2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Semantic-Aware Detail Enhancement for Blind Face Restoration

Huimin Zhao[1], Xiaoqiang Zhou[2], Jie Cao[2], Huaibo Huang[2], Aihua Zheng[1†], Ran He[2,3]
[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]NLPR & CRIPAC, Institute of Automation, Chinese Academy of Sciences, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, China

***Abstract*— The goal of Blind Face Restoration is to recover high-quality images from low-quality images suffering from unknown degradations, posing a significantly challenging problem. In recent years, numerous BFR methods have been proposed, achieving significant success. However, faces possess a unique facial topology, and subtle differences in texture, slight structural imbalances, and minimal asymmetry are easily perceptible in the restored face images. Previous methods often struggle to generate realistically high-quality images from real-world low-quality images and fail to preserve fine features. To more effectively restore image details and textures, providing a more natural and realistic restoration effect, we integrate facial semantic information as prior knowledge into the blind face restoration task. We employ a multi-head cross-attention mechanism to simultaneously consider facial semantic information and context information for modeling. Additionally, we introduce a local detail enhancement module specifically designed to enhance the processing capability of details around the eyes and mouth. Experimental results indicate that our proposed method recovers facial images on synthetic and real datasets more realistically and with higher fidelity.**

I. INTRODUCTION

Blind face restoration aims to recover high-quality face images from low-quality corresponding face images suffering from unknown degradations, including but not limited to low resolution, blurriness, noise, JPEG compression, or their combinations. This technology has potential applications in enhancing facial image quality and is pivotal in various fields. Early methods primarily relied on a substantial collection of low-quality (LQ) and high-quality (HQ) image pairs to learn the mapping from LQ images to HQ images. However, these image pairs often fail to comprehensively cover all types of image degradation encountered in real-world images. Consequently, the restored face images exhibit significant deviations from the ideal high-quality face images. This discrepancy is manifested in the generated facial images appearing unnaturally or not aligning accurately with the facial features of the original people depicted.

Most deep learning-based methods [16], [7] introduce various constraints or priors to mitigate the impact of discrepancies and enhance the quality of restoration. For instance, GFP-GAN [35] and GPEN [39] utilize pre-trained Generative Adversarial Networks (GANs) [10] as decoders to capture facial priors for simultaneous restoration and color enhancement. However, as depicted in Fig. 1, face images generated by these pre-trained GAN models often lack detailed information about facial skin texture, resulting in an overly smooth appearance. Recently, such as DR2 [38] and DifFace [40], have started to leverage the rich image priors and robust generative capabilities of pre-trained diffusion models to strengthen the robustness of blind image restoration. Denoising Diffusion Probabilistic Models (DDPM) [14] refine spatial content during the back-propagation process to enhance the realism of the images. However, these methods lack certain constraints to guide the generation process. Despite these advancements, such methods generally lack effective constraints to guide the generation process, leading to significant deficiencies in the fidelity of the generated images and difficulty in reproducing fine-grained facial details. For instance, as shown in Fig. 1, details like glasses and hair in the facial images generated, are prone to loss.

In this work, to generate realistic that aligned with user preferences while preserving fine-grained identity features, we incorporate face semantic information as a key prior knowledge into the task of blind face restoration. The objective is to transform degraded facial features into another set of features closely resemble real facial features based on the semantic priors. This involves using semantic priors to understand the basic structure of the face, such as the position and shape of the eyes, nose, and mouth. Subsequently, this facial semantic information is utilized as the foundational framework for constructing facial details, guiding the restoration process of facial details such as wrinkles, freckles, and skin textures. This approach not only enhances the details but also renders the face more natural and lifelike. Existing methods based on Vision Transformer (ViT) [23], [30], [17] typically employ multi-head self-attention mechanisms, which evenly focus on the entire input space but may fail to capture the subtle differences between different regions of the input space. This limitation becomes particularly evident in the restoration of facial textures and the reduction of artifacts. To effectively leverage semantic information, we designed a semantic-aware fusion module using a multi-head cross-attention mechanism. In this mechanism, degraded facial features are used as queries, and facial semantic features as key-value pairs. This allows for the capture of correlations between different regions, and enables fusion in both global and local spaces for more precise and targeted attention. This approach enhances the realism and fidelity of key facial region restoration.

In addition, we introduced a local detail enhancement

† Corresponding author.

This work was funded by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-03, the National Natural Science Foundation of China (Grants 62206277 and 61976003) and the Natural Science Foundation of Anhui Province under Grants 2308085Y40.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-9494-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/FG59268.2024.10581942

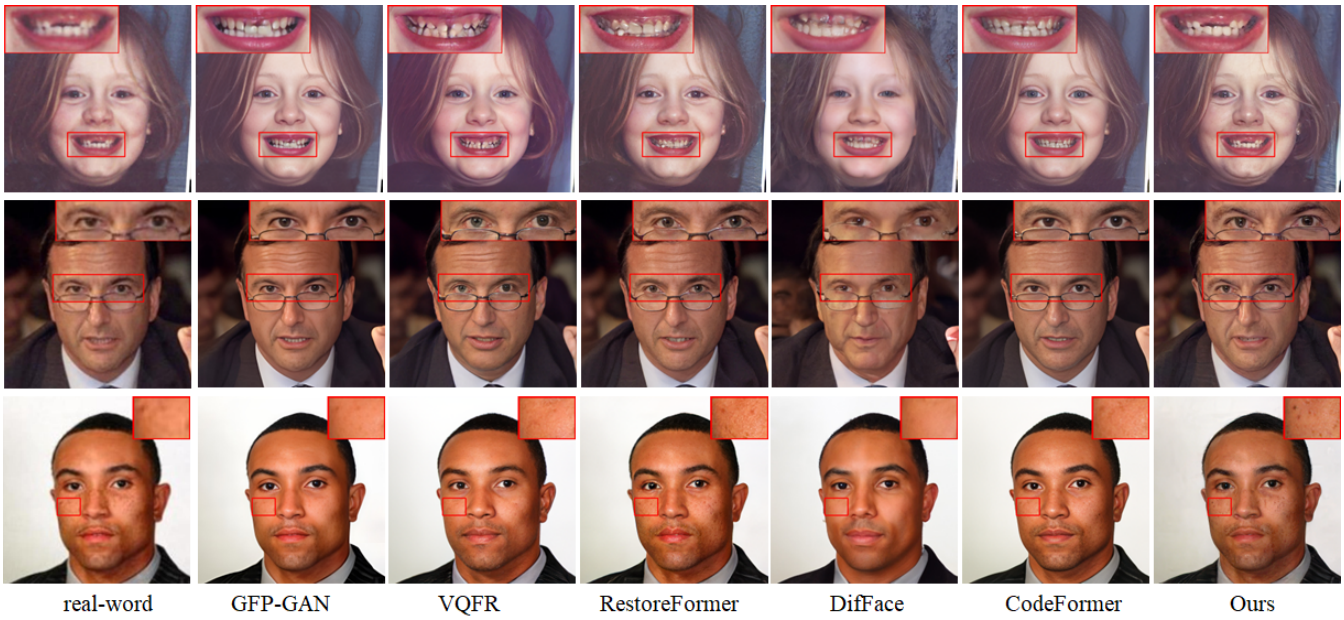


Fig. 1. Comparison of the restoration quality between our method and others on real-world datasets. We proposed a method that can restore high-quality facial details on various facial regions and keep the fidelity, while other methods lack realistic fine details. **Please zoom in for a better view.**

module, specifically designed for processing features at varying scales. This module particularly focuses on enhancing the detailed features of key areas such as the eyes and mouth. To elaborate, we segment the semantic information of the eye and mouth regions and embed it into the decoder. This integration allows for an effective fusion with the degraded facial features at different scales, thereby guiding the semantic repair of local details. Such a process ensures that the results of the repair are semantically consistent with the overall facial context, thereby avoiding unnatural imperfections. Our approach is capable of adapting to features at different scales, making it suitable for both local and global face restoration tasks, enabling the precise capture and restoration of subtle features, ultimately achieving naturalness and realism in facial restoration results. We evaluated the proposed method on real-world and synthetic datasets, and the experimental results demonstrate significant improvements in handling the micro-details and textures of facial skin.

II. RELATED WORK

A. Blind Face Restoration

Blind face restoration aims to recover high-quality faces from images that have undergone unknown and complex degradation. Early attempts utilized DNN-based [43], [16], [22] methods to directly restore high-quality faces from degraded ones. The Dual-Channel Convolutional Neural Network (BCCNN) proposed by Zhou et al. [43] directly maps LQ images to HQ images, utilizing a decoder to reconstruct the HQ face. Cascade Block Network (CBN) [45] adopts a cascaded framework to jointly optimize the performance limitations of previous methods when dealing with misaligned facial images. However, due to the limited information on degraded faces, researchers have begun seeking assistance from other priors.

Deep generative models [10], [14] are popular for their excellent performance in dealing with linear inverse problems such as super-resolution [32], [36], deblurring [21], [29], restoration [37], and colorization [24]. Instead of directly upsampling or reconstructing the input images, some approaches utilize the rich priors encapsulated in generative models, by embedding the pre-trained StyleGAN [1] decoder directly into the BFR network. PULSE [27] employs the latent feature space of pre-trained StyleGAN, identifying the latent vectors most relevant to the low-quality input face images in the feature domain of the pre-trained GAN for self-supervised face restoration. However, searching for the best match image in the latent space of the generative model does not ensure that the restored face will be consistent with the original face content. Then, GPEN [39], GFP-GAN [35], and GLEAN [3] utilize pre-trained GANs to capture facial priors, significantly enhancing restoration performance. These methods leverage the generative priors of GANs to guide the forward process of the network, effectively utilizing the input facial features to enhance the fidelity of the restoration. VQFR [11] also employs pre-trained VQGAN [8] to enhance facial details, significantly reducing uncertainty and blurriness. The main advantage of VQGAN lies in its vector quantization mechanism, allowing precise manipulation of specific features in the generated face images.

Recent works propose utilize the powerful generative priors in diffusion models [14] to address the issue of blind face restoration. DR2 [38] employs a diffusion model to transform degraded images into rough but degradation-invariant predictions, which are then restored to high-quality

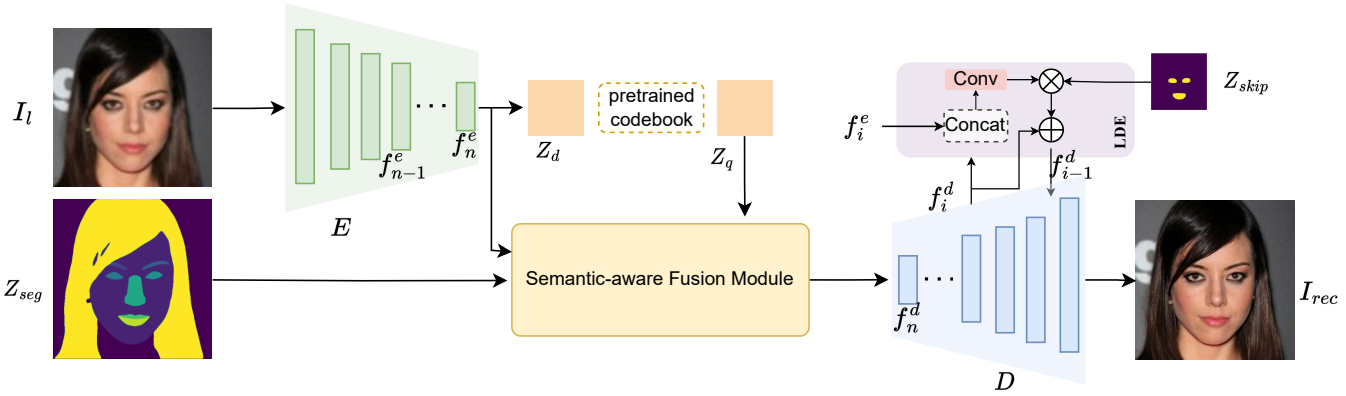


Fig. 2. **Overall framework of the proposed method.** The whole pipeline of our method primarily consists of three components: the encoder $\mathbf{E}$, the decoder $\mathbf{D}$, and the semantic-aware fusion module(SaFM). Given a low-quality image $\mathbf{I_l}$, the process begins with the encoder E extracting facial features $\mathbf{f^e_n}$. Subsequently, in the semantic-aware fusion module, these degraded facial features $\mathbf{f^e_n}$, high-quality facial details $\mathbf{Z_q}$, and facial semantic information $\mathbf{Z_{seg}}$ are integrated through feature fusion. Meanwhile, a local detail enhancement module (LDE) is introduced in the decoder, specifically targeting the enhancement of detail features in key areas such as the eyes and mouth. Finally, the decoder outputs a high-quality image $\mathbf{I_{rec}}$.

images using an enhancement module. Although this approach improves image quality, it falls short in maintaining the authenticity of the images. Subsequently, to balance the inherent realism priors in diffusion models with the fidelity requirements of image restoration tasks, DiffBIR [25] incorporates an additional degradation preprocessing module and a ControlNet-like conditional control module. SR3 [31] and DifFace [40] input low-quality images into a diffusion model to serve as a guiding condition for restoration during training. However, using only degradation information as a constraint still fails to enhance the realism and fidelity of the images. Therefore, this paper proposes the use of facial semantic information as an additional prior to enhance the fidelity of blind face restoration, thereby improving the detail of key facial components.

B. Vision Transformer

Vaswani et al. [34] first proposed the Transformer model, which achieved significant success in the field of natural language processing. Subsequently, the Transformer has been gradually introduced into various visual tasks. The Vision Transformer (ViT) model proposed by Dosovitskiy et al. [6] transforms images into a series of patches. Its core mechanism employs a self-attention mechanism to model the interactions among its inputs, demonstrating outstanding performance in visual tasks. VSR-Transformer [2] applies the self-attention mechanism to super-resolution tasks using a spatio-temporal convolutional self-attention layer that is theoretically comprehensible. HAT [4] combines channel attention with a window-based self-attention scheme, proposing a hybrid attention transformer for image restoration. Zhang et al. [41] first introduce the self-attention mechanism into the task of blind face restoration, achieving a favorable trade-off between quality and fidelity. Recently, the attention mechanism [23], [30] has also proven effective in the field of image restoration. Researchers have focused on integrating the attention mechanism to enhance the processing of critical facial areas. Methods based on attention primarily capture global facial information, resulting in superior performance.

III. METHOD

A. Framework Overview

In this section, we present a detailed description of the proposed model architecture. Our primary objective is to reconstruct high-quality facial images that not only possess lifelike facial details but also maintain the authenticity of the original degraded images. Given an input facial image $\mathbf{I_l}$, whose level of degradation is unknown, our blind restoration approach aims to estimate a high-quality image $\mathbf{I_{rec}}$, which closely approximates the real image $\mathbf{I_l}$ in terms of both authenticity and fidelity.

The overall framework of the restoration process is depicted in Fig. 2, primarily consists of three components: an encoder, a decoder, and a semantic-aware fusion module. Specifically, given a low-quality face image $\mathbf{I_l} \in \mathbb{R}^{\mathbf{H}\times\mathbf{W}\times\mathbf{3}}$ suffering from unknown degradation, where W and H represent the width and height of the image, respectively. Firstly, the encoder E is used to extract facial features from the low-quality image x, resulting in a feature $\mathbf{f^e_n}$, $(\mathbf{n} \in \{\mathbf{1}, \mathbf{2}, \ldots \mathbf{N}\})$, $\mathbf{N}$ denotes the number of scales used for subsequent feature fusion. Then, based on the extracted features $\mathbf{f^e_n}$, pre-trained codebook input features $\mathbf{Z_d}$ are obtained, leading to the generation of vector quantized features $\mathbf{Z_q}$.

After that, feature fusion is conducted in the Semantic Awareness Fusion Module, where degraded facial features $\mathbf{f^e_n}$, facial semantic information $\mathbf{Z_{seg}}$, and high-quality facial details $\mathbf{Z_q}$ are integrated to yield the fused feature $\mathbf{f_m}$. Beyond general reconstruction, our method also incorporates a local detail enhancement module specifically for key facial regions such as the eyes and mouth. This module plays a critical role in refining and accentuating subtle details, thereby significantly improving the overall perceptual quality of the restored facial images. Finally, these features are fed into the decoder $\mathbf{D}$ to reconstruct the high-quality face image $\mathbf{I_{rec}}$.

B. Semantic-aware Fusion Module

When refining image features with semantic priors, it is crucial to consider the differences between the two sources of information. Vision Transformer (ViT) is an effective method for context modeling in computer vision. Most methods based on ViT[23], [30] adopt a self-attention mechanism to capture global information in images. However, the self-attention mechanism, when applied to process a single input sequence, derives Query (Q), Key (K), and Value (V) solely from that sequence. Relying exclusively on degraded facial images as the only information source is insufficient to capture detailed features in images.

To mitigate this issue, in this work, we employ multi-head cross-attention to enhance the detailed facial information modeling process. Q, K and V, derived from different information sources, enable the model to effectively fuse facial semantic information with degraded facial features. This guides the model to more accurately restore the structure and expression characteristics of the face, providing finer repair details. Firstly, we introduce the multi-head cross-attention mechanism used, where $F_1, F_2 \in \mathbb{R}^{H' \times W' \times C}$ represent two different sources of information.

$$Q = F_1 W_q + b_q, K = F_2 W_k + b_k, V = F_2 W_v + b_v, \quad (1)$$

where $W_{q/k/v} \in \mathbb{R}^{C \times C}$, and $b_{q/k/v} \in \mathbb{R}^{C}$ are learnable parameters. Then, divide Q, K, and V into multiple heads along the channel dimension and calculate the attention scores for each head separately:

$$Z_i = \mathbf{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{C_h}}\right) V_i, i = 1, 2, \ldots, N_h, \quad (2)$$

where $C_h = C/N_h$,and then concatenate the outputs of each head to obtain the final output of multi-head attention:

$$Z_{attn} = \mathbf{Concat}\left(Z_1, Z_2, \ldots Z_{N_h}\right), \quad (3)$$

where N_h represents the number of attention heads.

The Semantic-aware Fusion Module aims to effectively integrate semantic information with degraded facial features, thereby enhancing the understanding and reconstruction capabilities of facial structure, texture, and expression details. As depicted in Fig. 3, SaFM receives three distinct inputs: facial semantic information $\mathbf{Z_{seg}}$, degraded facial features $\mathbf{f_n^e}$, and high-quality facial detail $\mathbf{Z_q}$. To fuse these diverse sources of information, we employ three cross-attention mechanisms. Initially, the first cross-attention module merges the degraded facial features $\mathbf{f_n^e}$ with high-quality detail features $\mathbf{Z_q}$, and the intermediate features obtained are further integrated with facial semantic priors $\mathbf{Z_{seg}}$ using the second cross-attention module. Concurrently, the third cross-attention module independently combines the facial semantic priors $\mathbf{Z_{seg}}$ with the degraded features $\mathbf{f_n^e}$. Finally, the fusion features generated from these two steps are summed to produce the ultimate composite feature $\mathbf{f_m}$. This composite effectively amalgamates inputs from varied information sources, ensuring an efficient interchange between semantic priors and degraded features, and thereby more precisely reconstructs and repairs facial details. This depth of fusion strategy significantly enhances the naturalness and accuracy of the restored face image.

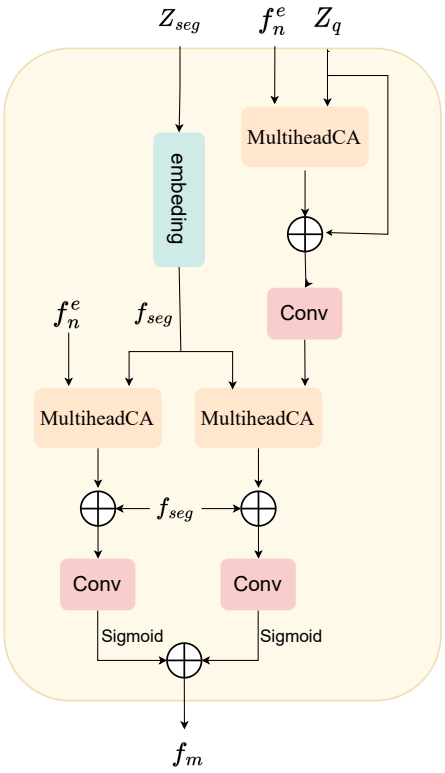


Fig. 3. Semantic-aware fusion module(SaFM) utilizes three multi-head cross-attention mechanisms to fuse the three input features, resulting in the semantically fused feature f_m.

C. Local Detail Enhancement

Although the use of semantic priors yields good results, the eyes and mouth of a human face often contain some detailed information. These details are crucial as they reflect a person's expressions and emotions. To enhance the quality of detail restoration in these areas, we introduce a Local Detail Enhancement module (LDE).

Specifically, we utilize a semantic segmentation map $\mathbf{Z_{skip}}$, based on the eyes and mouth to guide the restoration process, focusing on these areas. Initially, the degraded facial features $\mathbf{f_i^e}$ are concatenated with the current decoder features $\mathbf{f_i^d}$ along the channel dimension. This step is vital for leveraging convolutional operations to further process the concatenated features, thereby enhancing the details of key facial regions. Subsequently, the semantic segmentation map $\mathbf{Z_{skip}}$, which represents the eyes and mouth, is fused with the concatenated features to obtain a detail-enhanced feature map $\mathbf{f_{i-1}^d}$. This feature map is then used in the subsequent decoding process to generate more natural and realistic facial restoration results.

D. Training Objective

We utilize a set of synthesized LQ-HQ face images for training the whole network (the image synthesis process will

be detailed in the next section). Our training objectives consist of pixel reconstruction loss that constrains the restored high-quality face image $\mathbf{I_{rec}}$ to approximate the ground truth $\mathbf{I_h}$, adversarial loss for restoring realistic facial textures and identity loss.

1) **Pixel Reconstruction Loss.** We employ the widely used L1 loss and perceptual Loss [18] in pixel space as the reconstruction loss, denoted as:

$$\mathcal{L}_{pix} = ||I_h - I_{rec}||_1 + \lambda_{pix}||\phi\left(I_h\right) - \phi\left(I_{rec}\right)||_2^2, \quad (4)$$

where λ_{pix} represents the weight of the pixel reconstruction loss, ϕ is the pretrained VGG-19 [33] network and we use the $\{\text{conv1}, \ldots, \text{conv5}\}$ feature maps.

2) **Adversarial Loss.** We employ the adversarial loss to encourage the model to generate realistic textures. Due to the crucial facial components such as eyes and mouth playing significant roles in facial representation, in order to further enhance the perceptually important facial features, our adversarial loss is not only applied to the entire facial image but also separately addressed for the left eye, right eye, and mouth. The overall facial image loss is as follows, where D_d is the discriminator trained on facial images:

$$\mathcal{L}_{adv}^{global} = \lambda_{adv}\left[logD_d\left(I_h\right) + log\left(1 - D_d\left(I_{rec}\right)\right)\right]. \quad (5)$$

The definition of the loss for key facial components is as follows, where the first term is the discriminative loss, and the second term is the feature style loss:

$$\begin{aligned}\mathcal{L}_{adv}^{local} = \lambda_d \sum_r [logD_r\left(R_r\left(I_h\right)\right) + log\left(1 - D_r\left(R_r\left(I_{rec}\right)\right)\right)] \\ + \lambda_s \sum ||Gram\left(\varphi\left(R_r\left(I_h\right)\right)\right) - Gram\left(\varphi\left(R_r\left(I_{rec}\right)\right)\right)||_2^2,\end{aligned} \quad (6)$$

where D_r represents the discriminator for a specific region r of a facial image (left eye, right eye, mouth). The region r is obtained through ROI [12] alignment to acquire R_r. φ denotes the multi-resolution features of the discriminator Dr trained on region r. Gram represents the Gram matrix [9], which calculates feature correlations to measure style differences. λ_d and λ_s respectively signify the loss weights for local discriminative loss and feature style loss.

3) **Identity Preserving Loss.** We draw inspiration from GFP-GAN[35] and apply identity loss in our model to ensure that the restored image aligns with the original in terms of identity features, preventing any deviations from the characteristics of the original image during the restoration process:

$$\mathcal{L}_{id} = \lambda_{id}||\eta\left(I_h\right) - \eta\left(I_{rec}\right)||_2^2, \quad (7)$$

where η denotes the identity feature extracted from ArcFace [5] which is a well-trained face recognition model. λ_{id} denotes the weight of identity preserving loss.

The overall model training objective is a combination of all the loss functions proposed above:

$$\mathcal{L}_{total} = \mathcal{L}_{pix} + \mathcal{L}_{adv}^{global} + \mathcal{L}_{adv}^{local} + \mathcal{L}_{id}. \quad (8)$$

We will provide a detailed explanation of the hyperparameter settings for the loss functions in the next section.

IV. EXPERIMENTS

A. Datasets

Training Dataset. We train our models on the FFHQ [19] dataset, which consists of 70,000 high-quality face images with a resolution of 1024^2. We resize all the images to 512^2 during training, and then synthesize the LQ images following a typical degradation model:

$$I_l = \{JPEG_q\left(\left(I_h * k_\sigma\right) \downarrow r + n\delta\right)\} \uparrow_r \quad (9)$$

where I_l and I_h represent low-quality and high-quality images, respectively, k_σ is a Gaussian kernel with a width of σ, $\downarrow_r$ and $\uparrow_r$ are bicubic down-sampling or up-sampling operators with a given scale factor r, n_δ is additive Gaussian white noise with a standard deviation of σ, and JPEG_q denotes the JPEG compression process with a quality factor of q.

Testing Dataset. We evaluate the model on a synthetic dataset and three distinct real datasets from different sources. None of these datasets overlap with our training dataset. The synthetic dataset is represented as CelebA-Test, consisting of 3000 high-quality (HQ) images sourced from CelebA-HQ dataset [26]. The corresponding low-quality (LQ) images are synthesized using the degradation methods described above. For the real dataset LFW-Test, it comprises the first image of each identity from the original LFW [15] verification partition, totaling 1711 images. CelebChild-Test includes 180 images of celebrity children's faces collected from the internet, and WebPhoto-Test [35] consists of 392 faces from real-life situations

B. Implementation and Evaluation Metric

Implementation Details. In this work, the low-quality face images used are of dimensions $512 \times 512 \times 3$, and the segmentation maps of the generated face images Z_{seg} are of dimensions $512 \times 512 \times 1$. The weight factors for the loss function were set as $\lambda_{pix} = \lambda_d = 1$, $\lambda_s = 2000$, $\lambda_{adv} = 0.8$, and $\lambda_{id} = 3$. We employed the Adam optimizer [20] for training with a learning rate of $4.5e{-}6$, and the entire training process consisted of $60,000$ iterations. The training was conducted on four NVIDIA GeForce RTX 4090 GPUs.

To simulate various degrees of degraded images, for each training image pair, we randomly sample hyperparameters σ, r, δ, and q from the ranges $[0.1, 5]$, $[0.1, 8]$, $[0, 15]$, $[70, 100]$, respectively. Additionally, we introduce color jitter during training to enhance color diversity. Furthermore, to better handle mildly degraded images, we carefully select 10% of high-quality images to serve as inputs for low-quality images.

Metrics. Our evaluation metrics include two widely used non-reference perceptual metrics: Frchet Inception Distance (FID) [13] and Natural Image Quality Evaluator (NIQE) [28], and three widely used reference metrics Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [42]. Specifically, FID measures the Kullback-Leibler divergence between the feature distributions (assumed to be Gaussian) of the restored image and the ground truth image. LPIPS

Fig. 4. Qualitative comparisons on the **Synthetic CelebA-Test** dataset, our restoration results demonstrate enhanced realism and richer details, particularly in critical areas such as hair, mouth, eyes, and skin textures. Our method generates a more natural and lifelike appearance compared to other methods. **Please zoom in for a better view.**

is a learned perceptual similarity metric based on the computation of deep features using VGG [33]. Additionally, we introduce identity distance (IDD) to assess the fidelity of the restored facial images. IDD is the angular distance between the features of the restored facial image and its corresponding ground truth. We employ the pretrained ArcFace [5] face recognition model to extract features.

C. Comparison with State-of-the-art Methods

To validate the effectiveness of our proposed method, we compared its performance with several state-of-the-art face restoration methods, including GFP-GAN[35], VQFR[11], CodeFormer [44], RestoreFormer [37], and GPEN [39].

Synthetic CelebA-Test. The quantitative results of the above-mentioned method compared with our proposed method on the synthetic dataset CelebA-Test are presented in Tab. I. The results in Tab. I demonstrate that our method exhibits significant advantages across multiple metrics. We achieved the lowest LPIPS, indicating that our restoration results are perceptually closer to the real values in terms of perception, demonstrating higher visual similarity. We obtained the lowest IDD and NIQE, indicating that the output is minimally different from the distributions of real and natural face images, resulting in more realistic and fidelity face images. Additionally, our method preserves better identity features.

The visualization results, as shown in the Fig. 4, indicate that our proposed method generates higher quality facial components compared to other methods. The restoration results demonstrate a more natural appearance, especially in the detailing of key areas such as the eyes, mouth, and glasses. This suggests that the exceptional performance of our method in maintaining high fidelity. The faces restored by GFP-GAN [35] are overly smoothed, failing to recover clear facial components and textures, as seen in the excessively smooth skin texture and the loss of details like facial spots in the third row. Although VQFR [11] and RestoreFormer [37] achieve clearer facial images, they fall short in dealing with details in critical facial areas such as the mouth and eyes. As an example, the detail processing of teeth in the first row lacks naturalness, and the wrinkles and texture details around the eyes in the second row are not adequately addressed. While DifFace [40] performs better in terms of FID, resulting in high-quality facial images, its

TABLE I

QUANTITATIVE COMPARISON ON THE CELEBA-TEST DATASET FOR BLIND FACE RESTORATION. OUR METHOD PERFORMS BETTER ON NIQE, LPIPS, AND IDD METRICS, INDICATING THAT OUR RESULTS ARE PERCEPTUALLY CLOSER TO THE ACTUAL VALUES. THE RESTORED FACE IMAGES ARE MORE REALISTIC AND HAVE HIGHER FIDELITY. OUR APPROACH ALSO ACHIEVES BETTER RESULTS ON PSNR AND SSIM.

Methods	PSNR↑	SSIM↑	LPIPS↓	FID↓	NIQE↓	IDD↓
Input	29.15	0.7279	0.3969	94.01	7.933	0.6141
GFP-GAN [35]	27.01	0.6668	0.3076	70.61	4.472	0.3951
CodeFormer [44]	26.67	0.7174	0.3001	56.21	4.765	0.4724
GPEN [39]	**27.14**	**0.7188**	0.333	57.48	4.413	0.3631
VQFR [11]	24.13	0.6662	0.327	49.24	4.189	0.6476
RestoreFormer [37]	25.14	0.6601	0.3289	46.72	4.419	0.4245
DifFace [40]	24.50	0.6699	0.3668	**43.26**	4.226	0.7924
Ours	26.65	0.7156	**0.2914**	43.89	**3.861**	**0.3183**

Fig. 5. Qualitative comparisons on the three real-world datasets, LFW-Test, CelebChild-Test, and WebPhoto-Test, our restoration results demonstrate enhanced realism and richer details, particularly in critical areas such as hair, mouth, and skin textures. Our method generates a more natural and lifelike appearance compared to other methods. **Please zoom in for a better view.**

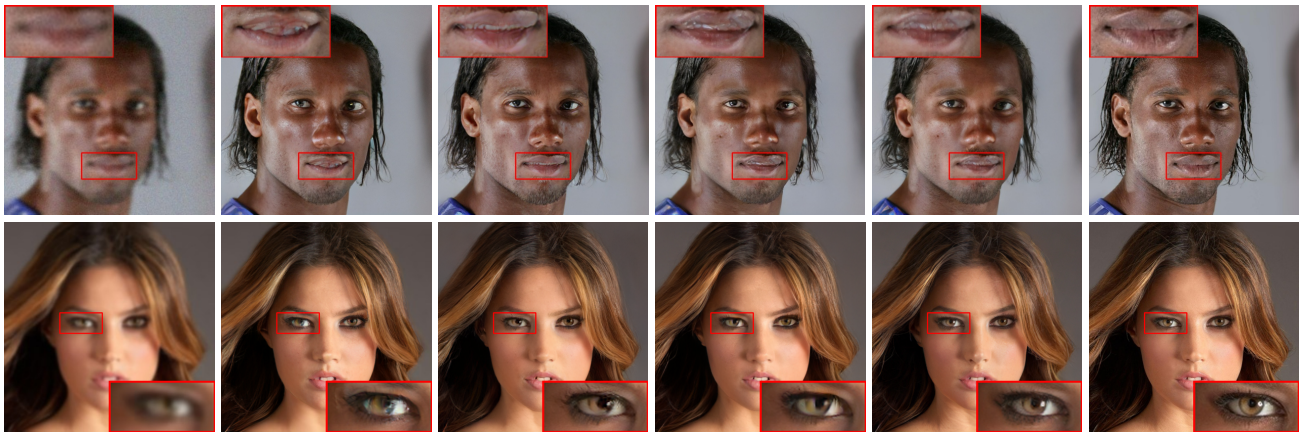
fidelity is significantly reduced. In comparison, our method successfully restores accurate facial expressions and detailed positioning.

Real-world Datasets. Our proposed method and comparative methods were quantitatively evaluated on three real-world test datasets: LFW-Test, CelebChild-Test, and WebPhoto-Test, with the results detailed in Tab. II. Analysis of the data presented in Tab. II reveals that our method demonstrated superior performance in terms of NIQE scores, significantly outperforming the runner-up. The NIQE metric is utilized for assessing the naturalness and overall quality of images, indicating that our method is capable of generating images that are not only more natural but also visually more satisfying.

The visualization results, as shown in Fig. 5, indicate that although most methods can obtain clear faces from slightly degraded damaged face images, the fidelity in key facial areas such as the eyes, mouth, and spotted skin regions is still insufficient. In contrast, the method we propose demonstrates a significant advantage in these critical areas, successfully capturing more details, effectively maintaining the individual's identity features, and producing natural results with rich details. This is attributed to our use of facial semantic priors combined with facial context information.

D. Ablation Study

Based on the description above, our proposed method effectively utilizes facial semantic priors to enhance the quality of facial details in image restoration. To better understand the role of semantic priors in our approach, we designed three variant networks: (1) w/o semantic prior, which refers to a model that does not use semantic priors; (2) w/o-SaFM, indicating the performance of using only the detail enhance-



(a) (b) (c) (d) (e) (f)

Fig. 6. Visual comparison between different variant networks of our method. (a) LQ input; (b) w/o semantic prior; (c) w/o-SaFM; (d) w/o-LDE; (e) Our proposed method; (f) Ground truth. **Please zoom in to see the details.**

TABLE II

QUANTITATIVE COMPARISONS ON THE THREE REAL-WORLD DATASETS, LFW-TEST, CELEBCHILD-TEST, AND WEBPHOTO-TEST. OUR METHOD PERFORMS BETTER IN TERMS OF NIQE SCORES AND ALSO ACHIEVES FAVORABLE RESULTS ON THE FID METRIC.

dataset	LFW-Test		CelebChild-Test		WebPhoto-Test	
Methods	FID↓	NIQE↓	FID↓	NIQE↓	FID↓	NIQE↓
Input	124.97	8.575	144.42	9.028	170.96	12.607
GFP-GAN [35]	67.09	4.649	119.63	4.879	101.52	5.400
CodeFormer [44]	51.94	4.519	116.25	4.981	84.89	4.724
GPEN [39]	51.37	4.548	109.3	4.632	96.90	5.591
VQFR [11]	49.43	3.902	114.71	4.503	86.34	4.728
RestoreFormer [37]	47.75	4.145	**101.18**	4.585	**78.28**	4.466
DifFace [40]	**45.44**	4.224	112.31	4.577	88.52	4.652
Ours	49.51	**3.501**	112.01	**3.991**	83.57	**3.941**

TABLE III

ABLATION STUDIES OF VARIANT NETWORKS.

Configurations	FID↓	NIQE↓	LPIPS↓	PSNR↑	IDD↓
a)w/o semanric piror	46.22	3.948	0.3263	25.23	0.4141
b)w/o-SaFM	44.48	3.928	0.3207	25.72	0.3739
c)w/o-LDE	**43.54**	3.857	0.2995	25.60	0.3631
Ours	43.89	**3.861**	**0.2914**	**26.65**	**0.3183**

ment module; and (3) w/o-LDE, denoting the absence of the local detail enhancement module. Our ablation study conducted on the synthesized celeba-test dataset, we utilized PSNR, FID, LPIPS, NIQE, and IDD as evaluation metrics to demonstrate the effectiveness of our proposed method. The detailed experimental results are listed in Tab. III. It is observable that our method outperforms other variant networks in terms of quantitative measurements.

The visualization results, as shown in Fig. 6,reveal that while the use of semantic priors can generate seemingly clean facial images, there are noticeable unnatural aspects in crucial facial structures. Specifically, in the first row, there's an additional white area in the middle of the mouth, and in the second row, the structure of the eyes has significantly deformed, with the color of the eyeballs shifting and generating unnatural artifacts. This may be attributed to the lack of semantic priors for key facial structures, leading to inaccuracies in capturing and reproducing the true structure of the eyes during the modeling process. Furthermore, the generated eyelashes appear particularly disordered, resulting in a rough and incoherent overall appearance of the eyes. This further confirms the importance of incorporating semantic priors in the complex process of facial feature reconstruction to enhance fidelity. The results from w/o-SaFM and w/o-LDE look slightly better than those without added semantic priors, but unnatural artifacts still occur in the eye area. Overall, our method demonstrates superior performance over its variants, validating the effectiveness of using semantic priors in the BFR task.

V. CONCLUSION

In this work, our objective is to tackle the challenging task of blind face restoration, enhancing the realism and fidelity of facial images in both synthetic and real-world images. Leveraging facial semantic information as semantic priors, we integrate the semantic priors into the facial restoration process using a multi-head cross-attention mechanism. We enhance the details of the eye and mouth regions based on the semantic and structural information in different resolution features, thereby improving the authenticity and fidelity of facial restoration. We conducted extensive experiments on real-world and synthetic datasets, demonstrating the outstanding capability of our approach in handling real-world images, surpassing existing techniques.

REFERENCES

[1] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] J. Cao, Y. Li, K. Zhang, and L. Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[3] K. C. Chan, X. Xu, X. Wang, J. Gu, and C. C. Loy. Glean: Generative latent bank for image super-resolution and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3154–3168, 2022.

[4] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[7] B. Duan, C. Fu, Y. Li, X. Song, and R. He. Cross-spectral face hallucination via disentangling independent factors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7930–7938, 2020.

[8] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[11] Y. Gu, X. Wang, L. Xie, C. Dong, G. Li, Y. Shan, and M.-M. Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[14] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[15] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[16] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017.

[17] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet domain generative adversarial network for multi-scale face hallucination. *International Journal of Computer Vision*, 127(6):763–784, 2019.

[18] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016.

[19] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[21] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019.

[22] Z. Lei, S. Liao, R. He, M. Pietikainen, and S. Z. Li. Gabor volume based local binary pattern for face representation and recognition. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.

[23] G. Li, J. Shi, Y. Zong, F. Wang, T. Wang, and Y. Gong. Learning attention from attention: efficient self-refinement transformer for face super-resolution. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1035–1043, 2023.

[24] J. Li, W. Li, Z. Xu, Y. Wang, and Q. Liu. Wavelet transform-assisted adaptive generative modeling for colorization. *IEEE Transactions on Multimedia*, 2022.

[25] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[27] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020.

[28] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[29] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[30] H. Qi, Y. Qiu, X. Luo, and Z. Jin. An efficient latent style guided transformer-cnn framework for face super-resolution. *IEEE Transactions on Multimedia*, 2023.

[31] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[32] T. R. Shaham, T. Dekel, and T. Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[35] X. Wang, Y. Li, H. Zhang, and Y. Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021.

[36] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision*, 2018.

[37] Z. Wang, J. Zhang, R. Chen, W. Wang, and P. Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022.

[38] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023.

[39] T. Yang, P. Ren, X. Xie, and L. Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021.

[40] Z. Yue and C. C. Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022.

[41] P. Zhang, K. Zhang, W. Luo, C. Li, and G. Wang. Blind face restoration: Benchmark datasets and a baseline model. *Neurocomputing*, page 127271, 2024.

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

[43] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Learning face hallucination in the wild. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[44] S. Zhou, K. Chan, C. Li, and C. C. Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022.

[45] S. Zhu, S. Liu, C. C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 614–630. Springer, 2016.