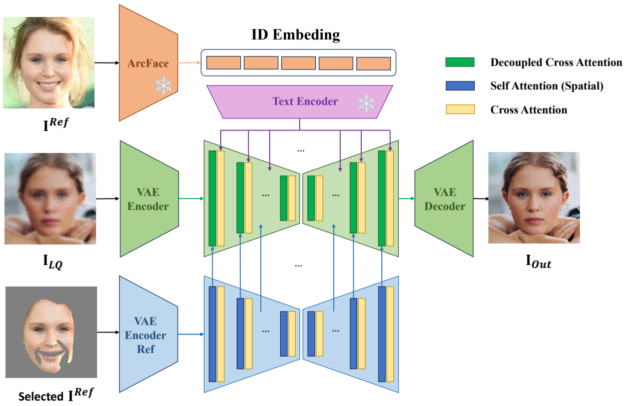

Figure 3: The network structure of our RefSTAR.

$$\sigma(\tfrac{Q_{in}K_{in}^\top}{\sqrt{d_k}})V_{in} + \sigma(\tfrac{Q_{in}K_{ref}^\top}{\sqrt{d_k}})V_{ref}. \tag{4}$$

In this way, the reference features are inevitably injected into the restoration model.

Mask-Compatible Cycle-Consistency Loss

Considering that the restoration is finally supervised by the ground truth $\mathbf{I}$, solely a decoupled cross-attention is less effective. To encourage the reference features to appear in the final output, we deploy a cycle-consistency loss, which in turn takes the degraded reference as new input and the original output as new reference to reconstruct the original reference $\mathbf{I}^{Ref}$. By applying the high-quality reference as the supervision, the detailed textures in the reference are required to be consistent with the original output. Notably, with the mask, not all features of the reference are transferred to the final output. Therefore, the cycle-consistency loss is modulated to be mask-compatible, *i.e.*,

$$\mathcal{L}_{cycle} = \Sigma_i \lambda_i \ell_i(\mathbf{M}\odot\mathbf{I}_{out}^{Ref}, \mathbf{M}\odot\mathbf{I}^{Ref}), \tag{5}$$

where $\odot$ denotes entry-wise multiplication, ℓ_i denotes loss functions, including ℓ_1 loss, perceptual loss and identity loss. λ_1, λ_{perc} and λ_{id} are 1, 0.8 and 1.2 separately. The other loss functions are still calculated with ground truth, which contains fidelity loss ℓ_1, perception loss ℓ_{perc} (Wang et al. 2021b), ID loss ℓ_{id} (Deng et al. 2019) and GAN loss ℓ_{adv} (Goodfellow et al. 2020),

$$\mathcal{L}_{GT} = \Sigma_i \lambda_i \ell_i(\mathbf{I}_{out}, \mathbf{I}), \tag{6}$$

where λ_1, λ_{perc}, λ_{id} and λ_{adv} are 1, 1, 1.2 and 0.2 separately.The total loss is defined by,

$$\mathcal{L} = \lambda_{GT}\mathcal{L}_{GT} + \lambda_{cycle}\mathcal{L}_{cycle}, \tag{7}$$

where λ_{GT} and λ_{cycle} are 1 and 1.5.

Experiments

Training Data. For the restoration model, we follow these reference-based methods and use the CelebRef-HQ (Li et al.

Methods			Celeb-Ref-Test						Real-World-Test		
			PSNR↑	LPIPS↓	FID↓	ID-GT↑	ID-Ref↑	MUSIQ↑	FID↓	ID-Ref ↑	MUSIQ ↑
w/o Reference	GAN Based	CodeFormer	23.81	0.375	26.82	66.38	38.75	74.30	158.55	39.13	73.08
		GFPGAN	23.69	0.390	27.04	68.00	39.28	**74.60**	162.87	42.88	73.54
		GPEN	23.43	0.429	31.12	69.85	40.40	74.26	162.77	45.38	**74.31**
	Diffusion Based	OSEDiff	23.25	0.360	36.99	57.26	33.47	74.36	167.29	37.67	73.45
		DiffBIR	23.99	0.429	33.67	69.89	40.42	74.56	166.95	44.30	72.73
w/ Reference	GAN Based	ASFFNet	23.90	0.368	27.12	67.22	46.74	73.85	173.63	45.74	71.71
		DMDNet	**24.34**	0.382	28.97	72.46	47.61	74.14	162.69	46.03	72.11
	Diffusion Based	FaceMe	23.84	0.417	31.56	62.47	35.73	72.90	158.74	43.37	67.44
		RefLDM	23.73	0.398	24.39	72.77	53.32	73.04	154.94	47.33	71.12
		RestoreID	23.46	0.433	33.08	68.67	46.77	73.31	161.46	46.24	71.10
		Ours	23.99	**0.350**	**22.21**	**78.04**	**64.68**	73.71	**154.51**	**55.73**	72.35

Table 1: Quantitative comparison on synthetic and real-world datasets. The methods are categorized into non-reference and reference-based face restoration for clarity. The best and second-best results are highlighted by **bold** and underline.

Methods	PSNR↑	LPIPS↓	FID↓	ID-GT↑	ID-Ref↑	MUSIQ↑
w/ all-ones mask	23.32	0.379	29.66	70.23	**69.34**	72.87
w/ all-zeros mask	23.88	0.380	25.01	69.24	46.11	**74.02**
w/o cycle loss	23.94	0.369	24.98	67.21	43.64	73.93
w/o DCA	23.79	0.365	25.29	67.09	47.22	73.50
Ours	**23.99**	**0.350**	**22.21**	**78.04**	64.68	73.71

Table 2: Ablation study on RefSel, Mask-Compatible Cycle-Consistency Loss and DCA.

	(1)	(2)	(3)	(4)	(5)	Average
Accuracy	0.90	0.90	0.86	0.94	0.82	0.88

Table 3: The accuracy of our RefSel module on five scenarios: (1) mouth open/close, (2) eyes open/close, (3) conflict wrinkles, (4) conflict beard, and (5) conflict glasses and patterns.

2022) dataset, which consists of 10,555 face images at a resolution of 512×512 across 1,005 identities. The degraded inputs are generated using the two-stage degradation pipeline of Real-ESRGAN (Wang et al. 2021b), with motion and defocus deblurring embedded for a more realistic degradation. The reference images are randomly selected from each identity to enhance generalization on real-world scenarios. For RefSel module, we apply the same degradation pipeline to convert ground-truth images into their corresponding degraded inputs.

Testing Data. For evaluating the restoration model, we select 1,000 identities from the CelebA (Liu et al. 2015) dataset, with one image randomly selected as the reference and another as the ground truth for each identity. We adopt the same degradation synthesis process as Real-ESRGAN (Wang et al. 2021b), with motion and defocus deblurring embedded for a more realistic degradation, to generate the testing data. To evaluate the performance in real-world scenarios, we collect 60 celebrity image pairs from the internet (*i.e.*, RealRef60), each consisting of one high-resolution image as reference and one low-resolution image as degraded input. These pairs are selected to ensure that there is no overlap with the CelebRef-HQ dataset. To assess the performance of RefSel, we use 250 pairs from the RefSel-HQ dataset as the test set.

Implementation Details. The RefSel module employs the AdamW optimizer (Loshchilov and Hutter 2017) with β_1 = 0.5 and β_2 = 0.999, and a learning rate of 1×10^{-3}. The restoration model is trained using a three-phase training strategy. In Phase I, we initialize the model pre-trained weights from Arc2Face (Papantoniou et al. 2024) and adapt it into a one-step diffusion-based image restoration model by training a set of LoRA adapters. In Phase II, we freeze the pretrained RefSel to predict a binary mask. Then we freeze the LoRA parameters and train only the ReferenceNet and our DCA module. Finally, the restoration model will be frozen and the RefSel will be fine-tuned with Eq. (6). All the phases use the AdamW optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, with a learning rate of 1×10^{-4}. All experiments were conducted on a server equipped with two NVIDIA A800 GPUs.

Comparison with Other Methods

To demonstrate the superiority of our RefSTAR, we compare it with several representative blind face restoration methods, including non-reference methods (*i.e.*, GPEN (Yang et al. 2021), GFP-GAN (Wang et al. 2021a), CodeFormer (Zhou et al. 2022), DiffBIR (Lin et al. 2024), and OSEDiff (Wu et al. 2024)) and reference-based methods (*i.e.*, ASFFNet (Li et al. 2020b), DMDNet (Li et al. 2022), RefLDM (Hsiao et al. 2024), FaceMe (Liu et al. 2025), and RestoreID (Ying et al. 2024)), which cover both GAN-based and Diffusion-based models.

Quantitative Comparison. We evaluate the model using PSNR, LPIPS (Zhang et al. 2018), FID (Heusel et al. 2017), and MUSIQ (Ke et al. 2021). To further assess its ability to preserve face identity, we compute identity similarity scores

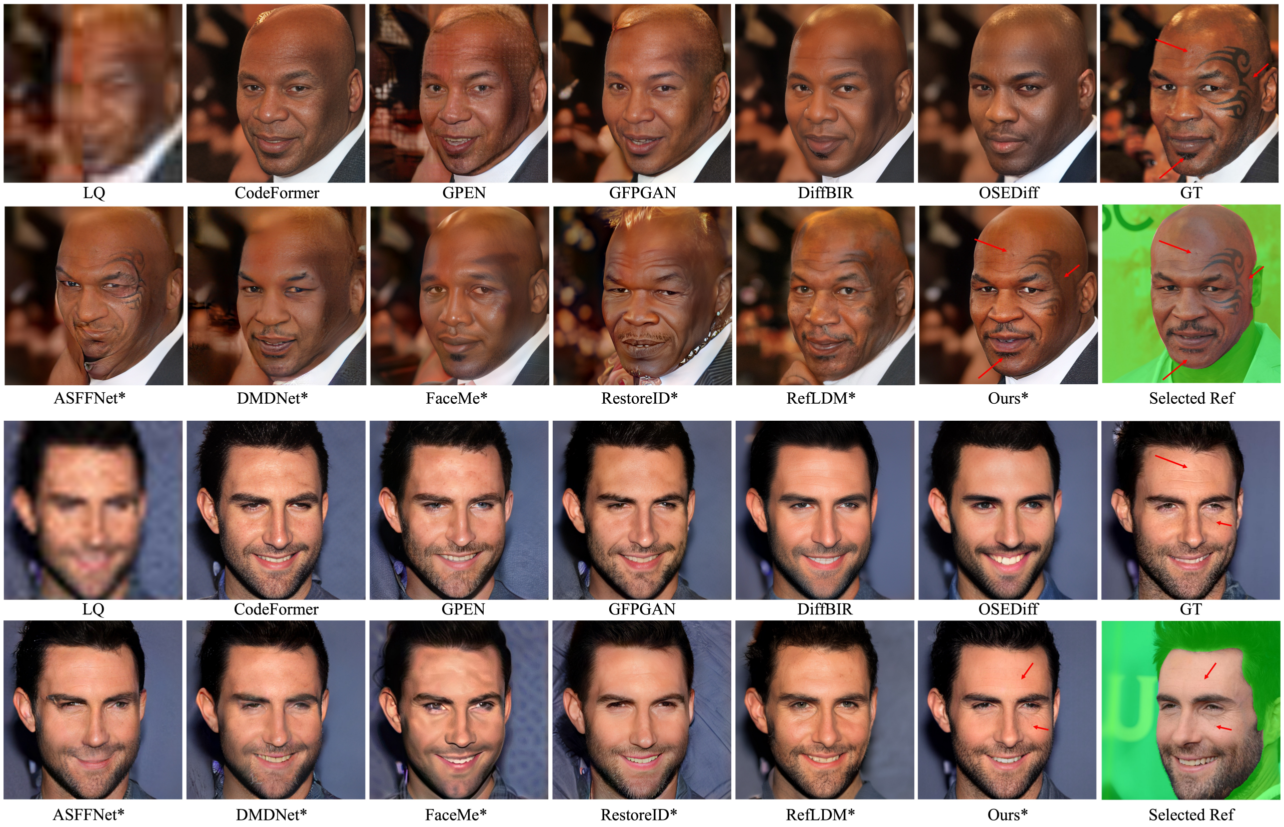

Figure 4: Visual comparison against state-of-the-art non-reference and reference-based face restoration methods on synthesized datasets. The methods with * denote the reference-based face restoration methods.

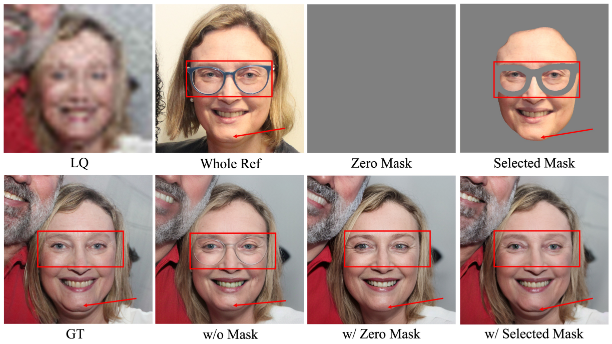

Figure 5: The visual comparison w/ or w/o RefSel.

using the ArcFace model (Deng et al. 2019). We provide two distances: one between the restored images and the ground truth, and the other between the restored images and the reference images, denoted as ID-GT and ID-Ref, respectively. Since the ground-truth image is unavailable for the real-world dataset RealRef60, we only use FID, ID-Ref, and MUSIQ as evaluation metrics. As for FID, it is calculated using ground-truth as a reference set for synthetic datasets, while for real-world datasets, it uses 70,000 high-resolution images from FFHQ as the reference set.

The quantitative results on both synthetic and real-world test sets are summarized in Tab. 1. For the synthetic test set, our approach achieves superior performance across most metrics, including LPIPS, FID, ID-GT, and ID-Ref. Our method performs on par with existing methods in PSNR and MUSIQ, while significantly outperforming them in LPIPS and FID, demonstrating superior visual quality in restored images. Notably, our method achieves the highest ID-GT and ID-Ref scores among all reference-based blind face restoration approaches, demonstrating the effect of our method in transferring reliable reference features. The highest ID-Ref shows that our method takes full advantage of the identity information in reference, while the highest ID-GT indicates that we utilize the correct identity information that is faithful to the ground truth. For the real-world dataset, our method consistently outperforms others in FID and ID-Ref, while achieving comparable MUSIQ scores. This demonstrates its effectiveness in handling complex degradations and highlights its great generalization capability across different real-world scenarios.

To evaluate the performance of RefSel in identifying texture-consistent regions, we categorize the analysis into five distinct scenarios: (1) mouth open/close, (2) eyes open/-close, (3) wrinkles conflicts, (4) beard conflicts, and (5) conflicts with glasses and patterns. Each scenario consists of 50 paired images. The evaluation results for RefSel are shown

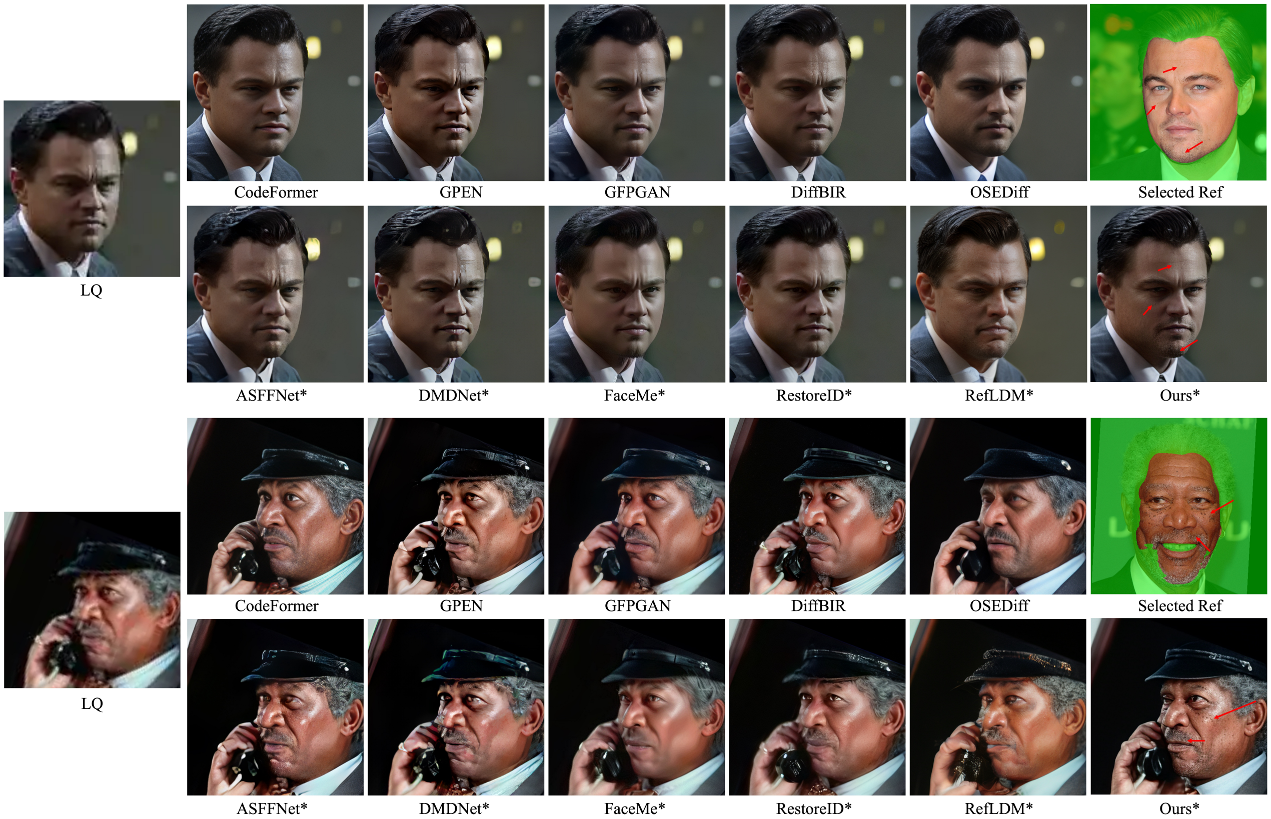

Figure 6: Visual comparison against state-of-the-art non-reference and reference-based face restoration methods on real-world datasets. The methods with * denote the reference-based face restoration methods.

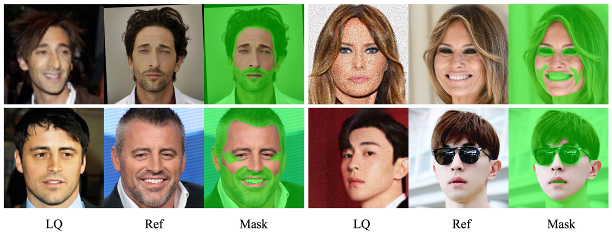

Figure 7: Visualization of RefSel module outputs on real-world low-quality input and high-quality reference pairs.

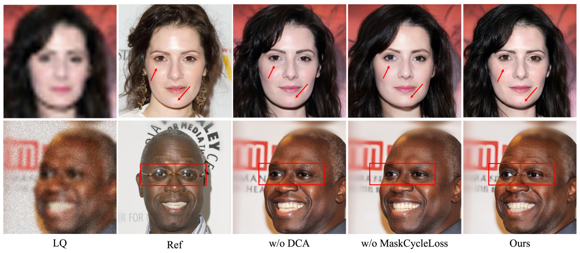

Figure 8: Comparison w/ or w/o DCA and Mask-Compatible Cycle-Consistency Loss.

in Tab. 3. The accuracy demonstrates the reliability and effectiveness of our RefSel module, showcasing its great ability to distinguish regions with consistent texture and effectively identify conflicts between the input and reference.

Qualitative Comparison. Figs. 4 and 6 show the restoration results on synthetic and real-world degraded images, respectively. We can find that our RefSTAR consistently demonstrates more faithful restoration performance. Specifically, for challenging personalized textures such as tattoos, RefSTAR excels in transferring features from the reference image. The framework also accurately replicates common skin textures, including wrinkles and age spots, with this precise texture transfer being crucial for superior identity retention. Notably, our model maintains robust texture transfer even under significant pose variations between the source and reference images, demonstrating the effectiveness of our RefSTAR in unconstrained scenarios.

Ablation Studies

The Effect of RefSel. We analyze the effect by applying different masks (all-zeros or all-ones) to reference images instead of RefSel during training. As shown in Fig. 5, there is an inconsistency in wearing the eyeglasses. With an all-ones mask (using the whole reference image), the unmasked eyeglasses area in the reference image is improperly integrated into the output. When all reference image is ignored by using an all-zeros mask for ReferenceNet, the output relies solely on the highly compressed ID embedding from ArcFace, leading to a loss of personalized textures and a slight decline in identity preservation. The accurate selection by RefSel ensures that only consistent reference features are incorporated into the restored result, preserving wrinkles while masking the glasses. Fig. 7 shows more prediction results on different scenarios. Tab. 2 further demonstrates that the model with the all-ones mask achieves the highest ID-Ref score, as it incorporates the most information from the reference. However, the model with the selected mask achieves the highest ID-GT, demonstrating that our RefSel effectively selects the features most relevant to the input, resulting in more reasonable personalized texture transfer.

The Effect of Decoupled Cross-Attention. Instead of using the cross-attention module with concatenated K and V, we propose a decoupled attention mechanism. Fig. 8 shows that the previous attention mechanism struggles to effectively incorporate reference features into the output, even when Mask-Compatible Cycle-Consistency Loss is applied. This issue arises because image detail restoration requires stricter alignment between the source and reference inputs. As a result, the network behavior becomes biased, and the concatenation of K and V of the input and reference causes the network to ignore the reference, even with the softmax function. Tab. 2 also indicates that it is challenging to properly leverage reference textures without DCA module.

The Effect of Mask-Compatible Cycle-Consistency Loss. To improve the integration of textural details from reference images, we propose a cycle-consistency loss with a masking mechanism during training. To evaluate the effect, we conduct an ablation study by removing it from the pipeline. As shown in Fig. 8, incorporating cycle-consistency loss significantly enhances the model to capture textures from reference images, while also improving identity-aware feature alignment. The cycle-consistency loss constraint ensures that the restored image more faithfully aligns with the reference, enabling better texture transfer and more accurate image restoration. Tab. 2 further validates that this loss function significantly enhances the performance of reference utilization. By ensuring the reconstruction aligns more closely with the reference image, the cycle-consistency loss not only improves texture transfer but also strengthens identity preservation. This results in more accurate feature alignment, effectively bridging the gap between the input and reference and thus boosting overall restoration quality.

Conclusion

In this work, we introduce RefSTAR, a reference-based blind face image restoration framework for effective reference selection, transfer, and reconstruction. Our RefSel module ensures accurate texture consistency selection, while the structured feature fusion mechanism boosts the reference transfer. Additionally, the cycle consistency constraint further enhances identity preservation and restoration fidelity. By incorporating manually annotated consistency regions, our RefSel-HQ dataset emphasizes the critical role of supervised guidance in enhancing reference-based restoration. Experiments demonstrate that RefSTAR outperforms existing methods in maintaining identity-related features and generating high-fidelity restorations.

References

Bulat, A.; and Tzimiropoulos, G. 2018. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *IEEE conference on computer vision and pattern recognition*, 109–117.

Chen, C.; Li, X.; Yang, L.; Lin, X.; Zhang, L.; and Wong, K.-Y. K. 2021. Progressive semantic-aware style transformation for blind face restoration. In *IEEE conference on computer vision and pattern recognition*, 11896–11905.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *IEEE conference on computer vision and pattern recognition*, 2492–2501.

Chong, M. J.; Xu, D.; Zhang, Y.; Wang, Z.; Forsyth, D.; Krishnan, G.; Wu, Y.; and Wang, J. 2025. Copy or Not? Reference-Based Face Image Restoration with Fine Details. In *Winter Conference on Applications of Computer Vision*, 9642–9651.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 139–144.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 6626–6637.

Hsiao, C.-W.; Liu, Y.-L.; Yang, C.-K.; Kuo, S.-P.; Jou, K.; and Chen, C.-P. 2024. ReF-LDM: A Latent Diffusion Model for Reference-based Face Image Restoration. *Advances in Neural Information Processing Systems*, 37: 74840–74867.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*, 5148–5157.

Kim, D.; Kim, M.; Kwon, G.; and Kim, D.-S. 2019. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*.

Li, W.; Wang, M.; Zhang, K.; Li, J.; Li, X.; Zhang, Y.; Gao, G.; Deng, W.; and Lin, C.-W. 2023. Survey on deep face restoration: From non-blind to blind and beyond. *arXiv preprint arXiv:2309.15490*.

Li, X.; Chen, C.; Zhou, S.; Lin, X.; Zuo, W.; and Zhang, L. 2020a. Blind Face Restoration via Deep Multi-scale Component Dictionaries. In *European Conference on Computer Vision*, 399–415.

Li, X.; Li, W.; Ren, D.; Zhang, H.; Wang, M.; and Zuo, W. 2020b. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2706–2715.

Li, X.; Liu, M.; Ye, Y.; Zuo, W.; Lin, L.; and Yang, R. 2018. Learning warped guidance for blind face restoration. In *European Conference on Computer Vision*, 272–289.

Li, X.; Zhang, S.; Zhou, S.; Zhang, L.; and Zuo, W. 2022. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–13.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Qiao, Y.; Ouyang, W.; and Dong, C. 2024. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, 430–448. Springer.

Liu, S.; Duan, Z.-P.; OuYang, J.; Fu, J.; Park, H.; Liu, Z.; Guo, C.-L.; and Li, C. 2025. FaceMe: Robust Blind Face Restoration with Personal Identification. *arXiv preprint arXiv:2501.05177*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision*, 3730–3738.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Papantoniou, F. P.; Lattas, A.; Moschoglou, S.; Deng, J.; Kainz, B.; and Zafeiriou, S. 2024. Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision*, 241–261. Springer.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510.

Sharipov, S.; Nutfullin, B.; and Maloyan, N. 2023. Blind Face Restoration Survey. *International Journal of Open Information Technologies*, 11(6): 11–28.

Shen, Z.; Lai, W.-S.; Xu, T.; Kautz, J.; and Yang, M.-H. 2018. Deep semantic face deblurring. In *IEEE conference on computer vision and pattern recognition*, 8260–8269.

Tao, K.; Gu, J.; Zhang, Y.; Wang, X.; and Cheng, N. 2024. Overcoming False Illusions in Real-World Face Restoration with Multi-Modal Guided Diffusion Model. *arXiv preprint arXiv:2410.04161*.

Wang, J.; Gong, J.; Zhang, L.; Chen, Z.; Liu, X.; Gu, H.; Liu, Y.; Zhang, Y.; and Yang, X. 2024a. OSDFace: One-Step Diffusion Model for Face Restoration. *arXiv preprint arXiv:2411.17163*.

Wang, Q.; Jia, X.; Li, X.; Li, T.; Ma, L.; Zhuge, Y.; and Lu, H. 2024b. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975*.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021a. Towards real-world blind face restoration with generative facial prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9168–9178.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021b. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *IEEE International Conference on Computer Vision*, 1905–1914.

Wang, Z.; Zhang, J.; Chen, R.; Wang, W.; and Luo, P. 2022. RestoreFormer: High-quality blind face restoration from undegraded key-value pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 17491–17500.

Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2024. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553.

Yang, L.; Wang, S.; Ma, S.; Gao, W.; Liu, C.; Wang, P.; and Ren, P. 2020. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of the 28th ACM international conference on multimedia*, 1551–1560.

Yang, P.; Zhou, S.; Tao, Q.; and Loy, C. C. 2023. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances in Neural Information Processing Systems*, 36: 32194–32214.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. Gan prior embedded network for blind face restoration in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 672–681.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arxiv:2308.06721*.

Ying, J.; Liu, M.; Wu, Z.; Zhang, R.; Yu, Z.; Fu, S.; Cao, S.-Y.; Wu, C.; Yu, Y.; and Shen, H.-L. 2024. RestorerID: Towards Tuning-Free Face Restoration with ID Preservation. *arXiv preprint arXiv:2411.14125*.

Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, 217–233.

Yuan, G.; Cun, X.; Zhang, Y.; Li, M.; Qi, C.; Wang, X.; Shan, Y.; and Zheng, H. 2023. Inserting Anybody in Diffusion Models via Celeb Basis. In *NeurIPS*.

Yue, Z.; and Loy, C. C. 2024. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, H.; Alaluf, Y.; Ma, S.; Kadambi, A.; Wang, J.; and Aberman, K. 2024. InstantRestore: Single-Step Personalized Face Restoration with Shared-Image Attention. *arXiv preprint arXiv:2412.06753*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhou, S.; Chan, K. C.; Li, C.; and Loy, C. C. 2022. Towards robust blind face restoration with codebook lookup transformer. *arXiv preprint arXiv:2206.11253*.