frontiers
in Neurorobotics

ORIGINAL RESEARCH
published: 04 February 2022
doi: 10.3389/fnbot.2022.797231

Blind Face Restoration *via* Multi-Prior Collaboration and Adaptive Feature Fusion

*Zi Teng, Xiaosheng Yu and Chengdong Wu**

Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China

Blind face restoration (BFR) from severely degraded face images is important in face image processing and has attracted increasing attention due to its wide applications. However, due to the complex unknown degradations in real-world scenarios, existing priors-based methods tend to restore faces with unstable quality. In this article, we propose a multi-prior collaboration network (MPCNet) to seamlessly integrate the advantages of generative priors and face-specific geometry priors. Specifically, we pretrain a high-quality (HQ) face synthesis generative adversarial network (GAN) and a parsing mask prediction network, and then embed them into a U-shaped deep neural network (DNN) as decoder priors to guide face restoration, during which the generative priors can provide adequate details and the parsing map priors provide geometry and semantic information. Furthermore, we design adaptive priors feature fusion (APFF) blocks to incorporate the prior features from pretrained face synthesis GAN and face parsing network in an adaptive and progressive manner, making our MPCNet exhibits good generalization in a real-world application. Experiments demonstrate the superiority of our MPCNet in comparison to state-of-the-arts and also show its potential in handling real-world low-quality (LQ) images from several practical applications.

Keywords: blind face restoration, generative adversarial network, facial priors transformation, multi-prior collaboration, deep neural networks

OPEN ACCESS

Edited by:
Hong Qiao,
University of Chinese Academy of Sciences, China

Reviewed by:
Qi Li,
Institute of Automation, Chinese Academy of Sciences (CAS), China
Huijie Fan,
Shenyang Institute of Automation, Chinese Academy of Sciences (CAS), China

****Correspondence:***
Chengdong Wu
wuchengdong@neu.edu.cn

Received: *18 October 2021*
Accepted: *10 January 2022*
Published: *04 February 2022*

Citation:
Teng Z, Yu X and Wu C (2022) Blind Face Restoration via Multi-Prior Collaboration and Adaptive Feature Fusion.
Front. Neurorobot. 16:797231.
doi: 10.3389/fnbot.2022.797231

1. INTRODUCTION

Face images are always one of the most popular types of images in our daily life, which record long-lasting precious memories and provide crucial information for identity analysis. Unfortunately, due to the limited conditions in the acquisition, storage and transmission devices, the degradations of face images are still ubiquitous in most real-world applications. The degraded face images not only impede human visual perception but also degrade face-related applications such as video surveillance and face recognition. This challenge motivates the restoration of high-quality (HQ) face images from the low-quality (LQ) face inputs which contain unknown degradations (e.g., blur, noise, compression), known as blind face restoration (BFR) (Chen et al., 2021; Wang et al., 2021; Yang et al., 2021). It has attracted increasing attention due to its wide applications.

Face images have face-specific geometry priors which include facial landmarks (Chen et al., 2018), facial parsing maps (Chen et al., 2018, 2021), and facial heatmaps (Yu et al., 2018). Therefore, many recent studies (Shocher et al., 2018; Zhang et al., 2018a, 2020; Soh et al., 2020) exploit extra face prior knowledge as inputs or supervision to recover accurate face shape and details. Benefiting from the incorporation of facial priors in deep neural networks (DNNs), these methods exhibit plausible and acceptable results on bicubic degraded faces. However, when applied to real-world scenarios, they are not applicable due to more complicated degradation. Additionally, the geometry priors estimated from LQ inputs contain very limited texture information for restoring facial details.

Other methods (Li et al., 2018, 2020b) investigate reference priors to generate realistic results. Reference priors can be only one face image, multiple face images, or facial component dictionaries, which can provide many identity-aware face details to the network. Nevertheless, when the identity of LQ is unavailable, the practical applications of referenced-based methods are limited. Additionally, the limited diversity and richness of facial component dictionaries also result in unrealistic restoration results.

Recently, with the rapid development of GAN techniques (Goodfellow et al., 2014), generative priors of pretrained face GAN models, such as StyleGAN (Karras et al., 2019, 2020), are exploited for real-world face restoration (Gu et al., 2020; Menon et al., 2020; Pan et al., 2021). Since face synthesis GANs can generate visually realistic faces with rich and diverse details, it is reasonable to incorporate such generative priors into the face restoration process. These methods first map the LQ input image to an intermediate latent code, which then controls the pretrained GAN at each convolution layer to provide generative priors such as facial textures and colors. This, however, leads to unstable quality of restored faces when dealing with the LQ face image. Due to the low-dimension of latent codes, such a decoupling control method is insufficient to guide the precise restoration process.

Another category of approaches involves performing degradation estimation (Michaeli and Irani, 2013; Bell-Kligler et al., 2019) to provide degradation information for the conditional restoration of LQ face images with unknown degradations. Although this design incorporates human knowledge about the degradation process and implies a certain degree of interpretability, the degradation process in the real world is too complex to be estimated, which fails to bring degradation estimation into full play.

In this article, we investigate the problem of BFR and aim at restoring HQ faces from LQ inputs with complicated degradation. For achieving a better trade-off between realness and fidelity, we propose a multi-prior collaboration network (MPCNet) to seamlessly integrate the advantages of generative priors and face-specific geometry priors. To be specific, we first pretrain an HQ face synthesis GAN and a parsing mask prediction network, and then embed them into a U-shaped DNN as decoder priors to guide face restoration. On the one hand, the encoder part of U-shaped DNN learns to map the LQ input to an intermediate latent space for global face reproduction, which then controls the generator of face synthesis GAN to provide the desired generative priors for HQ face images restoration. On the other hand, the decoder part of U-shaped DNN leverages the encoded intermediate spatial features and diverse facial priors to restore the HQ face in a progressive manner, during which the generative priors can provide adequate details and the parsing map priors provide geometry and semantic information. Instead of direct concatenation, we proposed multi-scale adaptive priors feature fusion (APFF) blocks to incorporate the prior features from pretrained face synthesis GAN and face parsing network in an adaptive and progressive manner. In each APFF block, we integrate generative priors and parsing maps priors with decoded facial features to generate the fusion feature maps for guiding face restoration. In this way, when applying to complicated degradation scenarios, the fusion feature maps can correctly find where to incorporate guidance prior features in an adaptive manner, making our MPCNet exhibits good generalization in a real-world application. The main contributions of this study include:

- We propose a MPCNet to seamlessly integrate the advantages of generative priors and face-specific geometry priors. We pretrain an HQ face synthesis GAN and a parsing mask prediction network, and then embed them into a U-shaped DNN as decoder priors to guide face restoration, during which the generative priors can provide adequate details and the parsing map priors provide geometry and semantic information.
- We propose an APFF block to incorporate the prior features from pretrained face synthesis GAN and face parsing network in an adaptive and progressive manner, making our MPCNet exhibits good generalization in a real-world application.
- Experiments demonstrate the superiority of our MPCNet in comparison to state-of-the-arts, and show its potential in handling real-world LQ images from several practical applications.

2. RELATED STUDY

Facial geometry prior knowledge: Face images have face-specific geometry prior information, which includes 3D facial prior, facial landmarks, face depth map, facial parsing maps, and facial heatmaps. To recover facial images with much clearer facial structure, researchers begin to utilize facial prior knowledge to design the effective face restoration network. Song et al. (2017) proposed to utilize a pre-trained network to extract facial landmarks to divide facial components and feed the five components into different branches to recover corresponding components. Jiang et al. (2018) developed a DNN denoiser and multi-layer neighbor component embedding for face restoration, which first recovered the global face images and then compensated missing details for every component. Wang et al. (2020) proposed the parsing map guided multi-scale attention network to extract the parsing map from LQ and then fed the concatenation of the parsing map and LQ into the subnetworks to produce HQ results. Supposed that the depth map could provide geometric information, Fan et al. (2020) built a subnetwork to learn the depth map from LR and then imported depth into the HQ network to facilitate the facial reconstruction. Benefiting from the incorporation of facial priors in DNNs, these methods exhibit plausible and acceptable results on bicubic degraded faces. However, when applied to real-world scenarios, they are not applicable due to more complicated degradation. Additionally, the geometry priors estimated from LQ inputs contain very limited texture information for restoring facial

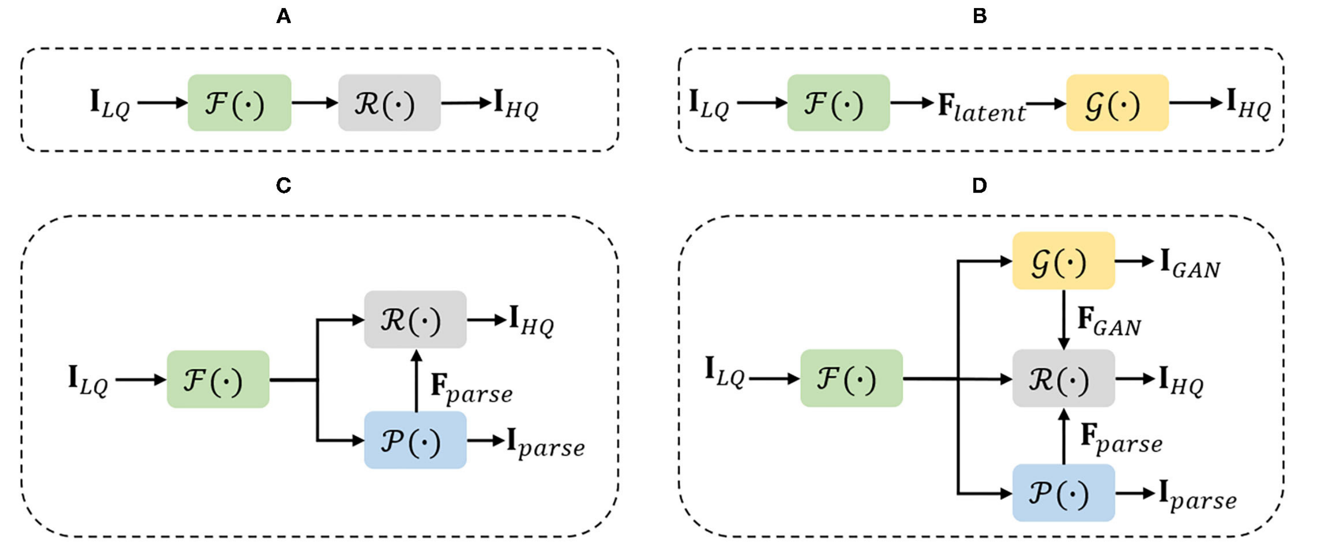


FIGURE 1 | Diagrams of the main ideas of four different paradigms applying for blind face restoration (BFR). In the figures, $\mathcal{F}(\cdot)$, $\mathcal{G}(\cdot)$, $\mathcal{P}(\cdot)$, and $\mathcal{R}(\cdot)$ denote the feature extraction, generative adversarial network (GAN) prior production, Parsing map prior production, and the restoration network, respectively. **(A)** Framework of basic CNN-based methods. **(B)** Framework of GAN prior embedded network. **(C)** Framework of parsing map guided network. **(D)** Framework of our multi-prior collaboration network.

details. Since face synthesis GANs can generate visually realistic faces with rich and diverse details, it is reasonable to incorporate such generative priors into the face restoration process.

Facial generative prior knowledge: Recently, with the rapid development of GAN techniques (Goodfellow et al., 2014), generative priors of pretrained face generative adversarial network (GAN) models, such as StyleGAN (Karras et al., 2019, 2020), are exploited for real-world face restoration (Gu et al., 2020; Menon et al., 2020; Pan et al., 2021). Generative Priors of pretrained GANs (Karras et al., 2017, 2019, 2020; Brock et al., 2018) are previously exploited by GAN inversion (Abdal et al., 2019; Gu et al., 2020; Zhu et al., 2020; Pan et al., 2021), whose primary aim is to map the LQ input image to an intermediate latent code, which then controls the pretrained GAN at each convolution layer to provide generative priors such as facial textures and colors. Yang et al. (2021) proposed to embed the GAN prior learned for face generation into a DNN for face restoration, then jointly fine-tuned the GAN prior network with the DNN. Therefore, the latent code and noise input can be well generated from the degraded face image at different network layers. Wang et al. (2021) proposed to utilize the rich and diverse generative facial priors that contained sufficient facial textures and color information to restore the LQ face images. However, extensive experiments have shown that, due to the low-dimension of latent codes, such decoupling control method is insufficient to guide the precise restoration process and leads to unstable quality of restored faces when dealing with the LQ face image. For achieving a better trade-off between realness and fidelity, we rethink the characteristic of the BFR task and turn to the direction of incorporating various types of facial priors for recovering HQ faces. To that end, we propose a novel multi-prior collaboration framework to seamlessly integrate the advantages of generative priors and face-specific geometry priors, which shows its potential in handling real-world LQ images from several practical applications (see **Figure 1**). For preserving high fidelity, we reform the GAN blocks in StyleGANv2 by removing the noise inputs to avoid the generation of extra stochastic facial details. Then, we design an APFF block to incorporate the prior features from pretrained face synthesis GAN and face parsing network in an adaptive and progressive manner. In general, our main contribution is to explore the solution of the BFR task from a different perspective and provide an effective method that can achieve promising performance on both synthetic and real degraded images.

3. METHODOLOGY

In this section, we first describe the degradation model and our framework in detail, then introduce the adaptive prior features fusion, and finally give the learning objectives used to train the whole network.

3.1. Problem Formulation

To tackle severely degraded faces in real-world scenarios, the training data is synthesized by a complicated degradation that can be formulated as follows:

$$\mathbf{x} = [(\mathbf{y} \circledast \mathbf{k}_{\sigma}) \downarrow_{r} + \mathbf{n}_{\delta}]_{JPEG_{q}} \tag{1}$$

where $\mathbf{x}$ is the LQ face, $\mathbf{y}$ is the HQ face image, $\mathbf{k}_{\sigma}$ is a blur kernel, $\circledast$ denotes convolution operation, $\downarrow_{r}$ represents the standard $\mathbf{r}$-fold downsampler, $\mathbf{n}_{\delta}$ refers to the Gaussian noise with SD δ, and the $JPEG_{q}$ denotes the JPEG compression operator with a quality factor q. In our implementation, for each training pair, we

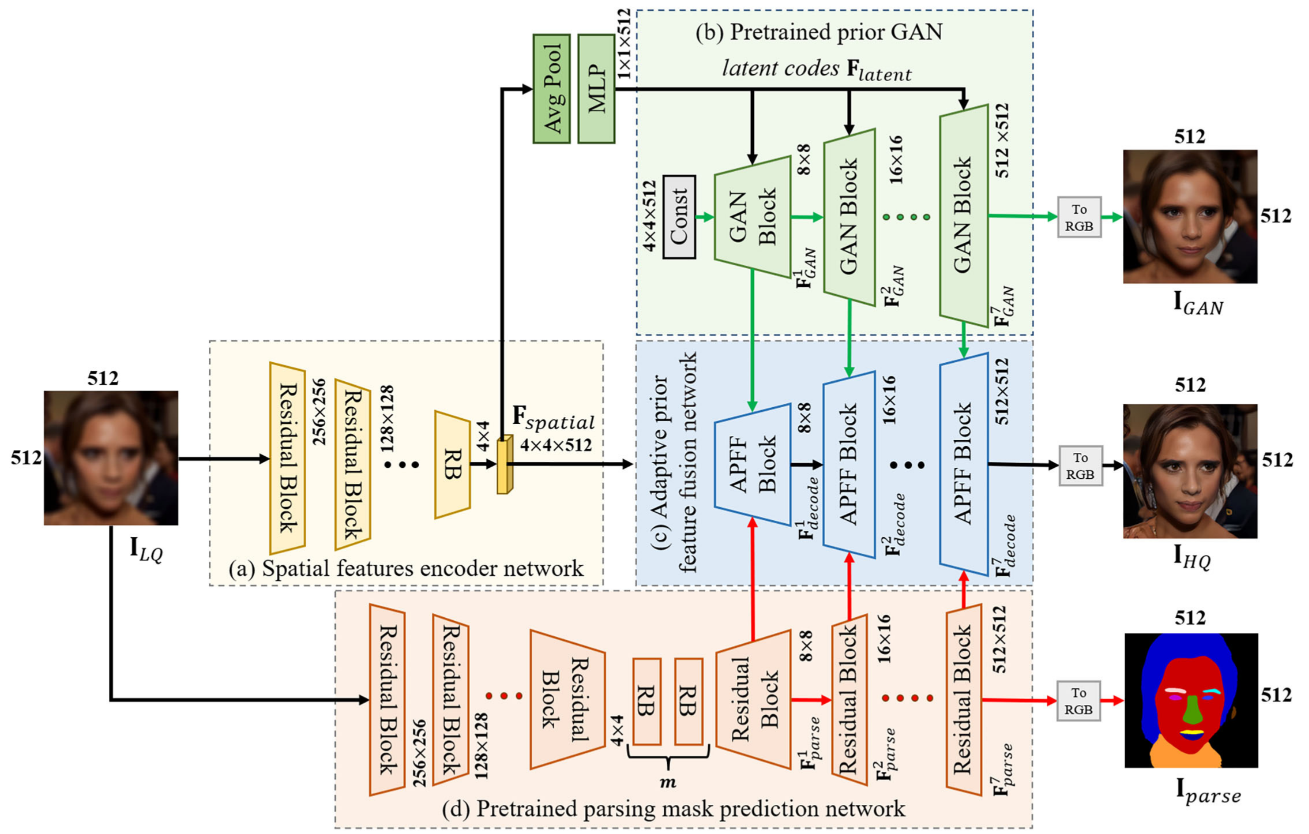


FIGURE 2 | The detailed architecture of the proposed network. **(a)** Spatial feature encoder network. **(b)** Pretrained prior GAn. **(c)** Adaptive prior feature fusion network. **(d)** Pretrained parsing mask prediction network.

randomly select the blur kernel $\mathbf{k}$ from the following four kernels: Gaussian Blur ($3 \leq \sigma \leq 15$), Average Blur ($3 \leq \sigma \leq 15$), Median Blur ($3 \leq \sigma \leq 15$), and Motion Blur ($5 \leq \sigma \leq 25$). The scale factor r is randomly sampled from $[4:16]$. The addictive white gaussian noise (AWGN) $\mathbf{n}_\delta$ is sampled channel-wise from a normal distribution with ($0 \leq \delta \leq 0.1 \times 255$). The compression level q is randomly sampled from $[10:65]$, where higher means stronger compression and lower image quality.

3.2. Overview of MPCNet

To begin with, BFR is defined as the task of reconstructing the HQ face image $\mathbf{y}$ from an LQ input facial image $\mathbf{x}$ suffering from unknown degradation. **Figure 2** illustrates the overall framework of the proposed MPCNet consisting of spatial features encoder network, adaptive prior fusion network, pretrained face synthesis GAN, and pretrained parsing mask prediction network.

3.2.1. U-Shape Backbone Network

The backbone of our MPCNet is composed of the spatial features encoder network and adaptive prior fusion decoder network. It starts with a degraded face image $\mathbf{I}_{LQ}$ of size $512 \times 512 \times 3$. When the input is of a different size, we simply resize it to 512×512 with bicubic sampling. Then, $\mathbf{I}_{LQ}$ goes through several down-sample residual groups to generate an intermediate latent space W which is shared by adaptive prior fusion decoder network and pretrained face synthesis GAN (such as StyleGANv2; Karras et al., 2020). To progressively fuse the decoded spatial features and multiple priors, we present the APFF blocks to construct the decoder part of the U-shape backbone network. The feature $\mathbf{F}^7_{decode}$ from the last APFF block is passed on to a single ToRGB convolution layer and predicts the final output $\mathbf{I}_{HQ}$. More details about the APFF block will be given in the next section.

3.2.2. Pretrained Face Synthesis GAN

Due to the high capability of GANs in generating HQ face images, we leverage pretrained StyleGAN2 prior to providing diverse and rich facial details for our BFR task. To utilize the generative priors, previous methods typically map the input image to its closest latent codes Z and then generate the corresponding output directly. However, due to the low-dimension of latent codes, such decoupling control method is insufficient to guide the precise restoration process and leads to unpredictable failures. Instead of generating the final HQ face image directly, we propose to exploit the intermediate convolutional features of pretrained GAN as priors and further combine them with other types of priors for better fidelity.

$$\mathbf{F}_{latent}, \mathbf{F}_{spatial} = Unet(\mathbf{x}) \tag{2}$$

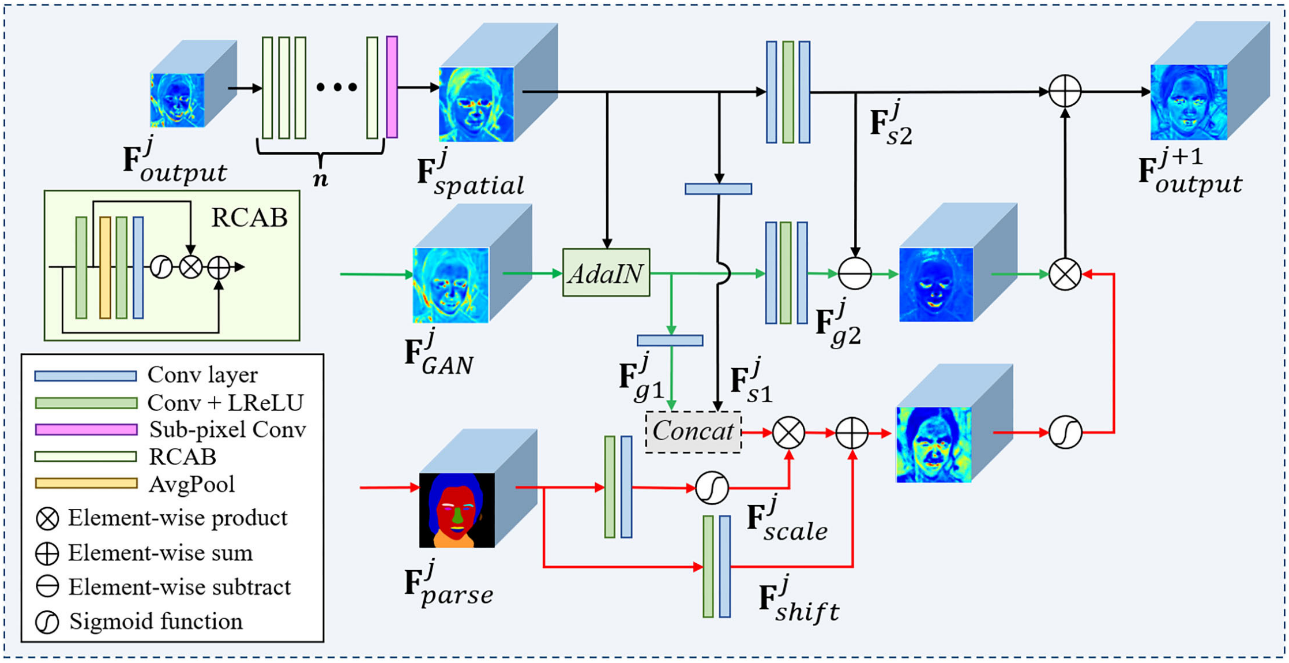


FIGURE 3 | The detailed structures of a GAN block. The definition of "Mod" and "Demod" can be found in Karras et al. (2020).

Specifically, given the encoded intermediate spatial features $\mathbf{F}_{spatial}$ of the input image (produced by the encoder part of the U-shape backbone network, Equation 2), we first map it to the latent codes $\mathbf{F}_{spatial}$ with global pooling operation and several multi-layer perceptron layers (MLP). The latent codes $\mathbf{F}_{latent}$ then pass through each convolution layer in the pretrained GAN and generate GAN features for each resolution scale.

$$\mathbf{F}_{latent} = MLP(\mathbf{F}_{spatial}), \mathbf{F}_{GAN} = StyleGAN(\mathbf{F}_{latent}), \qquad (3)$$

The structure of the GAN block is shown in **Figure 3**, which is consistent with the architecture in StyleGANv2. Additionally, the number of GAN blocks is equal to the number of APFF blocks in the U-shape backbone network, which is related to the resolution of the input face image. For the realness of the synthetic face, the original StyleGANv2 generates stochastic detail by introducing explicit noise inputs. However, the reconstructed HQ face image is required to faithfully approximate the ground-truth face image. For achieving a better trade-off between realness and fidelity, we abandon the noise inputs for all GAN blocks (see **Figure 4**).

3.2.3. Pretrained Parsing Mask Prediction Network

To further improve the fidelity of the restored face image, we pretrain a parsing mask prediction network to provide the geometry and semantic information for covering the deficiencies of GAN priors. As illustrated in **Figure 2D**, since learning the mapping from LQ→parsing maps is much simpler than face restoration, the parsing mask prediction network only employs an encoder-decoder framework. It begins with 7 downsample residual blocks, followed by 10 residual blocks, and 7 upsample residual blocks. The last feature $\mathbf{F}^{7}_{parse}$ is passed on to a single ToRGB convolution layer and predicts the final output $\mathbf{I}_{parse}$. Besides, we conduct extensive experiments to demonstrate the robustness of the parsing mask prediction network on LQ face images with unknown degradations.

3.3. Adaptive Feature Fusion

It is extremely complex to recover HQ faces from the LQ counterparts in real-world scenarios, due to the complicated degradation, diverse poses, and expressions. Therefore, it is natural to consider to combining the different facial priors and let them collaborate to improve the reconstruction quality. Since each facial prior has its shortcomings especially for a specific application, we propose a novel collaboration module that combines multiple facial priors, in which the feature translation, transformation, and fusion are considered for improving the restoration performance and generalization ability of our MPCNet. The APFF block is designed to integrate generative priors $\mathbf{F}^{j}_{GAN}$ and parsing maps priors $\mathbf{F}^{j}_{parse}$ with decoded facial features $\mathbf{F}^{j}_{spatial}$ to generate the fusion feature maps $\mathbf{F}^{j+1}_{output}$ for guiding face restoration. The rich and diverse details provided by $\mathbf{F}^{j}_{GAN}$ can greatly alleviate the difficulty of degradation estimation and image restoration. However, due to the deficiency of the decoupling control method in StyleGANv2, the style condition of $\mathbf{F}^{j}_{GAN}$ is unstable and inconsistent with $\mathbf{F}^{j}_{spatial}$, which should be considered before feature fusion.

AdaIN. AdaIN (Huang and Belongie, 2017) is first proposed to translate the content features to the desired style. Due to its efficiency and compact representation (Karras et al., 2020),

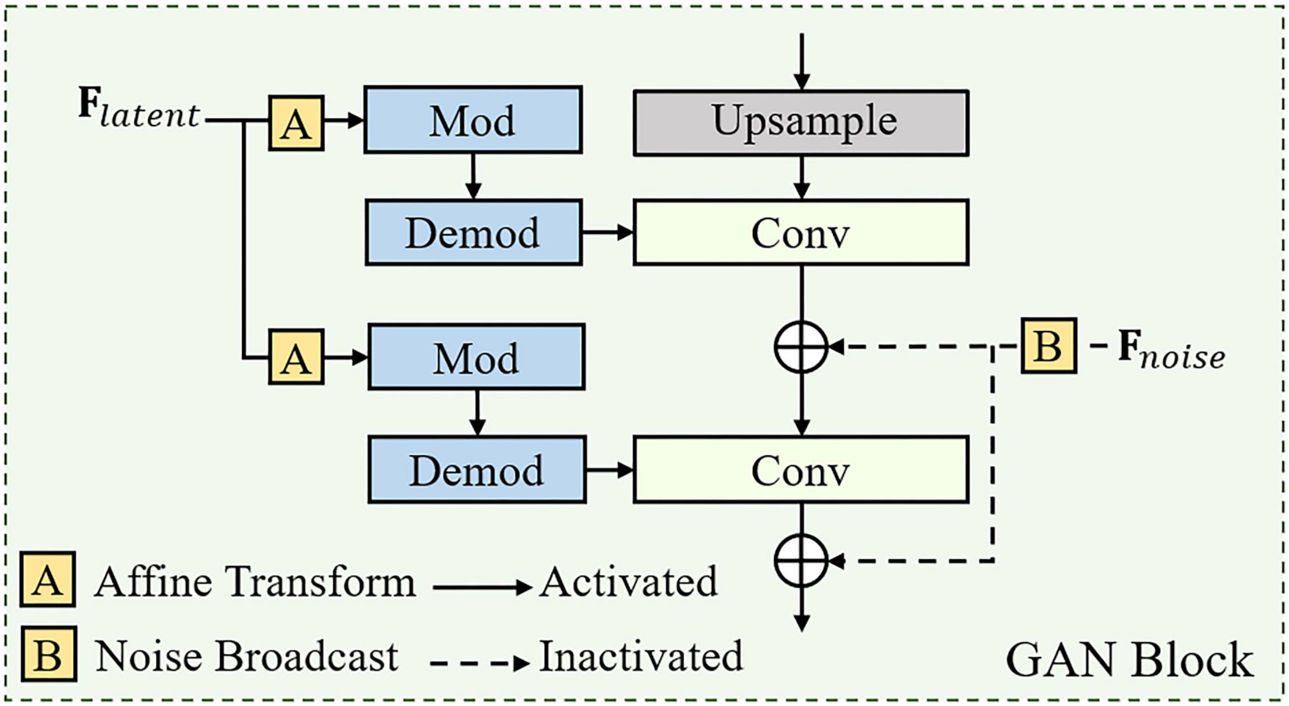


FIGURE 4 | The detailed structures of the adaptive prior feature fusion (APFF) block. The cascading residual channel attention blocks (RCABs) (Zhang et al., 2018c) can make the feature extraction focus on more informative components of the LQ features.

AdaIN is adopted to adjust $\mathbf{F}^j_{GAN}$ to have a similar style condition with the restored feature of degraded image. The AdaIN operation can be formulated as:

$$\begin{aligned} AdaIN(\mathbf{F}^j_{GAN}, \mathbf{F}^j_{spatial}) &= \sigma(\mathbf{F}^j_{spatial})\frac{\mathbf{F}^j_{GAN} - \mu(\mathbf{F}^j_{GAN})}{\sigma(\mathbf{F}^j_{GAN})} \\ &\quad + \mu(\mathbf{F}^j_{spatial}), \\ \mathbf{F}^j_{g1} &= f_{conv1}[AdaIN(\mathbf{F}^j_{GAN}, \mathbf{F}^j_{spatial})], \\ \mathbf{F}^j_{g2} &= f_{conv2}[AdaIN(\mathbf{F}^j_{GAN}, \mathbf{F}^j_{spatial})], \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ denotes the mean operation and $\mu(\cdot)$ denotes the SD operation. With AdaIN operation, $\mathbf{F}^j_{GAN}$ can, thus, be aligned with $\mathbf{F}^j_{spatial}$ by style condition such as color, contrast, and illumination. Intermediate generative features $\mathbf{F}^j_{g1}$ and $\mathbf{F}^j_{g2}$ are generated by $f_{conv1}(\cdot)$ and $f_{conv2}(\cdot)$ which denote 3 × 3 convolutions and are exploited to reduce the channel numbers and refine features, respectively. Besides, the intermediate spatial features $\mathbf{F}^j_{s1}$ and $\mathbf{F}^j_{s2}$ are also generated from $\mathbf{F}^j_{spatial}$ by the same process.

Spatial feature transform. Motivated by the observation that GAN priors are incapable to capture the geometry information of the overall face structure due to the decoupling control method, we propose to exploit the parsing map prior to providing the geometry and semantic information for covering the shortage of GAN priors. Specifically, we introduce the guidance features $\mathbf{F}^j_{guide}$ to direct the fusion process of $\mathbf{F}^j_{GAN}$ and $\mathbf{F}^j_{spatial}$. Additionally, the generation of $\mathbf{F}^j_{guide}$ considers the $\mathbf{F}^j_{GAN}$, $\mathbf{F}^j_{spatial}$, and $\mathbf{F}^j_{parse}$. For spatial-wise feature modulation, we employ Spatial Feature Transform (SFT), named $SFT(\cdot)$, Wang et al. (2018b) to generate the affine transformation parameters with $\mathbf{F}^j_{parse}$. At each resolution scale, the $SFT(\cdot)$ learns a mapping function $f(\cdot)$ that provides a modulation parameter pair $\boldsymbol{\alpha}, \boldsymbol{\beta}$ according to the parsing maps $\mathbf{F}^j_{parse}$, and then utilities $\boldsymbol{\alpha}, \boldsymbol{\beta}$ to provide spatially fine-grained control to the concatenation of $\mathbf{F}^j_{GAN}$ and $\mathbf{F}^j_{spatial}$.

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{F}^j_{parse}), \tag{5}$$

The concatenation of $\mathbf{F}^j_{GAN}$ and $\mathbf{F}^j_{spatial}$ is modified by scaling and shifting feature maps according to the transformation parameters:

$$\mathbf{F}^j_{guide} = SFT(Concat[\mathbf{F}^j_{g1}, \mathbf{F}^j_{s1}] \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha} \otimes Concat[\mathbf{F}^j_{g1}, \mathbf{F}^j_{s1}] + \boldsymbol{\beta}, \tag{6}$$

where $Concat[;]$ denotes the concatenation operation and $Concat[\mathbf{F}^j_{g1}, \mathbf{F}^j_{s1}]$ denotes the concatenated feature maps, which have the same dimension with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and $\otimes$ indicates element-wise multiplication.

On the one hand, the facial generative priors generally contain HQ facial texture details. On the other hand, the facial parse priors have more shape and semantic information and, thus, are

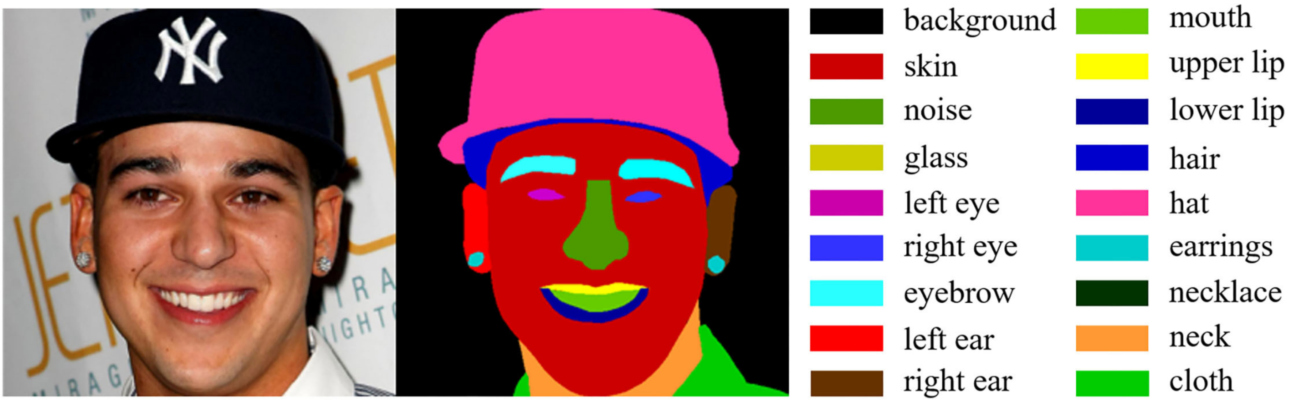


FIGURE 5 | A visual example of facial parsing map information.

more reliable for the global facial region. Considering that $\mathbf{F}^j_{GAN}$ and $\mathbf{F}^j_{parse}$ can mutually convey complementary information for each other, we combine them for better reconstruction of the HQ face image. We first calculate the errors between generative features and spatial features to highlight the inconsistent facial components that need correction. Then we exploit a gating module $softmax(\cdot)$ to generate the semantic-guided map from parse features. Finally, we combine the semantic-guided maps and the feature of inconsistent facial components to refine the initial spatial features in early layers for obtaining better results. The output of each APFF block can be written as,

$$\mathbf{F}^{j+1}_{output} = (\mathbf{F}^j_{g2} - \mathbf{F}^j_{s2}) \otimes softmax(\mathbf{F}^j_{guide}) + \mathbf{F}^j_{s2} \qquad (7)$$

As a result, this helps to make full use of the rich and diverse texture information from $\mathbf{F}^j_{GAN}$ as well as shape and semantic guidance from $\mathbf{F}^j_{parse}$ in an adaptive manner, thereby achieving a good balance between realness and faithfulness. Besides, we conduct APFF block at each resolution scale to facilitate progressive fusion and finally generate the restored face. In this way, when applying to complicated degradation scenarios, the fusion feature maps can correctly find where to incorporate guidance prior features in an adaptive manner, making our MPCNet exhibits good generalization in a real-world application.

3.4. Learning Objective

For achieving a better trade-off between realness and fidelity, following previous BFR methods (Chen et al., 2018; Wang et al., 2018a,c; Li et al., 2020a,b), we apply 1) reconstruction loss that constrains the outputs to faithfully approximate to the ground-truth face image, 2) adversarial loss that generates the visually realistic details for the photo-realistic face restoration, and 3) gram matrix loss that helps in better synthesize texture details.

Reconstruction loss. We combine the pixel and feature space mean square error (MSE) to constrain the network output $\hat{\mathbf{I}}_{HQ}$ close to the ground truth $\mathbf{I}_{HQ}$. As shown in follows, the second term is perceptual loss (Yu and Porikli, 2017; Wang et al., 2018b):

$$\mathcal{L}_{rec} = \lambda_{MSE} \parallel \mathbf{I}_{HQ} - \hat{\mathbf{I}}_{HQ} \parallel_1 + \lambda_{perc} \sum_{i=1}^{4} \parallel \varphi_i(\mathbf{I}_{HQ}) - \varphi_i(\hat{\mathbf{I}}_{HQ}) \parallel_1, \qquad (8)$$

where $\varphi_i(\cdot)$ represents the features from the i-th layer of the pretrained VGGFace model (Cao et al., 2018). λ_{MSE} and λ_{perc} denote the trade-off loss weights parameters. In this study, we set $i \in [1, 2, 3, 4]$.

Adversarial loss. Adversarial loss has been proved to be an effective and critical method in improving visual quality. In both generator and discriminator, we incorporate spectral normalization (Miyato et al., 2018) on the weights of each convolution layer to stabilize the learning. Furthermore, we adopt the hinge version of adversarial loss as the objective function (Brock et al., 2018; Zhang et al., 2019), defined as:

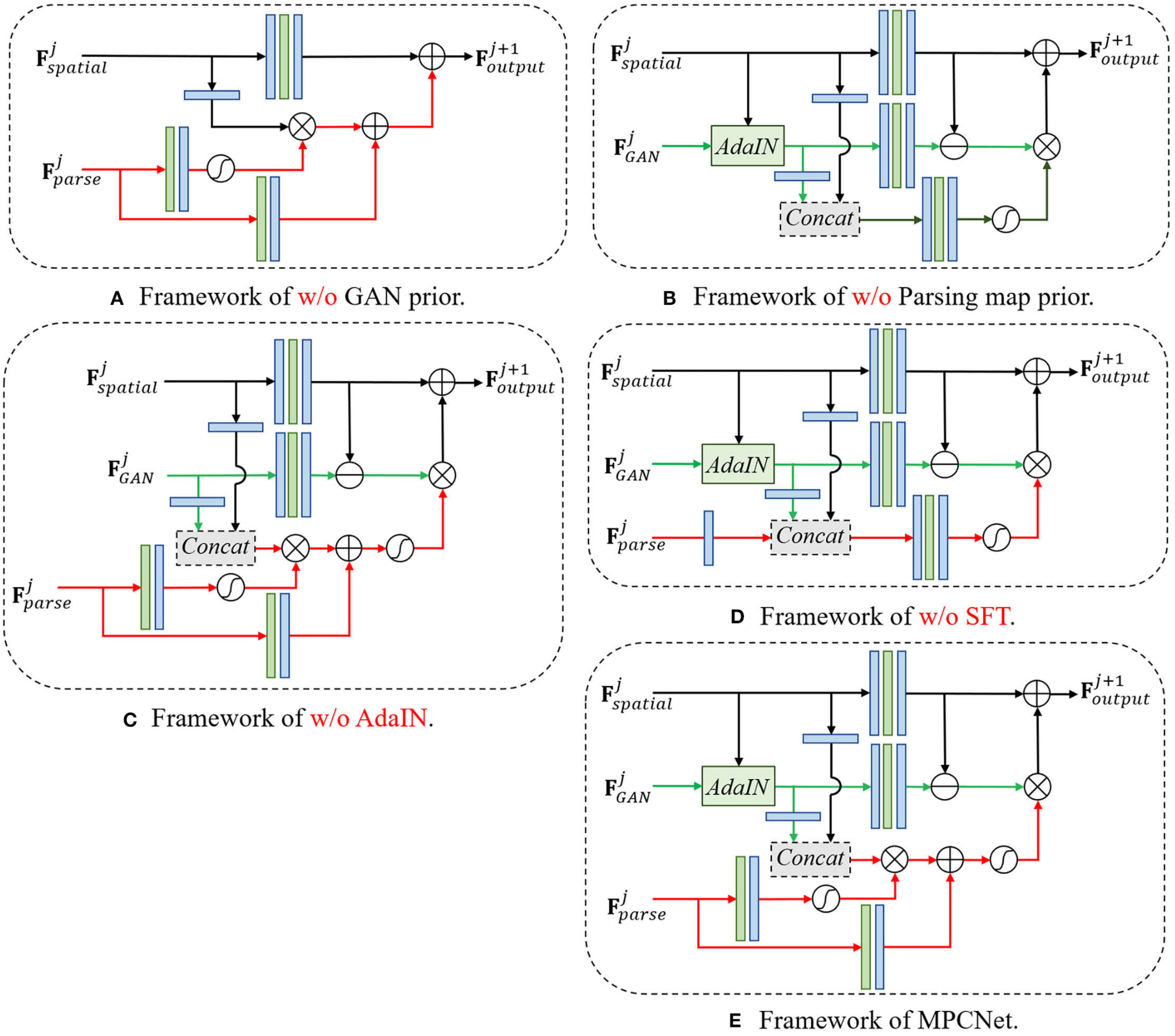
$$\begin{aligned}\mathcal{L}_{adv,D} &= \mathbb{E}[max(0, 1 - D(\mathbf{I}_{HQ}))] + \mathbb{E}[max(0, 1 + D(\hat{\mathbf{I}}_{HQ}))], \\ \mathcal{L}_{adv,G} &= -\mathbb{E}[D(\hat{\mathbf{I}}_{HQ})]\end{aligned} \qquad (9)$$

In this study, $\mathcal{L}_{adv,D}$ is used to update the discriminator, while $\mathcal{L}_{adv,G}$ is adopted to update the MPCNet for blind face restoration.

Gram matrix loss. Gram matrix loss (Gatys et al., 2016) has demonstrated that style transfer helps a lot in synthesizing visually plausible textures. We use pretrained VGGFace (Cao et al., 2018) features of layer relu2_1, relu3_1, relu4_1, and relu5_1 to calculate gram matrix loss, which is formulated as:

$$\mathcal{L}_{style} = \sum_{i=1}^{4} \frac{\parallel \varphi_i(\mathbf{I}_{HQ})^T \varphi_i(\mathbf{I}_{HQ}) - \varphi_i(\hat{\mathbf{I}}_{HQ})^T \varphi_i(\hat{\mathbf{I}}_{HQ}) \parallel^2}{C_i H_i W_i}, \qquad (10)$$

where $\varphi_i(\cdot)$ represents the features from the i-th layer of the pretrained VGGFace model.



A Framework of w/o GAN prior.

B Framework of w/o Parsing map prior.

C Framework of w/o AdaIN.

D Framework of w/o SFT.

E Framework of MPCNet.

FIGURE 6 | Structure diagrams of different network versions used in the ablation study. **(A)** Framework of w/o GAN prior. **(B)** Framework of w/o parsing map prior. **(C)** Framework of w/o AdaIN. **(D)** Framework of w/o SFT. **(E)** Framework of w/o MPCNet.

4. EXPERIMENT RESULTS

4.1. Dataset and Experimental Settings

Training datasets. We first adopt the CelebA-Mask-HQ (Lee et al., 2020) to pre-train the face parsing mask prediction network, which contains 30,000 HQ face images with a size of $1,024 \times 1,024$ pixels. As shown in **Figure 5**, each image of CelebA-Mask-HQ has a segmentation mask of facial attributes. To build the training set, we randomly choose 24,000 HQ images and resize all images to 512×512 pixels as ground-truth. Similar to Li et al. (2020a), we adopt the degradation model in section Problem formulation with randomly sampled parameters to synthesize the corresponding LQ images. Then we adopt the FFHQ dataset (Karras et al., 2019) to train the GAN prior network and the final MPCNet. FFHQ dataset contains 70,000 HQ face images with a size of $1,024 \times 1,024$ pixels. In the same way as CelebA-Mask-HQ, we synthesize the LQ inputs with Equation (1) during training.

Testing datasets. We construct one synthetic test dataset and one real-world LQ test dataset to validate the ability of the proposed method on handling the BFR. Additionally, all these test datasets have no overlap with the training datasets. For the synthetic test dataset, we first randomly choose 3,000 HQ images from the CelebA-HQ dataset (Karras et al., 2017). Then the generation way of testing pairs is the same as the training dataset, namely CelebA-Test. For the real LQ test dataset, we collect 1,000

LQ faces from CelebA (Liu et al., 2015) and 500 old photos from the web. We coarsely crop square regions in each image according to their face regions and resize them to 512 × 512 pixels using bicubic upsampling. In the end, we put all these images together and generate the real LQ test dataset containing 1,500 real LQ faces, namely **Real-Test**.

Implementation. We adopt Adam optimizer (Kingma and Ba, 2014) with $\delta 1 = 0.9$, $\delta 2 = 0.99$, and $\varepsilon = 10^{-8}$ to train our MPCNet with a batch size of 8. During training, we augment the training images by randomly horizontally flipping. The learning rate is initialized as $2 * 10^{-4}$ and then decreased to half when the reconstruction loss is no longer dropping on the validation set. Our proposed model is implemented on the Pytorch framework using two NVIDIA RTX 2080Ti GPUs.

4.2. Evaluation Index

For synthetic test datasets with ground truth, two widely used image quality assessment indexes, peak signal-to-noise ratio (PSNR) (Hore and Ziou, 2010) and structural similarity (SSIM) (Wang et al., 2004), are used as the criteria for evaluating the performance of models, which are defined as follows:

$$MSE(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{11}$$

where **x** is the target image; **y** is the HQ image which is generated from the LQ image; x_i and y_i represent the values of $i - th$ pixel in **x** and **y**, respectively, and n denotes the pixel number in the image. Then we calculate the PSNR as follows:

$$PSNR(\mathbf{x}, \mathbf{y}) = 10 \cdot \log_{10} \frac{MAX^2}{MSE(\mathbf{x}, \mathbf{y})} \tag{12}$$

where MAX denotes the maximum possible pixel value of the image. It is set to 255 in our experiments since the pixels of the images are represented using 8 bits per sample. PSNR is used to evaluate the performance of the proposed method in reconstructing HQ images. Instead of measuring the error between the ground-truth HQ image and the reconstructed HQ image, Wang et al. (2004) proposed an image quality assessment metric called SSIM to compute the SSIM of two images, and the SSIM value of the reconstructed HQ image **y** is computed as follows:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1)(2\sigma_{\mathbf{xy}} + C_2)}{(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C_1)(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + C_2)} \tag{13}$$

where $\mu_{\mathbf{x}}$, $\mu_{\mathbf{y}}$, $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$, and $\sigma_{\mathbf{xy}}$ represent the local means, SDs, and cross-covariance for images **x** and **y**, respectively. $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ are variables to stabilize the division with a weak denominator, where L is the dynamic range of the pixel values that are set to 255 and k_1 and k_2 are set to 0.01 and 0.03 in our experiments.

Besides, since pixel space metrics are only based on local distortion measurement and inconsistent with human perception, the Learned Perceptual Image Patch Similarity (LPIPS) score (Zhang et al., 2018b) is adopted to evaluate the perceptual realism of generated faces. For a real LQ test dataset without ground truth, the widely-used non-reference perceptual metrics: Fréchet Inception Distances (FID) (Heusel et al., 2017) is used as the criteria for evaluating the performance of the models. We choose 3,000 HQ images from the CelebA-HQ dataset as the reference dataset to evaluate the results of the real LQ test dataset.

TABLE 1 | Summary of model characteristics presented in the ablation study.

Configuration	GAN prior	AdaIN	Parsing map prior	SFT
w/o GAN prior	×	×	√	√
w/o Parsing map prior	√	√	×	×
w/o AdaIN	√	×	√	√
w/o SFT	√	√	√	×
MPCNet (ours)	√	√	√	√

4.3. Ablation Study

We further conduct an ablation study to verify the superiority of our multi-prior collaboration framework (see **Figure 6**). To demonstrate the superiority of our prior-integration method, we remove used modules separately and visualize some comparison results of different variants. The characteristics of different model variants used in the ablation study are summarized in **Table 1**.

Pretrained GAN prior: *w/o GAN prior* denotes the basic model that consists of the decoder part of U-shaped DNN which leverages the encoded intermediate spatial features and parsing map prior priors to restore the HQ face, during which the generative priors are abandoned. This model is in essence equivalent to a parsing map priors guided face restoration network and is included here to demonstrate the importance of generative priors. As the comparison between MPCNet and w/o GAN prior shown in **Figure 7** and **Table 2**, it is evident that the GAN priors can provide diverse and rich facial details for our BFR task.

Pretrained parsing map prior: *w/o Parsing map prior* denotes the model that consists of the decoder part of U-shaped DNN which leverages the encoded intermediate spatial features and generative priors to restore the HQ face, during which the parsing map prior are abandoned. This model is in essence equivalent to a generative priors guided face restoration network and is included here to demonstrate the importance of parsing map priors. As the comparison between MPCNet and w/o Parsing map prior shown in **Figure 7** and **Table 2**, it is evident that the Parsing map priors can provide the geometry and semantic information for covering the shortage of GAN priors and further improve the fidelity of restored face image.

AdaIN: *w/o AdaIN* denotes the model that consists of the decoder part of U-shaped DNN which leverages the encoded intermediate spatial features with types of facial priors to restore the HQ face, during which the AdaIN is abandoned. This model is included here to demonstrate the importance of AdaIN. As the comparison between MPCNet and w/o AdaIN shown in **Figure 7** and **Table 2**, it is evident that the AdaIN module can translate the content features to the desired style with effect and, thus, makes

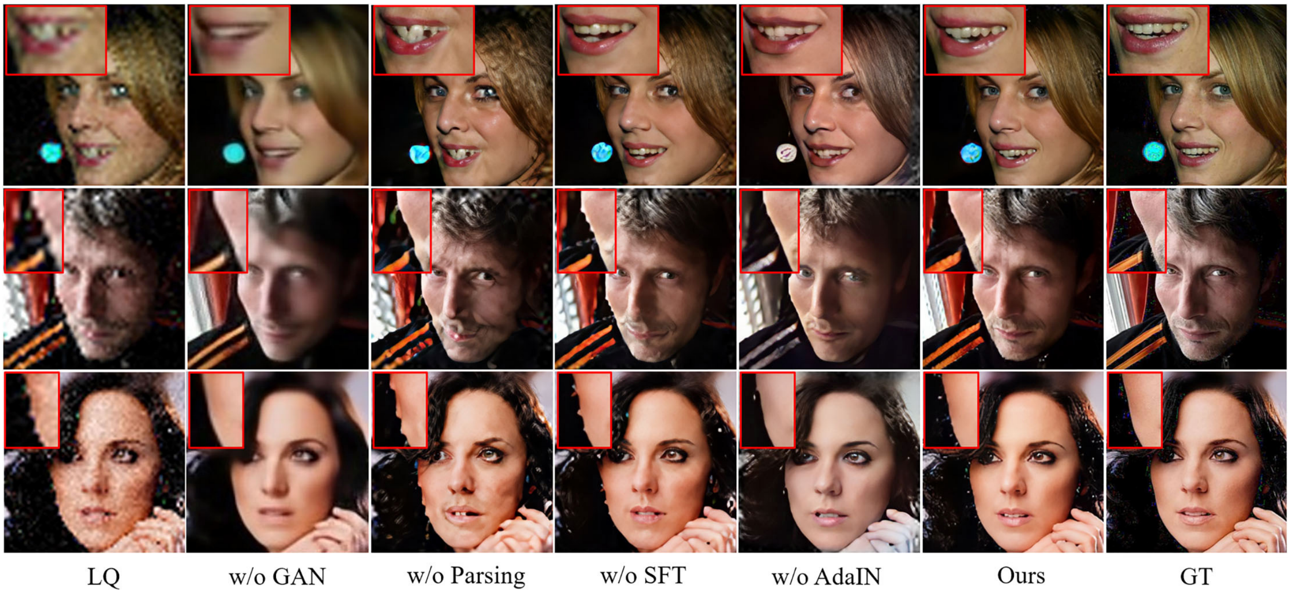


FIGURE 7 | Qualitative comparison of the effect of using different components to form the Blind Face Restoration (BFR) networks. Viewed best when zoomed in.

TABLE 2 | The quantitative performance of different variants on CelebA-Test.

Configuration	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
w/o GAN prior	120.14	0.5576	22.13	0.6165
w/o Parsing map prior	45.51	0.4133	21.54	0.6415
w/o AdaIN	29.49	0.3173	23.50	0.6629
w/o SFT	37.88	0.3838	22.64	0.6582
MPCNet (ours)	29.26	0.3097	23.82	0.6684

TABLE 3 | Quantitative comparison on CelebA-Test for blind face restoration (BFR).

Methods	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Input	158.72	0.6185	20.23	0.6823
WaveletSRNet	119.75	0.5351	22.87	0.6451
Super-FAN	100.23	0.5085	22.15	0.6377
DFDNet	43.74	0.3926	21.84	0.6439
HiFaceGAN	64.37	0.4795	21.05	0.5444
PSFRGAN	31.92	0.3226	23.17	0.6472
GPEN	31.41	0.3267	22.91	0.6428
MPCNet (ours)	29.26	0.3097	23.82	0.6684

the illumination condition of restored face consistent with the original input.

Spatial feature transform: *w/o SFT* denotes the model that consists of the decoder part of U-shaped DNN which leverages the encoded intermediate spatial features with types of facial priors to restore the HQ face, during which the SFT is abandoned. This model is included here to demonstrate the importance of SFT. As the comparison between MPCNet and w/o SFT shown in **Figure 7** and **Table 2**, it is evident that the SFT module can make full use of the parsing map priors to guide the face restoration branch to pay more attention to the essential facial parts reconstruction.

4.4. Comparison With the State-Of-The-Art

4.4.1. Comparison of Synthetic Dataset for BFR

To quantitatively compare MPCNet with other state-of-the-arts methods: WaveletSRNet (Huang et al., 2017), Super-FAN (Bulat and Tzimiropoulos, 2018), DFDNet (Li et al., 2020a), HiFaceGAN (Yang et al., 2020), PSFRGAN (Chen et al., 2021), and GPEN (Yang et al., 2021), we first perform experiments on synthetic images. Following the comparison experiments setting in Yang et al. (2021), we directly compared with these state-of-the-arts models trained by the original authors in the experiments. Except for Super-FAN, we adopt their official codes and finetune them on our face training set for fair comparisons. **Table 3** lists the perceptual metrics (FID and LPIPS) and pixel-wise metrics (PSNR and SSIM) results on the CelebA-Test testset. It can be seen that our MPCNet achieves comparable PSNR and SSIM indices to other competing methods, but it achieves significant performance gains over all the competing methods on FID and LPIPS indices, which are better measures than PSNR for the face image perceptual quality.

Figure 8 compares the BFR results on some degraded face images by the competing methods. One can see that the competing methods fail to produce reasonable face reconstructions. They tend to generate over-smoothed face images with distorted facial structures. Due to the powerful generative facial prior, it is obvious that our MPCNet is more

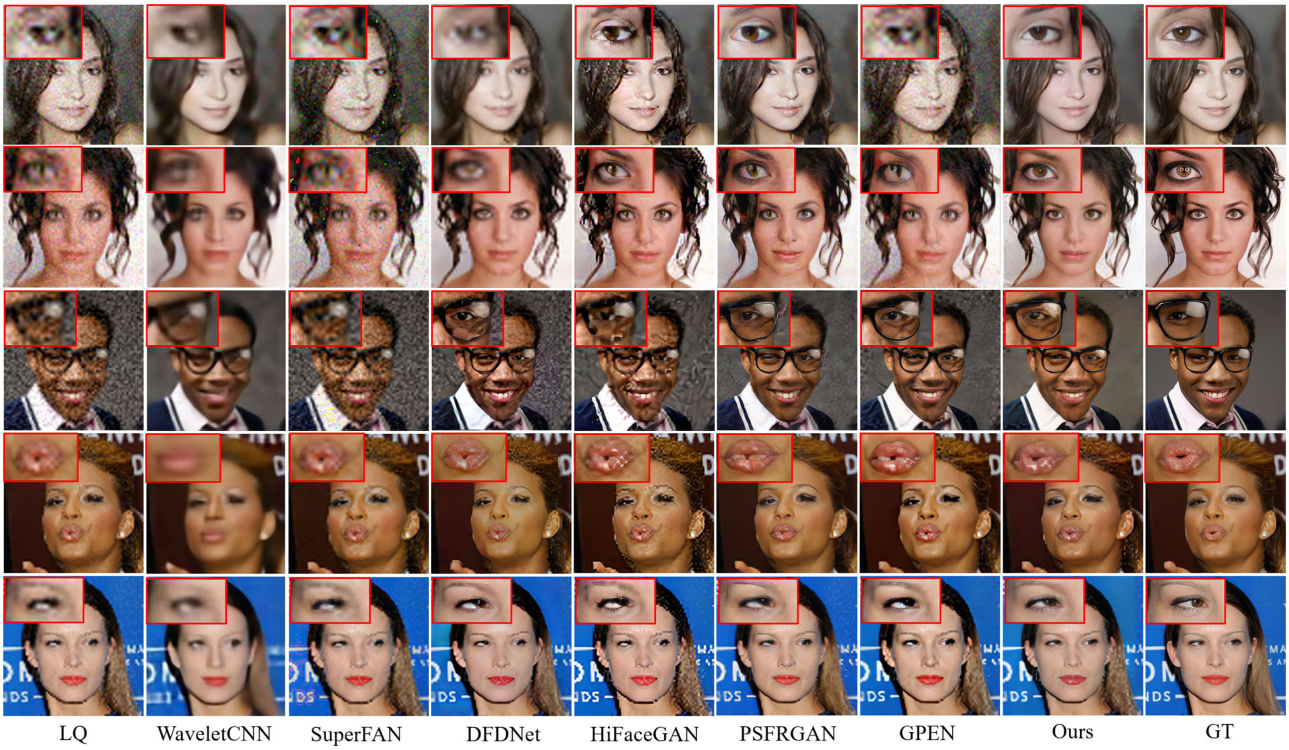


FIGURE 8 | Comparison of qualitative performance with state-of-the-art BFR methods from the literature.

TABLE 4 | Peak signal-to-noise ratio (PSNR) results achieved for ×4/ × 5/ × 6/ × 7/ × 8 face super-resolution (SR).

Methods	Scale				
	×4	×5	×6	×7	×8
Input	22.97	22.63	21.33	20.64	20.34
DFDNet	23.41	23.15	22.87	22.28	21.95
HiFaceGAN	23.12	22.85	22.68	22.19	21.17
PSFRGAN	24.52	24.14	23.92	23.44	23.24
GPEN	24.39	24.00	23.81	23.35	23.12
MPCNet (ours)	25.26	24.84	24.65	24.19	23.98

effective in restoring fine details while suppressing visual artifacts. In comparison with the competing methods, the results by MPCNet are visually photo-realistic and can correctly recover finer and identity-aware details, especially in eyes, nose, and mouth regions.

4.4.2. Experiments on Arbitrary Scales Face Super-Resolution (SR)

We can see from **Table 4** that our MPCNet achieves comparable performance to GPEN on all scale factors, with average PSNR improvements of 0.85. Compared to PSFRGAN, our MPCNet achieves notable performance improvements (23.98 vs. 23.24) for ×4 SR and (24.19 vs. 23.44) for ×7 SR. This clearly demonstrates that the proposed our MPCNet can enable scale-arbitrary SR without performance degradation on SR with fixed scale factors. **Figures 9**, **10** illustrate the qualitative SR results on two non-integral scale factors. As shown in these zoom-in regions, we can see that our MPCNet produces better visual results than other methods with fewer artifacts. For example, GPEN and PSFRGAN cannot recover the eyes and mouth regions reliably and suffer from obvious distorted artifacts. In contrast, our MPCNet produces finer details.

4.4.3. Experiments on Different Types Blur Kernels Degradations

We adopt 4 Gaussian blur kernels with different sizes and 4 motion blur kernels in four different directions to test the BFR performance of the competing methods. It can be observed from **Table 5** that HiFaceGAN produces relatively low performance on complex degradations. Since HiFaceGAN is sensitive to degradation estimation errors, its performance for complex degradations is limited. By incorporating the prior features from pretrained face synthesis GAN and face parsing network in an adaptive and progressive manner, our MPCNet exhibits good generalization on complex degradations. **Figure 11** further illustrates the visualization results produced by different methods. Our MPCNet achieves much better visual quality while other methods suffer obvious blurring artifacts.

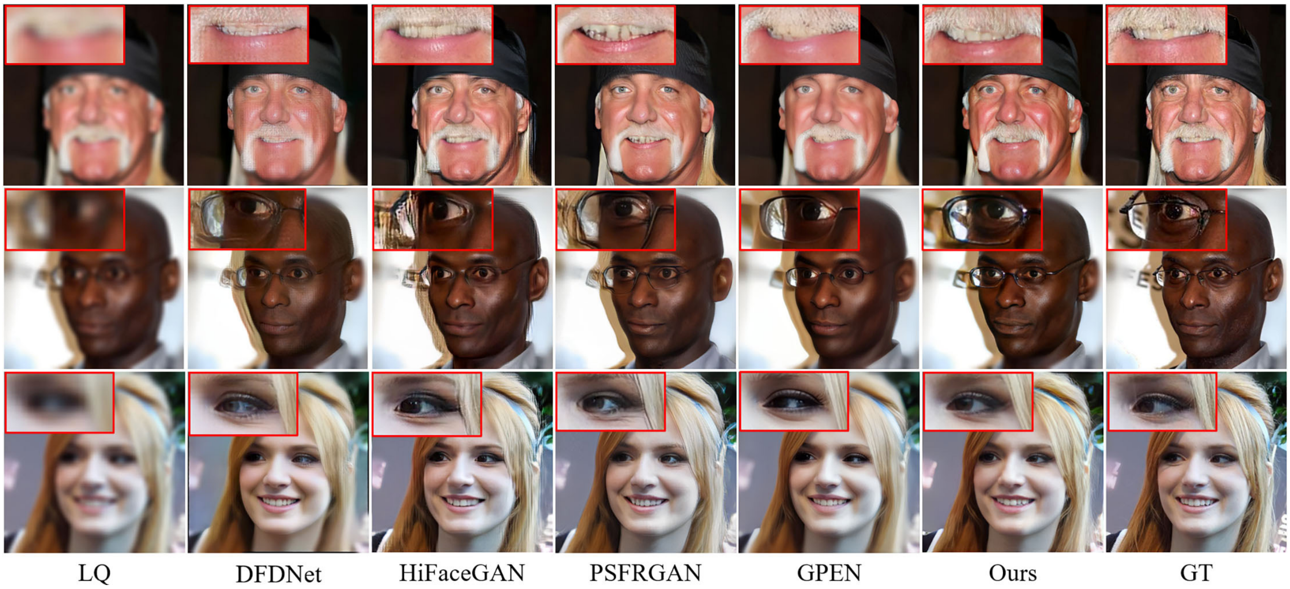


FIGURE 9 | Visual comparison for non-integer face super-resolution (SR) (i.e., ×6.5 SR, kernel width = 7).

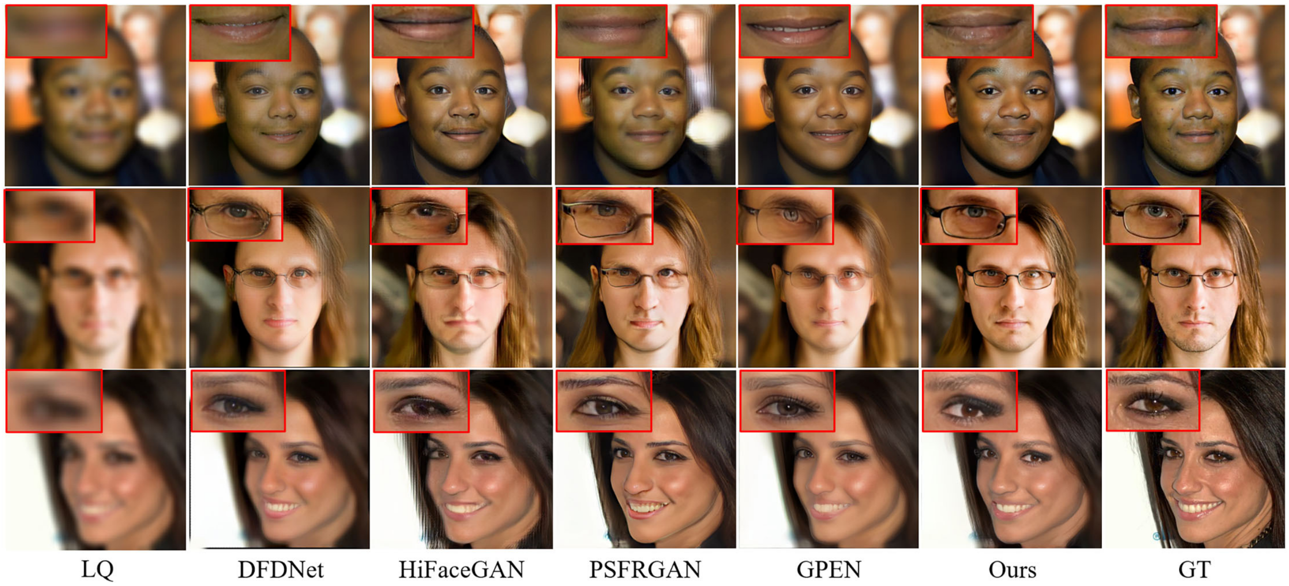


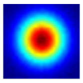
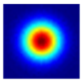
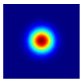
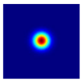
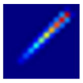
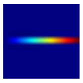
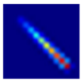
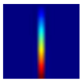
FIGURE 10 | Visual comparison for non-integer face SR (i.e., ×7.15 SR, kernel width = 10).

4.4.4. Experiments on Different Levels Noises Degradations

We set 6 noise levels to evaluate the restoration performance of the competing methods. In **Table 6**, we present the PSNR numbers for all noise levels. Since each APFF block can integrate generative priors and parsing maps priors to generate the fusion feature maps for guiding face restoration, when applying to complicated degradation scenarios, the fusion feature maps can correctly find where to incorporate guidance prior features in an adaptive manner, making our MPCNet outperform all the competitive algorithms for all noise levels.

Figures 12, **13** present the visual comparison outperforms all the other techniques published in **Table 6** and produces the best perceptual quality images. The closer inspections on the eyes, nose, and mouth regions reveal that our network generates textures closest to the

TABLE 5 | Peak signal-to-noise ratio results achieved on noise-free degradations with different blur kernels.

Method	Blur kernel							
								
Input	20.34	20.81	21.02	21.49	21.17	21.45	21.23	21.16
DFDNet	20.83	21.54	21.79	22.18	21.98	22.64	22.47	22.74
HiFaceGAN	20.41	21.19	21.65	21.89	21.64	22.41	22.21	22.49
PSFRGAN	22.56	22.92	23.14	23.58	23.12	23.09	22.76	23.27
GPEN	22.15	22.44	22.85	23.29	23.10	23.02	22.88	23.15
MPCNet (ours)	23.04	23.36	23.59	23.97	23.57	23.43	23.25	23.48

The kernel widths are set to 10.

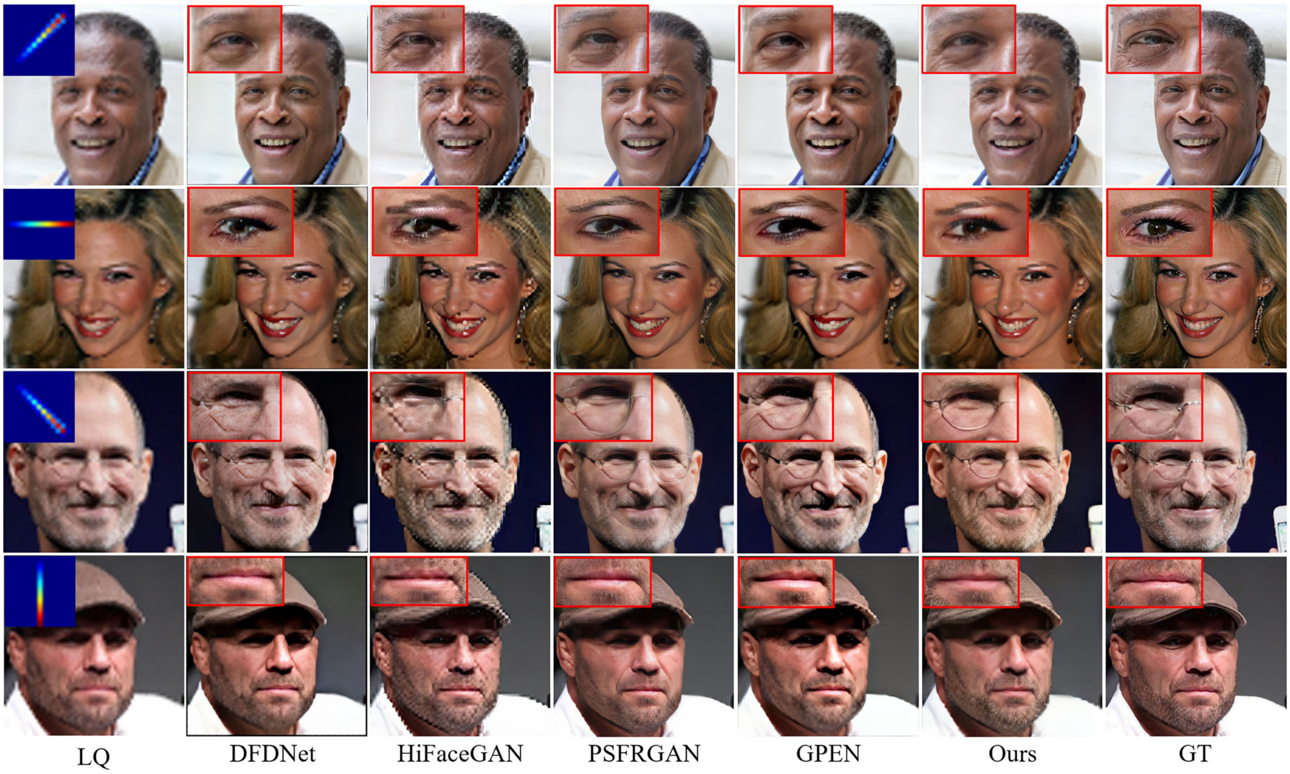


FIGURE 11 | Visual comparison achieved on noise-free degradations with different blur kernels. The blur kernels are illustrated with green boxes.

TABLE 6 | Peak signal-to-noise ratio results achieved on CelebA-Test degraded by different level noises.

Methods	Noise					
	0	**5**	**10**	**15**	**20**	**25**
Input	21.54	21.25	20.91	20.67	20.45	20.21
DFDNet	22.26	21.94	21.77	21.39	21.08	20.70
HiFaceGAN	21.93	21.59	21.25	20.98	20.74	20.33
PSFRGAN	23.66	23.50	23.38	23.06	22.79	22.42
GPEN	23.31	23.17	22.94	22.78	22.31	22.09
MPCNet (ours)	24.03	23.81	23.65	23.43	23.24	22.95

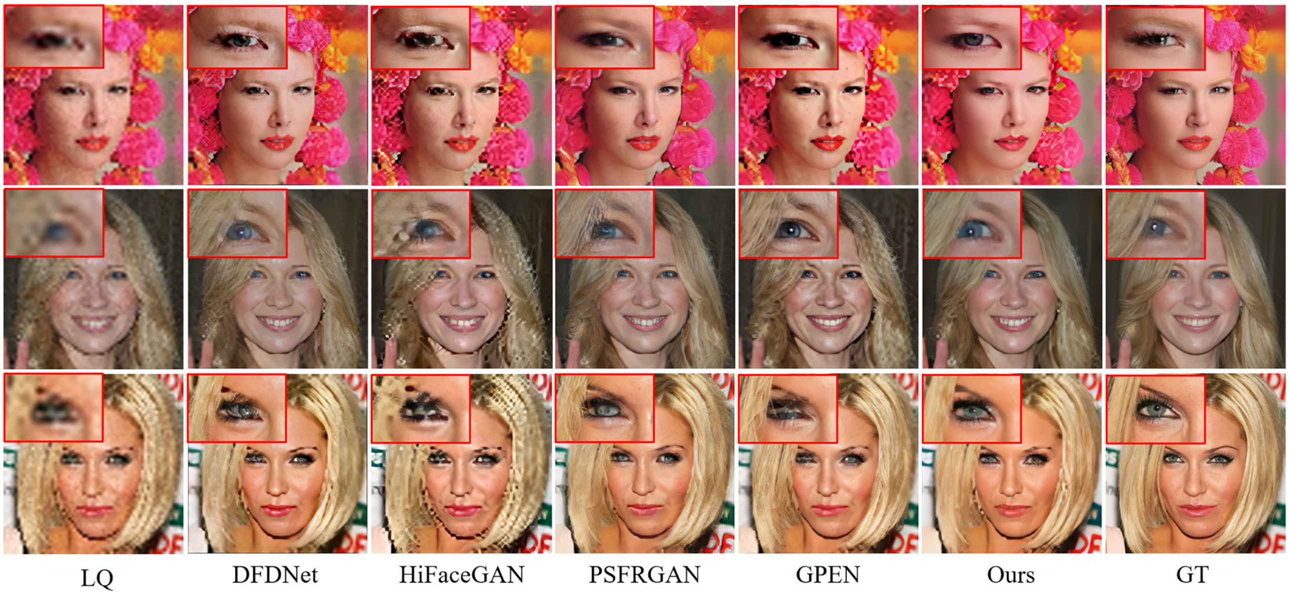


FIGURE 12 | Visual comparison achieved on CelebA-Test with the noise level set at 5.

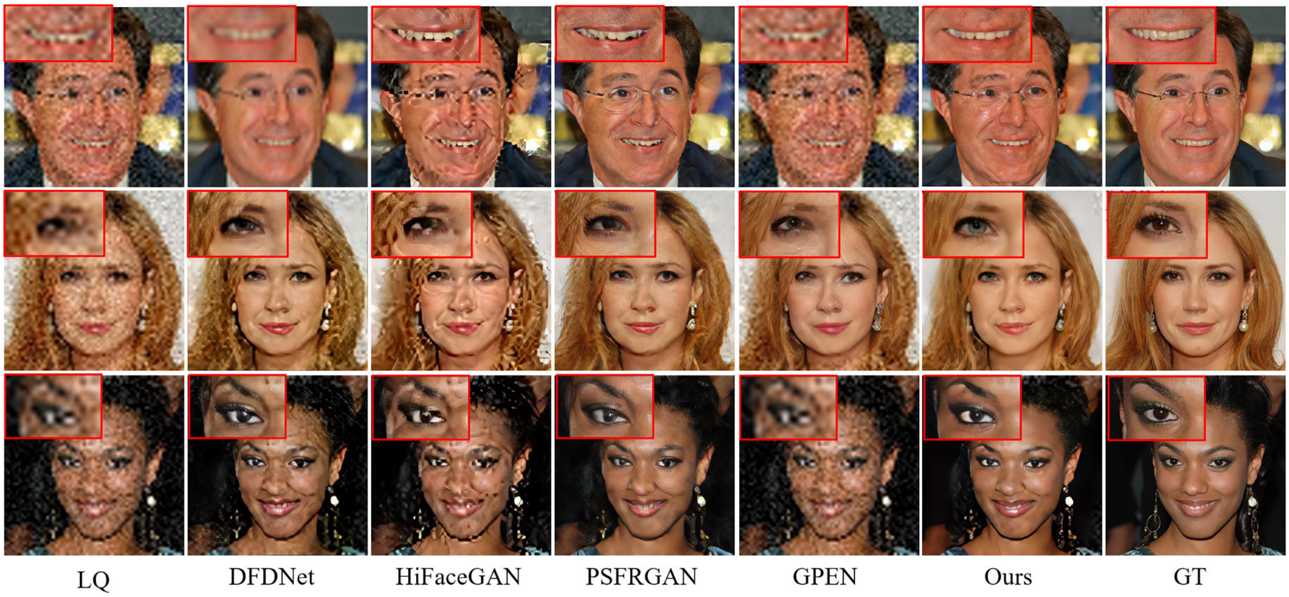


FIGURE 13 | Visual comparison achieved on CelebA-Test with the noise level set at 15.

ground-truth with fewer artifacts and more details for all noise levels.

4.4.5. Comparison of Real World LQ Images

To test the generalization ability, we evaluate our model on the real-world dataset. The quantitative results are shown in **Table 7**. Our MPCNet achieves superior performance and shows its remarkable generalization capability. Although GPEN also obtains comparable perceptual quality, it still fails in recovering the faithful face details as shown in **Figures 14**, **15**.

The qualitative comparisons are shown in **Figures 14**, **15**. The cropped LR face images from real-world images in **Figures 14**, **15** are 24 × 24 pixels and 36 × 36 pixels, and then we rescale the LR images to a fixed input size for MPCNet of 512 × 512 pixels. Thus, the scale factors of the visual comparisons are 21.4× and 14.2×, respectively. MPCNet seamlessly integrates the advantages of generative priors and face-specific geometry priors for restoring real-life photos with faithful facial details. Since the generative priors can provide adequate details and the parsing map priors provide geometry and semantic information, our method could produce plausible and realistic faces on complicated real-world degradation while other methods fail to recover faithful facial details or produce artifacts. Not only can our method perform well in common facial components like mouth and nose, but it can also perform better in hair and ears, as the parsing map priors can take the whole face into consideration rather than separate parts.

TABLE 7 | Quantitative comparison on the real-test.

Methods	NIQE ↓	FID ↓
Input	14.623	183.73
DFDNet	5.824	107.36
HiFaceGAN	6.141	124.12
PSFRGAN	5.783	94.51
GPEN	5.597	91.63
MPCNet (ours)	4.849	89.18

5. CONCLUSION

We have proposed a MPCNet to seamlessly integrate the advantages of generative priors and face-specific geometry priors. Specifically, we pretrained an HQ face synthesis GAN and a parsing mask prediction network and then embedded them into a U-shaped DNN as decoder priors to guide face restoration, during which the generative priors can provide adequate details and the parsing map priors provide geometry and semantic information. By designing an adaptive priors feature fusion (APFF) block to incorporate the prior features from pretrained face synthesis GAN and face parsing network in an adaptive and progressive manner, our MPCNet exhibited good generalization in a real-world application. Experiments demonstrated the superiority of our MPCNet in comparison to state-of-the-arts and also showed its potential in handling real-world LQ images from several practical applications.

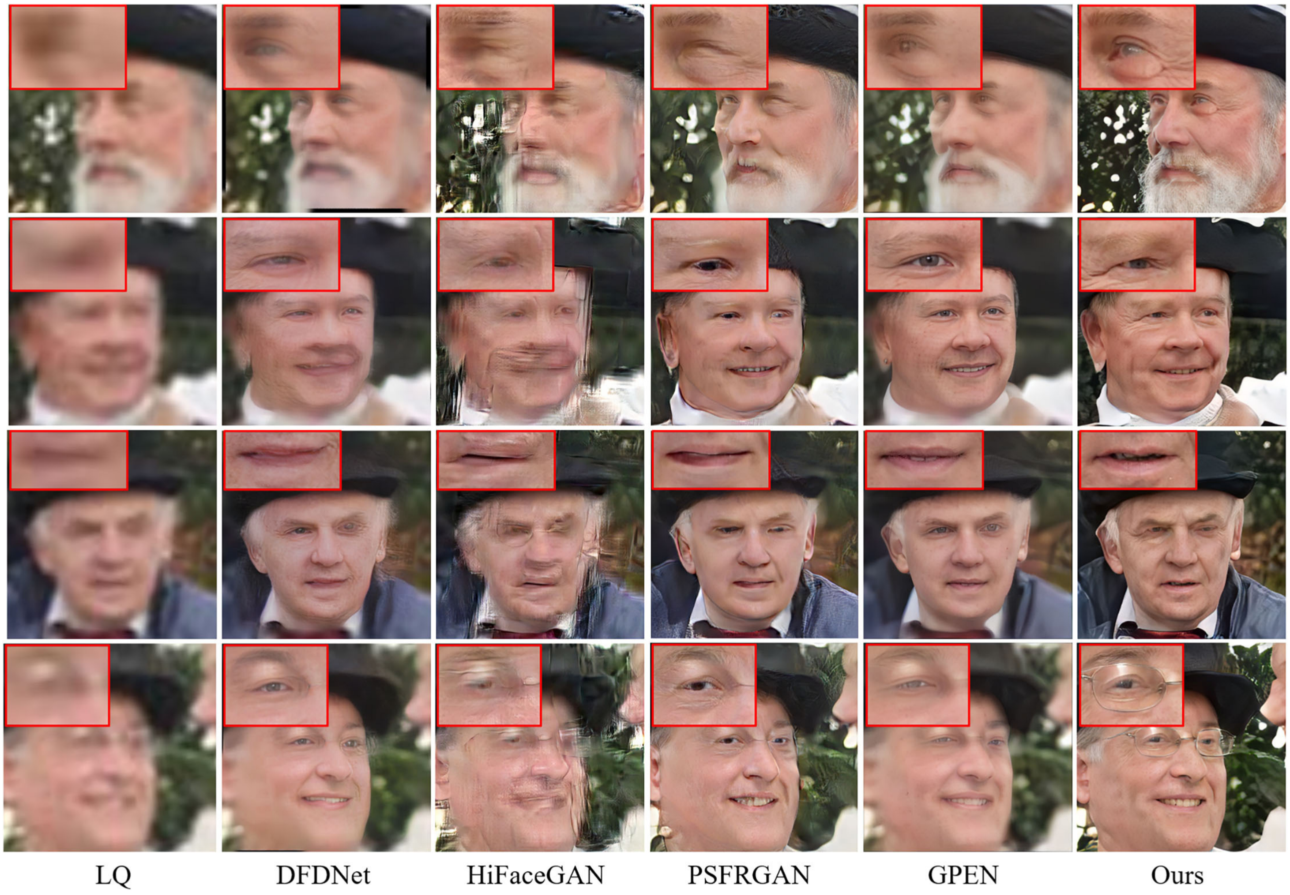


FIGURE 14 | Visual comparisons of competing methods with top performance on real-world low-quality (LQ) images (×21.4 SR).

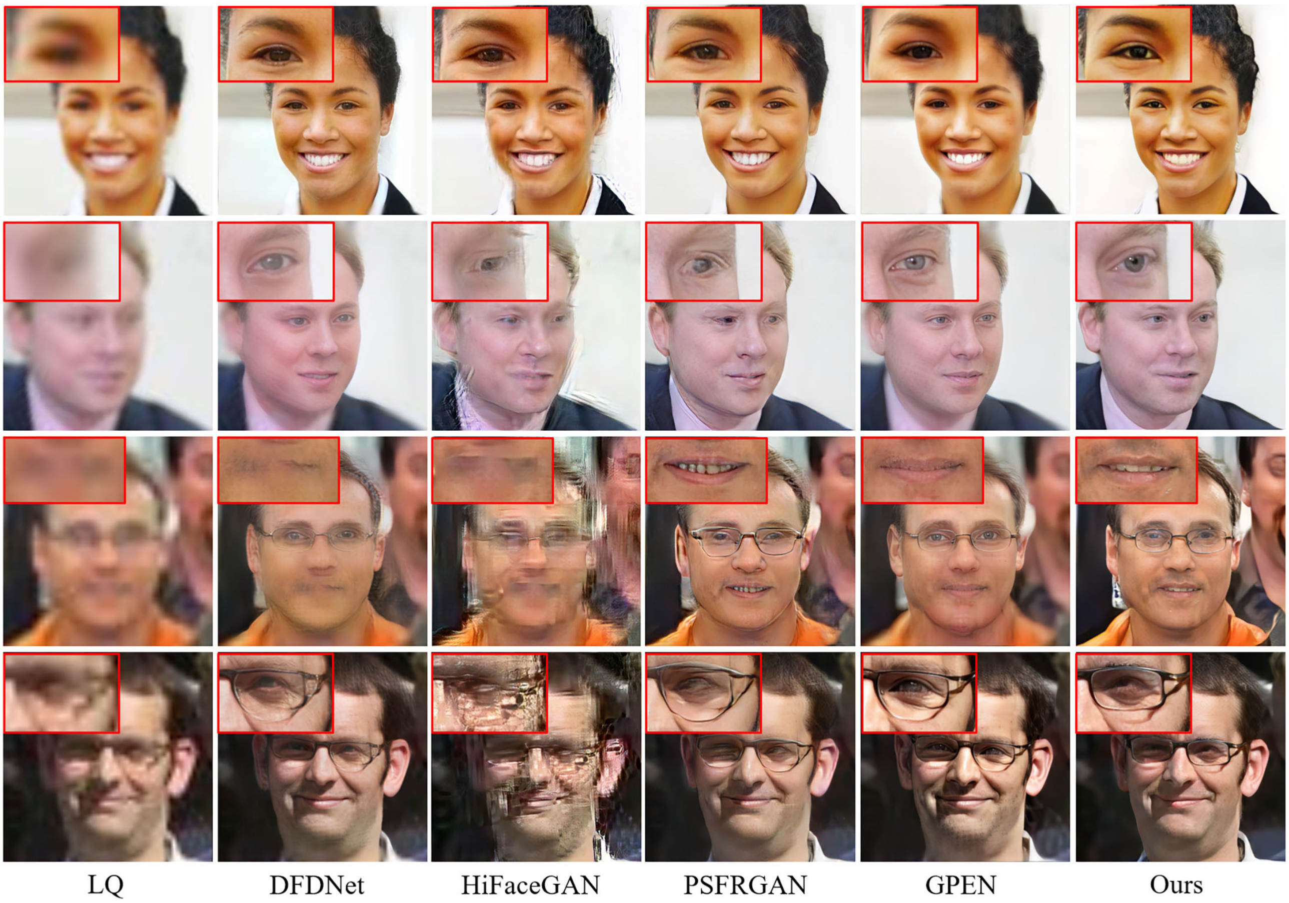


FIGURE 15 | Visual comparisons of competing methods with top performance on real-world LQ images (×14.2 SR).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZT: conceptualization, methodology, software, writing–original draft preparation, and data curation. XY: software, validation, visualization, and supervision. CW: investigation, writing–reviewing, and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (under Grant Nos. U20A20197 and 61973063), Liaoning Key Research and Development Project (2020JH2/10100040), Natural Science Foundation of Liaoning Province (2021-KF-12-01), and the Foundation of National Key Laboratory (OEIP-O-202005).

REFERENCES

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

Bell-Kligler, S., Shocher, A., and Irani, M. (2019). Blind super-resolution kernel estimation using an internal-gan. *arXiv preprint* arXiv:1909.06581.

Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint* arXiv:1809.11096.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 109–117.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). "Vggface2: a dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)* (Xi'an: IEEE), 67–74.

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 *Proceedings of the IEEE/CVF 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages* 2492–2501. doi: 10.1109/CVPR.2018.00264

Fan, Z., Hu, X., Chen, C., Wang, X., and Peng, S. (2020). Facial image super-resolution guided by adaptive geometric features. *EURASIP J. Wireless Commun. Networking* 2020, 1–15. doi: 10.1186/s13638-020-01760-y

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 2414–2423.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst*. 27.

Gu, J., Shen, Y., and Zhou, B. (2020). "Image processing using multi-code gan prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3012–3021.

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. *Adv. Neural Inf. Process. Syst*. 30.

Hore, A., and Ziou, D. (2010). "Image quality metrics: psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition* (Istanbul: IEEE), 2366–2369.

Huang, H., He, R., Sun, Z., and Tan, T. (2017). "Wavelet-srnet: a wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 1689–1697.

Huang, X., and Belongie, S. (2017). "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 1501–1510.

Jiang, J., Yu, Y., Hu, J., Tang, S., and Ma, J. (2018). Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination. *arXiv preprint* arXiv:1806.10726. doi: 10.24963/ijcai.2018/107

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint* arXiv:1710.10196.

Karras, T., Laine, S., and Aila, T. (2019). "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 8110–8119.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.

Lee, C.-H., Liu, Z., Wu, L., and Luo, P. (2020). "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5549–5558.

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* (Berlin; Heidelberg: Springer), 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 2706–2715.

Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., and Yang, R. (2018). "Learning warped guidance for blind face restoration," in *Proceedings of the European conference on Computer Vision (ECCV)*, 272–289.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 3730–3738.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. (2020). "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*, 2437–2445.

Michaeli, T., and Irani, M. (2013). "Nonparametric blind super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney, NSW: IEEE), 945–952.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint* arXiv:1802.05957.

Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C. C., and Luo, P. (2021). Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Trans. Pattern Anal. Mach. Intell*. doi: 10.1109/TPAMI.2021.3115428

Shocher, A., Cohen, N., and Irani, M. (2018). "zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3118–3126.

Soh, J. W., Cho, S., and Cho, N. I. (2020). "Meta-transfer learning for zero-shot super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3516–3525.

Song, Y., Zhang, J., He, S., Bao, L., and Yang, Q. (2017). Learning to hallucinate face images via component generation and enhancement. *arXiv preprint* arXiv:1708.00223. doi: 10.24963/ijcai.2017/633

Wang, C., Zhong, Z., Jiang, J., Zhai, D., and Liu, X. (2020). "Parsing map guided multi-scale attention network for face hallucination," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 2518–2522.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018a). "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 8798–8807.

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

Wang, X., Yu, K., Dong, C., and Loy, C. C. (2018b). "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 606–615.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., et al. (2018c). "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process*. 13, 600–612. doi: 10.1109/TIP.2003.819861

Yang, L., Wang, S., Ma, S., Gao, W., Liu, C., Wang, P., et al. (2020). "Hifacegan: Face renovation via collaborative suppression and replenishmen," in *Proceedings of the 28th ACM International Conference on Multimedia*, 1551–1560.

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

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.

Yu, X., and Porikli, F. (2017). "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 3760–3768.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). "Self-attention generative adversarial networks," in *International Conference on Machine Learning* (Venue: PMLR), 7354–7363.

Zhang, K., Gool, L. V., and Timofte, R. (2020). "Deep unfolding network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 3217–3226.

Zhang, K., Zuo, W., and Zhang, L. (2018a). "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3262–3271.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018b). "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 586–595.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. (2018c). "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhu, J., Shen, Y., Zhao, D., and Zhou, B. (2020). "In-domain gan inversion for real image editing," in *European*

Conference on Computer Vision (Berlin; Heidelberg: Springer), 592–608.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potentialconflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Teng, Yu and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.