FaceFormer: Scale-aware Blind Face Restoration with Transformers

Aijin Li¹, Gen Li¹*, Lei Sun¹, Xintao Wang²
Platform Technologies, Tencent PCG ¹
ARC Lab, Tencent PCG ²

Abstract. Blind face restoration usually encounters with diverse scale face inputs, especially in the real world. However, most of the current works support specific scale faces, which limits its application ability in real-world scenarios. In this work, we propose a novel scale-aware blind face restoration framework, named FaceFormer, which formulates facial feature restoration as scale-aware transformation. The proposed Facial Feature Up-sampling (FFUP) module dynamically generates upsampling filters based on the original scale-factor priors, which facilitate our network to adapt to arbitrary face scales. Moreover, we further propose the facial feature embedding (FFE) module which leverages transformer to hierarchically extract diversity and robustness of facial latent. Thus, our FaceFormer achieves fidelity and robustness restored faces, which possess realistic and symmetrical details of facial components. Extensive experiments demonstrate that our proposed method trained with synthetic dataset generalizes better to a natural low quality images than current state-of-the-arts.

1 Introduction

Blind face restoration can be defined as the ability to recovering the high quality (HQ) realistic details from its low quality (LQ) ones without knowing the degradation types, such as low-resolution, blur, noisy and lossy compression. In many real-world scenarios, the restoration performance of the face images is degraded by variant posture, occlusion, cluttered background, etc..

To handle the aforementioned issues, various deep learning-based approaches have been proposed and have shown impressive performance in face restoration. These approaches can be divided into three categories, such as facial prior guided, facial attribute information and identity information [1]. Unlike general image restoration, the face prior guided approaches utilize extra facial prior knowledge to recover facial details such as wrinkles, pupils, and eyelashes, etc.. Hence, many related literature adopts facial parsing maps, landmarks and heatmaps. However, these prior information are usually difficult to correctly extract from LQ image due to severe degradation. In addition, the facial attribute-based approaches utilize a single or multiple semantic information, such as gender, hair color,
age, expression and mustache, etc. The restored face images bring in different artifacts due to one-to-many maps from LQ images to HQ ones. For the identity information, those methods are always used for keeping the identity consistency between its restorations and ground truths. However, those methods are confined to the capacity of face dictionaries or high quality references.

To our best knowledge, although recent approaches have achieved promising performance, most the face restoration approaches are developed for better super-resolving under a specific resolution of low-quality faces. To guarantee the specific input resolution, conventional upsampling operations, such as bicubic interpolation [2,3], are usually employed. However, the interpolation-based upsamplers enlarge the image resolution only based on own image signals, without bringing any more information. Instead, they often introduce some side effects, such as artifacts, noise amplification and blurring restorations [1].

Therefore, to overcome these drawbacks, this work proposes a novel scale-aware blind face restoration framework, named **FaceFormer**. Our method consists of facial feature up-sampling, facial feature embedding and facial feature generator.

Specifically, in this work, we first integrate a super-resolving architecture with a fractional sampler module to refine more accurate scale-aware features, which makes the proposed method applicable and flexible in the arbitrary scale factor for the blind face restoration. After that, inspired by recent success of transformer [4], we leverage the hierarchical transformer blocks for the facial feature embedding with some different non-overlapping local windows and cross-window connection to extract diversity and robustness of facial latent feature. Moreover, the generated face images by proposed method usually possess a realistic and symmetrical facial details, such as pupil type, eyebrow and eyelash etc..

We summarize the our contributions as follows.

1. We propose a novel fractional up-sampler for blind face restoration. The module manages to generate scale-aware super-resolutions of low quality faces in the real-world scenarios.

2. Furthermore, we propose the FaceFormer framework with the facial feature up-sampling module and facial feature embedding module to improve fidelity and robustness. The facial feature embedding module utilizes some different non-overlapping local windows and cross-window connection to extract the more robust and diverse tiny facial component features.

3. Extensive experiments demonstrate that our proposed method trained with synthetic datasets generalizes better to natural low quality images than current state-of-the-arts.

## 2 Related Works

### 2.1 Face Super-Resolution

Face super resolution is often studied as a basic task for face restoration. Research [5] proposed a bidirection-framework that alternately optimizes facial restoration and dense corresponding field estimation in a cascade manner. Research
Fig. 1. The framework of proposed FaceFormer. It is mainly composed of Facial Feature Up-sampling, Facial Feature Embedding and Facial Feature Generator.

[6] employed generative adversarial networks (GAN) to directly super-resolve low resolution (LR) inputs. More researchers further exploited challenging problems like unaligned faces [7], noisy faces [8] and faces with different attributes [9]. Wavelet-SRNet [10] manages to generate high resolution (HR) wavelet coefficient sequence for HR image reconstruction instead of directly inferring HR image. Recent researches utilize additional face priors to improve performance. FSRNet [11] employs face landmark heatmaps and parsing maps to refine the super-resolution results, especially for misaligned LR faces. Instead, SuperFAN [12] forms face SR and landmark prediction as multi-task learning. In addition to the LR super-resolution branch, research [13] extra exploits a face component heat map to keep the face structure. Research [14] employs the difference heatmap as weights for multi-scale facial attention loss to better landmark restorations. Nonetheless, most of these researches succeed a general encoder-decoder framework without full application of face priors. It is not well suitable to handle real-world scenarios.

2.2 Blind Face Restoration

To deal with blind face restoration, recent works typically tried to adopt face-specific information, geometry priors, facial attribute information and identity information. The geometry priors include facial landmarks [11,14,15], face parsing maps [15,13,11] and facial component heatmaps [13]. However, these prior information are usually tough to correctly extract from LQ image due to severe degradation in real-world scenarios.
For attribute information, it is accessible without LQ face images, which is distinct from face structure prior information. AGCycleGAN [16], FSRSA variants [9] and ATNet [17] versions are directly concatenate attribute information and LR face image. The aforementioned methods take no consideration for the absence of attributes, limiting their applicability in real world scenes. RAAN [18] constructs multi-branch residual attention to further aggregate attribute information. Instead, FACN [19] proposes a capsule generation block for attribute integration.

Identity information usually refers to HQ face images of the same identity with LR faces. GFRNet [20] employs single HQ face images to share the same identity. while the variable facial expressions and poses may suppress the restoration of face images. ASFFNet [21] is the forerunner to explore multi-face guided FSR, which selects the best reference image by guidance selection module. Its variants [22] further cope with misalignment and illumination differences of multi-face guided face restoration. According to the observation that different people may have similar facial components, research [23] builds a component dictionary to advance face restoration. DFDNet [24] builds multi-scale component dictionaries based on features of the entire dataset.

2.3 Vision Transformer

Recently, natural language processing model Transformer [25] received wide attention in the computer vision field. It learns to pay attention to important image regions by exploring the global interactions between different regions, and is widely available for image classification [4], object detection [26], seg-
Fig. 3. The computation of coordinate \((x', y')\) and relative distances \((R(x)\) and \(R(y)\)) in LR feature space

3 Methodology

3.1 Overview of Our FaceFormer

This section describes the proposed FaceFormer architecture. Given an input face image \(I\) suffering from unknown degradation, the aim of blind face restoration is to estimate a high-quality image \(\hat{y}\), which is as similar as possible to the ground-truth image \(y\) in terms of both realistic and fidelity.

The architecture of the proposed FaceFormer is shown in Fig. 1. The FaceFormer is mainly composed of facial feature up-sampling (FFUP), facial feature embedding (FFE) and facial feature generator (such as pretrained Style-GAN2 [31]). Sequentially, to extract scale-aware upsampled facial features, the arbitrary scale facial image \(I\) is firstly fed into the FFUP module with scale factor \((s_h, s_v)\) and relative distance \((R(x), R(y)\) for scale-aware feature up-sampling. Then, the FFE module is proposed to extracted facial semantic features \(F_{\text{semantic}}\) which removing complicated and severer degradations of the output of FFUP module. The FFE module leverages transformer module that captures long-distance dependencies of the face images, such as symmetrical details of components.

Finally, the facial feature generator (FFG) module is employed to generate diverse and rich facial details. The facial semantic feature \(F_{\text{semantic}}\) is mapped to the closest latent features to the generative priors, which is provided by the pre-trained GAN. The latent features contain two kinds informative features: latent features \(F_{\text{latent}}\) and spatial features \(F_{\text{spatial}}\).
To summarise, the proposed FaceFormer achieves realistic results while preserving high fidelity in a coarse-to-fine manner.

3.2 Facial Feature Up-sampling

In the real world, face restoration explicitly deals with arbitrary resolution LQ images. Many attempts [2,3] employ interpolation-based upsamplings to handle multiple scale factors without considering the scale information during feature learning. Intuitively, since the degradation is directly related to various scale factors, scale information can further be adapted to learn discriminative features to improve HQ image performance. Moreover, these upsampling methods often introduce some side effects, such as artifacts, noise amplification, blurring restorations [32]. To reduce these effects, we propose a novel scale-aware facial feature up-sampling module instead of interpolation upsamplings of the single image super resolution (SISR). The proposed module dynamically generates upsampling filters conditioned on scale information. The proposed architecture is depicted in Fig. 2.

To extract dense facial features $F_{dense}$, given a LQ facial image $I$, then, $I$ is fed into the backbone module of a classical super resolution architecture without any upsamplings and reconstructions, such as SRResNet [33], and ESRGAN [34], etc.. Then, following the $F_{dense}$ features, a scale-aware fractional up-sampling layer is designed for arbitrary-scale upsampling, instead of subpixels [35] or de-convolution layer. Besides, the proposed upsampling layer can be applied for a fractional super-resolving through a simple yet effective modification. This upsampling layer consists of two stages, spatial-sampling stage and channel-varying stage with fractional scale factor inputs.

Given a pixel ($x, y$) in the high resolution (HR) feature space, the pixel is firstly projected into the low resolution (LR) feature space. As shown in Fig. 3, its corresponding LR coordinate ($x', y'$) and relative distances ($R(x)$ and $R(y)$) are calculated as follows:

\[
\begin{align*}
  x' &= x + 0.5 \frac{s_h}{s_v} - 0.5 \\
  y' &= y + 0.5 \frac{s_v}{s_h} - 0.5 \\
  R(x) &= x' - \left\lfloor \frac{x + 0.5}{s_h} \right\rfloor \\
  R(y) &= y' - \left\lfloor \frac{y + 0.5}{s_v} \right\rfloor
\end{align*}
\]

(1)

where the $s_h$ and $s_v$ denote horizontal and vertical scale factor, respectively. For a pixel ($x, y$), its corresponding LR coordinate ($x', y'$) are theoretically projected as equation (1). However, the generated coordinate ($x', y'$) is non integer
in most cases. Thus, the floor operation $\lfloor \cdot \rfloor$ are often employed for coordinate quantization. The relative distances $(R(x)\text{ and } R(y))$ denotes the projection bias from HR feature space to LR feature space.

As shown in Fig. 2, the relative distances $(R(x), R(y))$ and horizontal and vertical scale factors $(s_h, s_v)$ are firstly concatenated and fed into two Fully Connection (FC) layers for scale-aware feature extraction. The relative distances $(R(x), R(y))$ reflect the information interaction between HR and LR feature space for more sufficient scale-aware feature extraction. Then, offset filter and scale-aware filter are adopt to the extracted feature for offsets $(\delta_x, \delta_y)$ and scale-varying weights $W_{\text{scale}}$, respectively. Here, to reduce the complexity and improve the generalization ability, these two kernels refer to two FC layers as follows:

$$(\delta_x, \delta_y), W_{\text{scale}} = FCs(s_h||s_v||R(x)||R(y))$$

where the $||$ denotes concatenation.

Subsequently, the scale-varying weights serve as scale guidance to combine squeeze and expand kernels, bringing about a pair of squeeze Conv$_{sq}$ and expansion Conv$_{ex}$ convolution layers. The weights of these convolution layers are adaptively learned conditioned on scale information, while the interpolation-based upsamplings, such as bicubic upsampling adopts a fixed weight.

In the spatially-sampling stage, given the predicted horizontal and vertical offsets $\delta_x$ and $\delta_y$, a $k \times k$ neighborhood centered at $(x'+\delta_x, y'+\delta_y)$ in the previous dense facial feature $F_{\text{dense}}$ are upsampled using Grid Sample [36] as follows:

$$F_{\text{spatial\_varying}} = \text{GridSample}(x'+\delta_x, y'+\delta_y)_{k\times k}$$

in the channel-varying stage, the spatially-sampled feature $F_{\text{spatial\_varying}}$ is fed into with the scaled channel-varying filters for scale information interaction in channels. The output are fused with $F_{\text{spatial\_varying}}$ to generate $F_{\text{upsample}}$ in a residual way as follows:

$$F_{\text{upsample}} = \text{Conv}_{ex}(\text{Conv}_{sq}(F_{\text{spatial\_varying}})) + F_{\text{spatial\_varying}}$$

In general, a conventional up-sampling operator adopts its adjacent pixel filling algorithm, which mainly considers a spatially-wise correlation and does not well consider a channel-wise correlation. However, the proposed FFUP module takes care of the both spatially-wise and channel-wise correlation via a proposed fractional upsampler, which significantly improve performance of arbitrary scale face restoration.

### 3.3 Facial Feature Embedding

Typically, real-world blind face restoration faces suffer from complicated and severer degraded variations in resolution, blur, noise and JPEG artifacts. The
Facial Feature Embedding module is designed to explicitly strengthen aforementioned issues and extracted the facial semantic features $F_{\text{semantic}}$, providing reliable information for subsequent modules.

As shown in Fig. 1, we first extract facial attribute feature maps $F_{\text{attribute}}$ from FFUP module upsampled features $F_{\text{upsample}}$ as:

$$F_{\text{semantic}} = \text{Transformer}(F_{\text{upsample}})$$  \hspace{1cm} (6)

where $\text{Transformer}(\cdot)$ is a transformer-based feature extraction module (such as swin transformer, vision transformer etc.) and it contains a position linear embedding and $K$ residual Swin Transformer blocks (STB). The multi-resolution spatial features $F_{\text{spatial}}$ are used to modulate the GAN prior features. Note that we use the last convolution layer follow by the STB. Using a convolution layer at the end of feature extraction can bring the inductive bias of the convolution operation into the Transformer-based network, and lay a better foundation for the later aggregation of shallow and deep features.

### 3.4 Facial Feature Generator

The pre-trained GAN models capture the facial feature distribution, named generative priors, from its learned weight of convolutions. In this work, following GFP-GAN, we adopt the facial semantic feature from proposed FFE module and then map these feature to the pre-trained styleGAN2 to generate realistic and diverse face texture details. Specifically, given the facial semantic feature $F_{\text{semantic}}$, the latent feature $F_{\text{latent}}$ are generated as:

$$F_{\text{latent}} = \text{MLP}(F_{\text{semantic}})$$ \hspace{1cm} (7)

where $\text{MLP}(\cdot)$ denotes multi-layer perceptron layers. The $F_{\text{latent}}$ then pass through each convolution layer in the pre-trained GAN (such as styleGAN2) and generate GAN features for each resolution scale.

### 3.5 Model Objectives

The learning objective of training our FaceFormer consists of: 1) upsampling loss $L_{\text{up}}$ that constraints the FFUP outputs $y_{\text{up}}$ of arbitrary scale inputs close to the ground-truth $y$. 2) restoration loss $L_{\text{restoration}}$ for restored output $\hat{y}$ close to $y$.

**Upsampling Loss** We employ the widely-used L1 loss to further enhance the scale-aware significant face upsampling. The upsampling loss $L_{\text{up}}$ are defined as follows:

$$L_{\text{up}} = \lambda_1 \|y_{\text{up}} - y\|_1$$ \hspace{1cm} (8)
Restoration Loss  Restoration loss $L_{rest}$ are adopted to constraint the outputs $\hat{y}$ close to the ground-truth $y$. Typically, following the GFP-GAN [2], it contains reconstruction loss $L_{rec}$, adversarial loss $L_{adv}$, component loss $L_{comp}$ and identity loss $L_{id}$ as follows:

$$L_{rest} = \lambda_{rec} L_{rec} + \lambda_{adv} L_{adv} + \lambda_{comp} L_{comp} + \lambda_{id} L_{id}$$  \hspace{1cm} (9)

The restoration loss hyper-parameters are set as: $\lambda_{rec} = 0.1, \lambda_{adv} = 0.1, \lambda_{comp} = 0.1, \lambda_{id} = 1$.

The overall model objective is a combination of the above losses:

$$L_{total} = \lambda_1 L_{up} + \lambda_2 L_{rest}$$  \hspace{1cm} (10)

The total loss hyper-parameters $\{\lambda_1, \lambda_2\}$ are set as $\{0.01, 0.1\}$, respectively.

4 Experiments

4.1 Datasets and Implementation

Training Datasets. We adapt the FFHQ [37] as our training set, which consists of 70,000 high-quality images with size of $1024 \times 1024$. We resize the images to $512 \times 512$ with bilinear downampling as the ground-truth high quality images.

The proposed FaceFormer is trained on the synthetic data under the degradation model:

$$I = \left[ (y \otimes K_\sigma) \downarrow r + n_\delta \right]_{\text{JPEG}q}$$  \hspace{1cm} (11)

Given high quality image $y$, Gaussian blur kernel $K_\sigma$ is firstly employed to convolve with $y$ under parameter $\sigma$. Then the convolved image is downsampled with a scale factor $r$. Subsequently, the resulting image suffers from additive white Gaussian noise $n_\delta$ and finally it is compressed by JPEG with quality factor $q$. Following the practice in GFP-GAN [2], we randomly sample the parameters $\sigma, \delta, r$ and $q$ from $\{0.2 : 10\}, \{1 : 8\}, \{0 : 15\}$ and $\{60 : 100\}$, respectively. The color jittering is also added during training for color enhancement.

Testing Datasets. The FaceFormer is tested on one synthetic dataset and three different real-world datasets. All these datasets have no overlap with our training dataset. The following provide a brief introduction here.

CelebA-Test is the synthetic dataset with 3,000 CelebA-HQ images from its testing partition [38]. The generation is the same way as that during training.

LFW-Test. LFW [39] involves low-quality images in the wild. Here, the first image of each identity in the validation partition is assembled as test set, which contains 1711 images.

CelebChild-Test [2] contains 180 child faces of celebrities collected from the Internet. They are low-quality and many of them are black-and-white old photos.
Fig. 4. Qualitative comparison on the CelebA-Test for blind face restoration. Our GFP-GAN produces faithful details in eyes, mouth and hair. Zoom in for best view.

WebPhoto-Test [2] consists of 407 facial images cropped from 188 low-quality photos in real life. These photos have diverse and complicated degradation. Some of them are old photos with very severe degradation on both details and color.

Implementation. The Transformer for degradation removal module consists of encoder and decoder. The encoder contains four downsamples with \{2, 4, 6, 2\} Swin Transformer [4] blocks, respectively. And the decoder contains seven up-samples, each with a residual block. Following the work [2], we also adopt the pretrained StyleGAN2 [31] with 512×512 outputs as our generative facial prior.

Table 1. Quantitative comparison on CelebA-Test for blind face restoration. Red and blue indicates the best and the second best performance. (∗) denotes finetuning on our training set.

| Methods         | LPIPS ↓ | FID ↓ | NIQE ↓ | PSNR ↑ | SSIM ↑ |
|-----------------|---------|-------|--------|--------|--------|
| Input           | 0.4866  | 143.98| 13.440 | 25.35  | 0.6848 |
| DeblurGANv2(*)  | 0.4001  | 52.69 | 4.917  | 25.91  | 0.6952 |
| Wan et al. [41] | 0.4826  | 67.58 | 5.356  | 24.71  | 0.6320 |
| HiFaceGAN [42]  | 0.4770  | 66.09 | 4.916  | 24.92  | 0.6195 |
| DFDNet [24]     | 0.4341  | 59.08 | 4.341  | 23.68  | 0.6622 |
| PSFRGAN [3]     | 0.4240  | 47.59 | 5.123  | 24.71  | 0.6557 |
| mGANprior [43]  | 0.4584  | 82.27 | 6.422  | 24.30  | 0.6758 |
| PULSE [44]      | 0.4851  | 67.56 | 5.305  | 21.61  | 0.6200 |
| GFP-GAN [2]     | 0.3646  | 42.62 | 4.077  | 25.08  | 0.6777 |
| **FaceFormer(ours)** | **0.3237** | **13.17** | **4.369** | **26.94** | **0.7386** |

For training, the proposed FaceFormer is conducted with Adam optimizer of total 800k iterations. We use a learning rate of \(2e^{-3}\) with a MultiStepLR decay policy by a factor of 2 at the 700k-th, 750k-th iterations. We implement our models with the PyTorch framework and train them using two NVIDIA Tesla A100 GPUs.
4.2 Comparisons with State-of-the-art Methods

We compare the proposed FaceFormer with some state-of-the-art methods designed for different restoration tasks: HiFaceGAN [42], DFDNet [24], PSFRGAN [3], Super-FAN [12], Wan et al. [41], GFP-GAN [2], PULSE [44] and mGANprior [43] for face restoration; DeblurGANv2 [40] for natural image super-resolution.

Specially, HiFaceGAN, DFDNet and GFP-GAN are conducted for blind face restoration. Note that the above methods are adopted their official codes except for Super-FAN, for which we use a re-implementation.

Table 2. Quantitative comparison on the real-life LFW, CelebChild, WebPhoto. Red and blue indicates the best and the second best performance. (*) denotes finetuning on our training set.

| Dataset       | LFW-Test       | CelebChild     | WebPhoto       |
|---------------|----------------|----------------|----------------|
|               | FID ↓ NIQE ↓   | FID ↓ NIQE ↓   | FID ↓ NIQE ↓   |
| Input         | 137.56 11.214  | 144.42 9.170   | 170.11 12.755  |
| DeblurGANv2(*) | 57.28 4.309    | 110.51 4.453   | 100.58 4.666   |
| Wan et al.    | 73.19 5.034    | 115.70 4.849   | 100.40 5.705   |
| HiFaceGAN     | 64.50 4.510    | 113.00 4.855   | 116.12 4.885   |
| DFDNet        | 62.57 4.026    | 111.55 4.414   | 100.68 5.293   |
| PSFRGAN       | 51.89 5.096    | 107.40 4.804   | 88.45 5.582    |
| mGANprior     | 73.00 6.051    | 126.54 6.841   | 120.75 7.226   |
| PULSE         | 64.86 5.097    | 102.74 5.225   | 86.45 5.146    |
| GFP-GAN       | 49.96 3.882    | 111.78 4.349   | 87.35 4.144    |
| FaceFormer(ours) | 49.36 4.166 | 72.27 4.293 | 137.59 4.134 |

For the evaluation, we employ the widely-used non-reference perceptual metrics: FID [45] and NIQE [46] for real-life low quality faces without Ground-Truth (GT). These metrics measure the realism and statistic distance between the restorations and inputs. Smaller values indicate better restorations. In addition, we also adopt pixel-wise metrics (PSNR and SSIM) and the perceptual metric (LPIPS [47]) for the CelebA-Test with GT. The LPIPS evaluates the similarity of deep features between restorations and GTs to simulate human perceptual similarity. Smaller values indicate closer similarity.

**Synthetic CelebA-Test** The comparisons are conducted under these setting that blind face restoration whose inputs and outputs have the same resolution. Table 1 reports the quantitative results among comparison methods. It is can be observed that our FaceFormer get the lowest LPIPS score, indicating that FaceFormer is perceptually close to GT. Our FaceFormer also surpasses other methods by a large margin in terms of FID, and is 29% higher than the second best result of GFPGAN. It shows that the restoration is close to the real face distribution in statistical similarity of Inception [48] feature embedding. Besides the perceptual performance, our method also reports the pixel-wise metrics PSNR
and SSIM. It owns to the proposed Facial Feature Embedding, which learns detailed geometrical structure features.

Qualitative results are presented in Fig. 4. Thanks to the robustness FFE module, our FaceFormer recovers faithful details in the eyes (pupils and eye-lashes), nose, etc. Our FaceFormer is capable of retaining fidelity, e.g., it produces naturally closed mouth without forced addition of teeth as PSFRGAN does. Moreover, the mouth generated by FaceFormer perceptually looks more similar to GT than GFP-GAN.

![Fig. 5. Qualitative comparisons on three real-life datasets. Zoom in for best view.](image)

![Fig. 6. Ablation studies on the effectiveness of FFUP module, FFE module. Zoom in for best view.](image)
Real-Life LFW, CelebChild and WedPhoto-Test The final target of all methods is to restore real world LQ face images. To evaluate the generalization ability of comparison methods, we also evaluate the performance of FaceFormer on three different real-world datasets. The quantitative results are shown in Table 2. Our method achieves surpassing performance on LFW-test and CelebChild datasets, showing its remarkable generalization capability.

Fig. 5 reports the qualitative comparisons overall comparison methods. Our method could produce plausible and realistic faces on complicated real-world degradation while other methods fail to recover faithful facial details or produce transparent artifacts (especially in CelebChild-Test in Fig 5).

Our method takes arbitrary scale face as input and generates scale-aware realistic restorations through FFUP module. Thus, the restorations largely avoid the transparent artifacts caused by scale drift, especially at the boundary of face components. In addition, FaceFormer keep the same pupil color in the 4th row owing to FFE module.

It can be observed that results of HiFaceGAN, Wan et al. and GFPGAN contains artifacts in the mouth and eyes. This is most likely because of the unstable training of GAN without face prior guidance. Different from PSFR-GAN which utilizes parsing map, DFDNet needs to first detect the LQ face components and then matches them to a HQ dictionary. It may fail to find the correct reference when the LQ are too blurry or with large pose. For example, the results of DFDNet in the 3-nd row seems to generate blur teethes.

| Configuration                  | LPIPS ↓ | FID ↓ | NIQE ↓ | PSNR ↑ | SSIM ↑ |
|-------------------------------|---------|-------|--------|--------|--------|
| 1) Our FaceFormer with FFUP   | 0.3237  | 13.17 | 4.369  | 26.94  | 0.7386 |
| 2) Our FaceFormer w/o FFUP    | 0.3502 (↑) | 17.82 (↑) | 4.481 (↑) | 25.36 (↓) | 0.7182 (↓) |
| 3) Our FaceFormer w/o FFE     | 0.3646 (↑) | 42.62 (↑) | 4.077 (↓) | 25.08 (↓) | 0.6777 (↓) |

4.3 Ablation Studies
For ablation study, we test the proposed FaceFormer on CelebA-test set to discuss the impact of each module.

Impact of the FFUP. As shown in Table 3 (configuration 2) and Fig. 6, when we remove the Facial Feature Up-sampling layer, it can be observed that the restored faces produce apparent artifacts, especially in the teeth. The appearance of image artifacts is usually caused by the reduction of high-frequency spatial frequency, which is usually accompanied by the change of object edge width, mainly in the process of image re-sampling. Thus, the face identity cannot be well retained without the FFUP module (high perceptual metrics).
Fig. 7. Results for different input resolutions while fixing other degradation parameters. The ‘(·)’ denotes the input resolution.

Robustness to multi-scale Face Inputs. We conduct comparisons of the arbitrary scale inputs. Specifically, the scale of \( \{ \times 1.5, \times 2.4, \times 4.0, \times 6.0, \times 8.0 \} \) are adopt to the FFUP module as the upsampling scale, respectively. As shown in Fig. 7, our FaceFormer works well for upscale factors and can produce more reasonable results, such as more realistic eyebrows and skin texture. Note that when the upscale factor is large (\( \times 6.0 \), \( \times 8.0 \) scale), the FaceFormer restorations are perceptually superior to the GFP-GAN ones. We further provide additional visualizations of diverse upscale factors in supplemental material.

Impact of the FFE module. The FFE module is designed to provide diversity and robustness of facial latent feature, and strengthen the restoration ability for complicated degradation in the real world. Without the FFE module, as shown in Table 3 [configuration 3] and Fig. 6, it can be observed that the facial components are incomplete for the lack of local shape information, especially in teeth. Hence, the generated face has suboptimal fidelity in the absence of FFE module (high perceptual metrics and low pixel-wise metrics).

5 Conclusions

This paper proposes a scale-aware blind face restoration framework FaceFormer. The FaceFormer formulates facial feature restoration as scale-aware transformation procedure and then adopts hierarchical transformer embedding blocks to achieve high fidelity, realistic and symmetrical details of facial components with arbitrary scale inputs. Extensive comparisons demonstrate the superior capability of FaceFormer in joint face restoration for real-world images, outperforming prior art.
References

1. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence 43 (2020) 3365–3387
2. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 9168–9178
3. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 11896–11905
4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 10012–10022
5. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: European conference on computer vision, Springer (2016) 614–630
6. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: European conference on computer vision, Springer (2016) 318–333
7. Yu, X., Porikli, F.: Face hallucination with tiny unaligned images by transformative discriminative neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 31. (2017)
8. Yu, X., Porikli, F.: 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. (2017) 3760–3768
9. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 908–917
10. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 1689–1697
11. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2492–2501
12. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 109–117
13. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European conference on computer vision (ECCV). (2018) 217–233
14. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. arXiv preprint arXiv:1908.08239 (2019)
15. Shen, Z., Lai, W.S., Xu, T., Kautz, J., Yang, M.H.: Deep semantic face deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 8260–8269
16. Lu, Y., Tai, Y.W., Tang, C.K.: Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European conference on computer vision (ECCV). (2018) 282–297
17. Li, M., Zhang, Z., Yu, J., Chen, C.W.: Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement. IEEE Transactions on Multimedia 23 (2020) 468–483
18. Xin, J., Wang, N., Gao, X., Li, J.: Residual attribute attention network for face image super-resolution. In: Proceedings of the AAAI conference on artificial intelligence. Volume 33. (2019) 9054–9061
19. Xin, J., Wang, N., Jiang, X., Li, J., Gao, X., Li, Z.: Facial attribute capsules for noise face super resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 12476–12483
20. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European conference on computer vision (ECCV). (2018) 272–289
21. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 2706–2715
22. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. (2017) 1501–1510
23. Li, X., Duan, G., Wang, Z., Ren, J., Zhang, Y., Zhang, J., Song, K.: Recovering extremely degraded faces by joint super-resolution and facial composite. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE (2019) 524–530
24. Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., Zhang, L.: Blind face restoration via deep multi-scale component dictionaries. In: European Conference on Computer Vision, Springer (2020) 399–415
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
26. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision, Springer (2020) 213–229
27. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) 6881–6890
28. Liang, D., Chen, X., Xu, W., Zhou, Y., Bai, X.: Transcrowd: Weakly-supervised crowd counting with transformer. arXiv preprint arXiv:2104.09116 (2021)
29. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12299–12310
30. Cao, J., Li, Y., Zhang, K., Van Gool, L.: Video super-resolution transformer. arXiv preprint arXiv:2106.06847 (2021)
31. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 8110–8119
32. Aganj, I., Yeo, B.T.T., Sabuncu, M.R., Fischl, B.: On removing interpolation and resampling artifacts in rigid image registration. IEEE Transactions on Image Processing 22 (2012) 816–827
33. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 4681–4690
34. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. (2018) 0–0
35. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1874–1883
36. Wang, L., Wang, Y., Lin, Z., Yang, J., An, W., Guo, Y.: Learning a single network for scale-arbitrary super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 4801–4810
37. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 4401–4410
38. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. (2015) 3730–3738
39. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. (2008)
40. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 8878–8887
41. Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., Wen, F.: Bringing old photos back to life. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 2747–2757
42. Yang, L., Wang, S., Ma, S., Gao, W., Liu, C., Wang, P., Ren, P.: Hifacegan: Face renovation via collaborative suppression and replenishment. In: Proceedings of the 28th ACM International Conference on Multimedia. (2020) 1551–1560
43. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 3012–3021
44. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: Proceedings of the icee/cvf conference on computer vision and pattern recognition. (2020) 2437–2445
45. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017)
46. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. IEEE Signal processing letters 20 (2012) 209–212
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 586–595
48. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence. (2017)