IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 April 2023, accepted 27 April 2023, date of publication 10 May 2023, date of current version 17 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274841

RESEARCH ARTICLE

RPF: Reference-Based Progressive Face Super-Resolution Without Losing Details and Identity

JI-SOO KIM[1], (Student Member, IEEE), KEUNSOO KO[1], (Student Member, IEEE), HANUL KIM[2], (Member, IEEE), AND CHANG-SU KIM[1], (Fellow, IEEE)

[1]School of Electrical Engineering, Korea University, Seoul 02841, South Korea
[2]Department of Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Chang-Su Kim (changsukim@korea.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grants funded by the Korea Government (MSIT) under Grant NRF-2021R1A4A1031864, Grant NRF-2022R1A2B5B03002310, and Grant RS-2022-00166922.

ABSTRACT Face super-resolution involves generating a high-resolution facial image from a low-resolution one. It is, however, quite a difficult task when the resolution difference between input and output images is too large. In order to tackle this challenge, many approaches use generative adversarial networks that are pre-trained on a large facial image dataset, but they often generate fake details and distort the person's original face, leading to a loss of identity. Hence, in this paper, we propose a progressive face super-resolution network, called RPF, to super-resolve a facial image without losing details and personal identity by progressively exploiting the same person's high-resolution image as a reference image. First, we remove unnecessary detail information, such as hair and background, from the reference image, which may be different from the low-resolution input. Next, we align the high-resolution reference image to the low-resolution input image and blend them to generate a synthesized image. Finally, we refine the synthesized image to generate a faithful super-resolved image containing both details and identity information. Experimental results demonstrate that the proposed RPF algorithm outperforms recent state-of-the-art methods in terms of detail restoration and identity preservation, with improvements of 0.0098 and 0.0478 in LPIPS and ISC, respectively, on the CelebA-HQ dataset.

INDEX TERMS Face super-resolution, reference-based super-resolution, convolutional neural networks, generative adversarial networks.

I. INTRODUCTION

Face super-resolution (SR) is a problem of reconstructing a high-resolution (HR) facial image from a low-resolution (LR) one. Recent face SR researches [1], [2], [3], [4] have focused on a challenging scenario in which the scale factor — the resolution ratio between HR and LR images — is very large, *e.g.*, 64 as illustrated in Figure 1. This is because input images to many problems in face analysis, such as face recognition [5], face attribute recognition [6], and face alignment [7], often have very low resolutions. In order not to degrade the performance of such analysis severely, the input images should be sufficiently upsampled. However, it is not easy to restore HR images with large scale factors. A typical SR network is trained to reduce pixelwise differences between a restored HR image and the ground truth (GT), but it may suffer when input is too small and thus does not provide meaningful cues for restoration. In such a case, the network tends to produce blurry results.

To overcome the limitation, many techniques [1], [2], [3], [4] adopt generative adversarial networks (GANs) [8], [9] pre-trained on a large scale facial image dataset. They generate high-quality images by exploiting facial details learned by the pre-trained GANs. However, as shown in Figure 1, GANs may reconstruct fake details and alter the person's original face severely, resulting in identity loss. This misses

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

the objective of face SR to restore a specific person's HR facial image. Face SR should not only provide high-quality details but also preserve the person's identity.

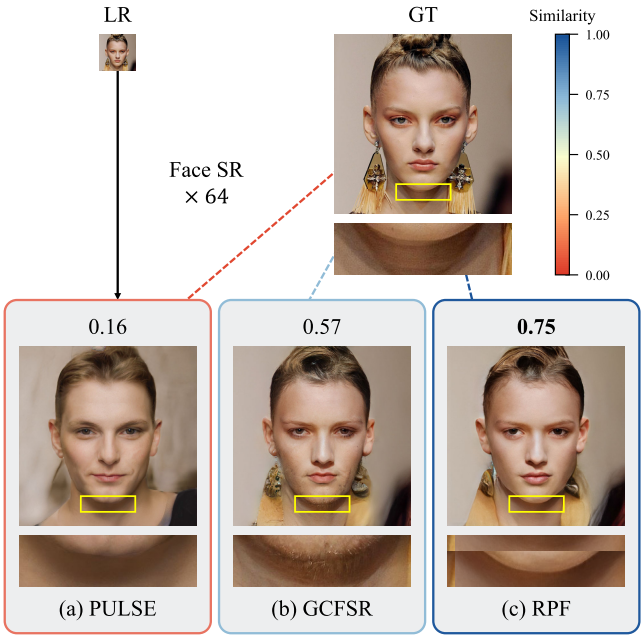


FIGURE 1. Examples of face SR results at a scale factor of 64: An input LR image is super-resolved by PULSE [1], GCFSR [2], and the proposed RPF algorithm. Each dotted line presents the similarity between the super-resolved image and GT. RPF yields the highest similarity both quantitatively and qualitatively, indicating that it preserves the identity information in GT faithfully.

There have been some attempts [10], [11], [12], [13] to exploit the same person's HR reference image containing fine details and identity information. Note that HR references are readily available in many cases. For example, someone may zoom in on her tiny face in a group photo using her HR portrait in the same personal album [13]. These reference-based techniques should align a reference image to an input image, for they may have different facial angles and expressions. However, even though they are successfully aligned, it is not straightforward to use the reference information for face SR. Non-facial regions in the reference may exhibit quite different characteristics from those in the input, depending on hairstyles, accessories, and shooting locations. If such differences are not taken into account systematically in the SR process, they may cause unpleasing artifacts in SR results.

In this paper, we propose a novel reference-based face SR algorithm to reconstruct a high-quality face image without losing the person's identity. First, we mask a reference image based on non-facial region removal to prevent adverse impacts of non-facial regions due to their variability. Second, we propose the reference-based progressive face SR (RPF) network, which gradually magnifies the input resolution by employing reconstruction blocks repeatedly. Each reconstruction block contains three modules for warping, synthesis, and refinement. First, the warping module estimates the motion field between the masked HR reference and the LR input to align the reference to the input. Second, the synthesis module reconstructs facial details finely and faithfully using the reference information. Third, the refinement module generates plausible details for non-facial regions and also further enhances facial details by fine-tuning facial regions. Extensive experiments demonstrate the effectiveness of the proposed RPF algorithm. Also, it is shown that RPF outperforms existing face SR techniques in both detail restoration and identity preservation.

This paper has the following contributions:

- We propose a novel reference-based face SR algorithm, called RPF, that reconstructs high-quality details without losing identity information.
- We propose masking reference images based on the observation that non-facial regions have different characteristics from facial ones.
- The proposed RPF algorithm meaningfully outperforms existing face SR techniques in detail restoration and identity preservation.

The remainder of this paper is organized as follows: Section II reviews related work. Section III explains the proposed RPF algorithm, and Section IV discusses experimental results. Finally, Section V draws conclusions.

II. RELATED WORK

A. FACE SR WITH SMALL SCALE FACTORS

With the great success of convolutional neural networks (CNNs), many CNN-based face SR algorithms, *e.g.* [14], [15], [16], have been proposed. Zhou et al. [14] developed the bi-channel CNN consisting of only a few layers. Yu and Porikli [15] adopted a generative adversarial loss to reconstruct realistic images. Huang et al. [16] designed a wavelet-based CNN to prevent over-smoothing.

To interpolate facial images effectively, attempts have been made to use prior knowledge of facial configurations or characteristics [14], [17], [18], [19], [20], [21], [22]. Zhu et al. [14] adopted the high-frequency prior extracted from HR face samples. Chen et al. [17] developed an end-to-end SR network using landmark and parsing maps. Also, spatial configurations of faces were extracted and used for face SR in [19], [20], and [23]. Xin et al. [18] extended the capsule network [24] to encode facial attributes such as gender, age, and beard style. Yu et al. [21] developed an attribute-embedded upscaling network to exploit facial attributes. Hu et al. [22] extracted 3D priors to achieve more reliable SR. However, these algorithms are mainly for relatively small scale factors ($\times 8$, $\times 16$).

B. FACE SR WITH LARGE SCALE FACTORS

Generative models, such as StyleGANs [8], [9], can produce detailed and realistic facial images randomly, but they cannot generate what a user wants. To exploit StyleGANs for face SR, some algorithms [1], [2], [3], [4] adopt encoders to extract features from an LR input image and use them as input to StyleGANs. For instance, Menon et al. [1] iteratively optimized latent features to obtain desired SR results.

Yang et al. [3] developed an encoder to extract variables from an LR image and then employed StyleGAN as a decoder to generate an HR image. Similarly, Chan et al. [4] extracted variables, refined the variables using StyleGAN, and used their decoder to produce an HR image. He et al. [2] proposed a controllable, as well as generative, face SR algorithm. These algorithms [1], [2], [3], [4] provide rich details for face SR, but those details may be fake ones rather than interpolated from an input image. Such fake details often distort identity information, so Wang et al. [25] fine-tuned StyleGAN with the identity preserving loss. However, since identity information cannot be used in testing, it still has the weakness of ambiguous identity restoration. In contrast, the proposed RPF algorithm uses the reference image of the same person to enforce the identity information to be used in testing as well as in training.

C. REFERENCE-BASED FACE SR

Some algorithms [10], [11], [12], [13] use an HR reference image, which can be selected from the same person's album, for face SR. Liu et al. [12] adopted a conditional variational autoencoder to restore details using the same person's HR reference. They embedded input and reference images into a joint latent space and generated an SR image using a generative decoder. Li et al. [10] warped a reference image to an input image based on optical flow and reconstructed an SR image. They used the information in the entire reference image, including non-facial regions such as hair, accessories, and background, for the SR reconstruction, but the non-facial regions may be quite different from those in the input image, causing visual artifacts. For more reliable SR, Li et al. [11] selected the most similar reference image among multiple candidates. Kim et al. [13] partitioned both input and reference feature maps into patches and matched the most similar pairs of input and reference patches to restore details. These reference-based SR algorithms, however, may cause unpleasing artifacts in non-facial regions. Furthermore, they are designed to use reference HR images at relatively small scale factors ($\times 8$, $\times 16$).

III. PROPOSED ALGORITHM

Suppose that an LR image $\mathbf{x} \in \mathbb{R}^{L\times L\times 3}$ is a degraded version of an original HR image $\mathbf{y} \in \mathbb{R}^{\kappa L\times \kappa L\times 3}$, where κ is the scale factor. In this paper, face photos have the aspect ratio of $1:1$, so the spatial resolution of $\mathbf{x}$ is denoted by $L \times L$. Also, it is assumed that a reference image $\mathbf{r} \in \mathbb{R}^{\kappa L\times \kappa L\times 3}$ is given, which differs from $\mathbf{y}$ but contains the same person's face. In reference-based face SR, we aim to restore the HR image $\mathbf{y}$ from the LR one $\mathbf{x}$ by exploiting the information in the reference $\mathbf{r}$.

Figure 2 shows an overview of the proposed RPF algorithm. Through the non-facial region removal, we obtain the masked reference image $\hat{\mathbf{r}}$ from $\mathbf{r}$. Then, we gradually magnify $\mathbf{x}$ with $\hat{\mathbf{r}}$ through multiple reconstruction blocks.

A. NON-FACIAL REGION REMOVAL

A face photo is composed of facial regions and non-facial regions, such as hair, accessories, and background. In comparison with non-facial regions, the facial regions, containing the identity information, are relatively invariant across different photos. In general, the facial regions in two photos of the same person are different due to the misalignment of facial landmarks, which are caused by different facial angles or expressions. Such differences can be compensated by an alignment process. In contrast, even for the same person, non-facial regions vary greatly depending on hairstyles, fashion items, or shooting locations. Therefore, rather than being helpful for SR, non-facial regions often cause visual artifacts in SR. It is hence essential to erase non-facial regions from a reference image to achieve face SR reliably.

We develop a scheme for non-facial region removal illustrated in Figure 3. We parse a reference image $\mathbf{r}$ into 18 classes (*e.g.*, skin, nose, and hair) using the bilateral segmentation network [26]. We then generate a binary face mask $\mathbf{m}$ based on the semantics of those 18 classes. Specifically, we assume that the five classes of 'neck,' 'hat,' 'hair,' 'necklace,' and 'clothing' are non-facial, while the others are facial. We set the mask value to 1 if a pixel belongs to one of the facial classes, and 0 otherwise. Finally, we obtain the masked reference image $\hat{\mathbf{r}}$ by

$$\hat{\mathbf{r}} = \mathbf{m} \odot \mathbf{r} \tag{1}$$

where $\odot$ denotes the Hadamard product.

B. RPF NETWORK

As shown in Figure 2, we progressively restore an HR image with a scale factor $\kappa = 2^N$ through the proposed RPF network consisting of N reconstruction blocks. Each reconstruction block doubles the resolution. When a scale factor is large, it is difficult to train a neural network that produces an HR image at once. It is known that the progressive approach facilitates more reliable training at a large scale factor [27], [28]. In Section IV, we also validate the effectiveness of the progressive approach experimentally.

For progressive face SR, we downsample the masked reference image by a factor of 2 repeatedly and feed them to the corresponding reconstruction blocks, as shown in Figure 2(a). Then, each reconstruction block in Figure 2(b) increases the resolution by a factor of 2 using the warping, synthesis, and refinement modules. More specifically, the nth reconstruction block processes an image $\mathbf{x}^{(n)} \in \mathbb{R}^{2^{n-1}L\times 2^{n-1}L\times 3}$ and the corresponding masked reference $\hat{\mathbf{r}}^{(n)} \in \mathbb{R}^{2^nL\times 2^nL\times 3}$ to yield an output image $\mathbf{x}^{(n+1)} \in \mathbb{R}^{2^nL\times 2^nL\times 3}$, where $1 \le n \le N$. Note that $\mathbf{x}^{(1)}$ is the LR input $\mathbf{x}$, and $\mathbf{x}^{(N+1)}$ is the SR output. Let us describe each module subsequently.

1) WARPING MODULE

To super-resolve the input $\mathbf{x}^{(n)}$ faithfully, we first align the masked reference $\hat{\mathbf{r}}^{(n)}$ to $\mathbf{x}^{(n)}$ using the warping module in Figure 4. The warping module estimates the motion field

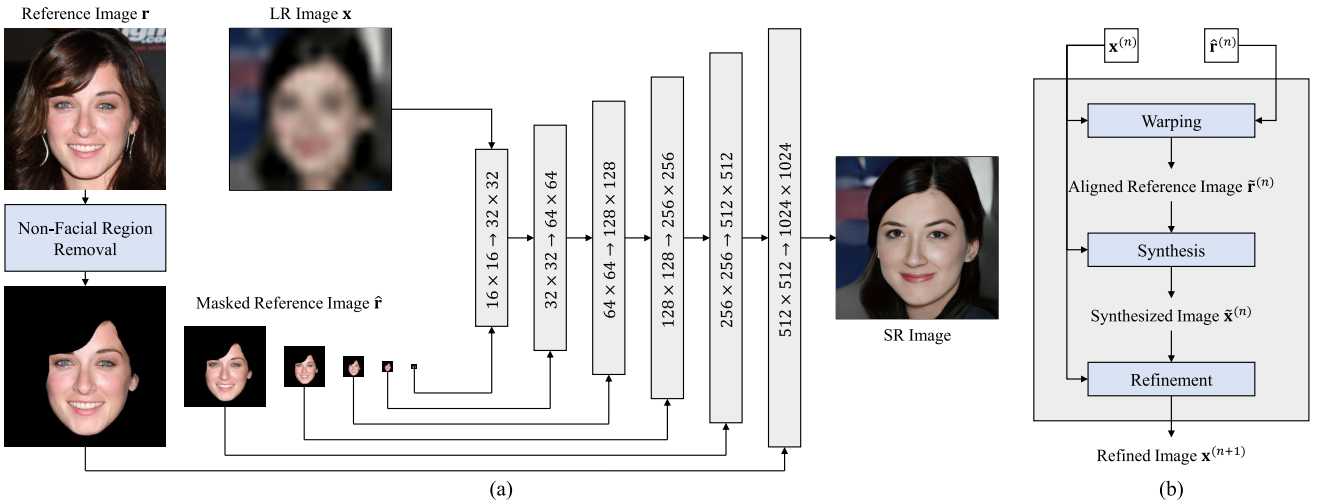


FIGURE 2. (a) An overview of the proposed RPF algorithm. In this example, the LR input x has the spatial resolution of 16 × 16, which is super-resolved to 1024 × 1024 using six reconstruction blocks. For visualization, x is resized to the same size as the SR result via bicubic interpolation. (b) The *n*th reconstruction block.

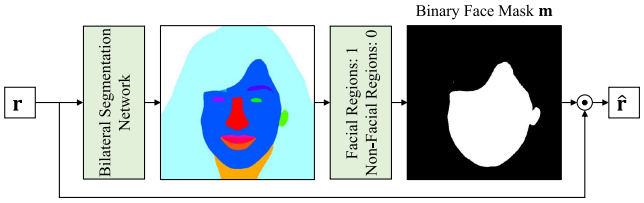


FIGURE 3. Illustration of the non-facial region removal process.

$\mathbf{u}_{\mathbf{x}\to\hat{\mathbf{r}}}$, which indicates the matching pixel in $\hat{\mathbf{r}}^{(n)}$ corresponding to each pixel in $\mathbf{x}^{(n)}$. However, the spatial resolution of $\mathbf{x}^{(n)}$ is $2^{n-1}L \times 2^{n-1}L$, while that of $\hat{\mathbf{r}}^{(n)}$ is $2^nL \times 2^nL$. Thus, we first double the resolution of $\mathbf{x}^{(n)}$ through a convolution layer and a pixel shuffle layer. Then, we concatenate the input and reference information and perform the optical flow estimation.

Similar to U-Net [29], the warping module contains the encoder, composed of downscale blocks, and the decoder, composed of upscale blocks. In Figure 4, each downscale block halves the spatial resolution using an ordinary convolution layer and a stride convolution layer. The number of downscale blocks is determined so that the encoder output, *i.e.*, feature map, has the spatial resolution of 8 × 8. Then, through the same number of upscale blocks, each of which contains a convolution layer and two deconvolution layers, we gradually refine the feature map to obtain the motion field $\mathbf{u}_{\mathbf{x}\to\hat{\mathbf{r}}}$ of resolution $2^nL \times 2^nL$. Finally, we obtain the aligned reference image $\tilde{\mathbf{r}}^{(n)}$ by

$$\tilde{\mathbf{r}}^{(n)} = \phi_{\mathrm{B}}(\hat{\mathbf{r}}^{(n)}, \mathbf{u}_{\mathbf{x}\to\hat{\mathbf{r}}}) \tag{2}$$

where ϕ_{B} is the backward warping operator [30].

2) SYNTHESIS MODULE

The synthesis module aims to restore fine details of facial regions in $\mathbf{x}^{(n)}$ using the aligned reference image $\tilde{\mathbf{r}}^{(n)}$ in (2). Figure 5 shows the structure of the synthesis module. First, we double the resolution of $\mathbf{x}^{(n)}$ through a convolution layer and a pixel shuffle layer and process $\tilde{\mathbf{r}}^{(n)}$ through another convolution layer. Since both input and reference are aligned, we simply concatenate them and use four residual blocks to transfer high-frequency information in the reference to the input and finally yield the synthesized image $\tilde{\mathbf{x}}^{(n)}$. Note that the synthesis module focuses on the detailed reconstruction of facial regions.

3) REFINEMENT MODULE

Even though we can restore facial regions faithfully through the synthesis module, it is not straightforward to reconstruct missing details of non-facial regions because no reliable information is available in the reference image. Hence, we design the refinement module to generate plausible details for non-facial regions. To this end, we adopt the encoder-decoder architecture [2]. The encoder extracts features from the input image $\mathbf{x}^{(n)}$ and the synthesized one $\tilde{\mathbf{x}}^{(n)}$, and then the decoder processes the features to yield the final reconstruction $\mathbf{x}^{(n+1)}$. Specifically, the encoder is composed of K downscale blocks, each of which halves the spatial resolution using an ordinary convolution layer and a stride convolution layer, and an embedding block containing two stride convolution layers and K fully connected (FC) layers, as shown in Figure 6 (a). Here, K is determined so that the intermediate feature map $\mathbf{e}_1$ has the spatial resolution of 4 × 4. As shown in Figure 6 (b), the decoder consists of K upscale blocks, each of which doubles the resolution via style modulation and feature modulation [2]. In this work, we train multiple feature modulation models: one for each specific resolution.

In the refinement module within the *n*th reconstruction block, we first upsample the input image $\mathbf{x}^{(n)}$ via bicubic interpolation. Then, from the average of the upsampled image and the synthesized one $\tilde{\mathbf{x}}^{(n)}$, we extract the intermediate

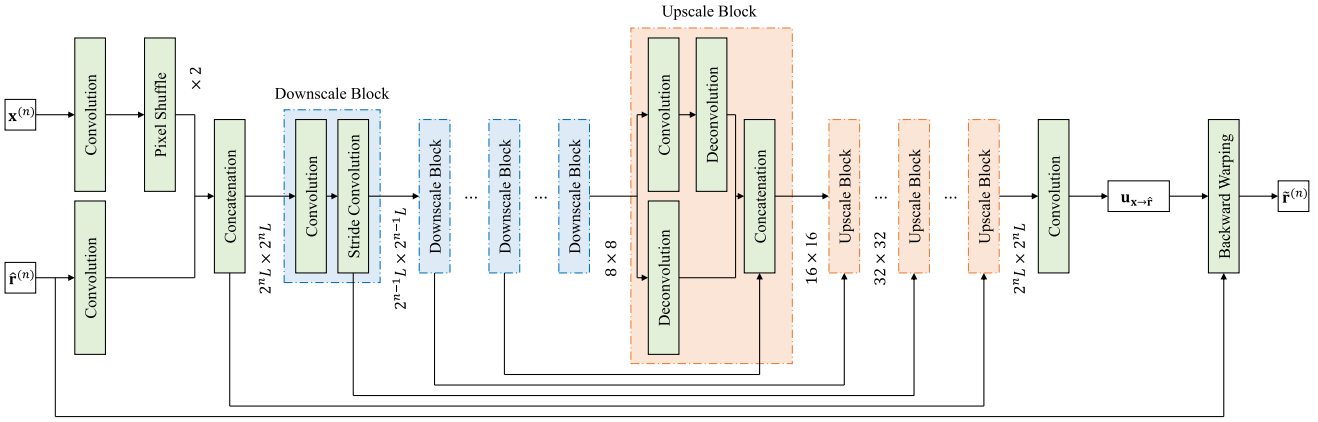


FIGURE 4. The network structure of the warping module.

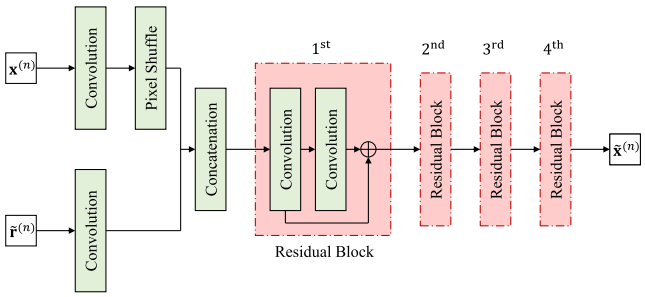


FIGURE 5. The network structure of the synthesis module.

feature maps $\{\mathbf{e}_k\}_{k=1}^{K}$ and the latent vectors $\{\boldsymbol{v}_k\}_{k=1}^{K}$. Next, through K upscale blocks in the decoder, we obtain the feature map $\mathbf{d}_K$, which is processed by the last convolution layer to yield $\mathbf{x}^{(n+1)}$. Specifically, at the kth upscale block, the style modulation outputs the feature map $\mathbf{s}_k$ from $\mathbf{d}_{k-1}$ and $\boldsymbol{v}_k$, where $\mathbf{d}_0 = \mathbf{e}_1$. Then the feature modulation blends $\mathbf{s}_k$ and $\mathbf{e}_k$ using learnable mixing coefficients α_k and β_k to construct the feature map $\mathbf{d}_k$, as shown in Figure 6 (c).

For example, Figure 7 shows intermediate results for super-resolving an image.

C. LOSS FUNCTIONS

To train the three modules in each reconstruction block, we design the warping loss $\mathcal{L}_{\text{warp}}$, the synthesis loss $\mathcal{L}_{\text{syn}}$, and the refinement loss $\mathcal{L}_{\text{ref}}$. We do sequential training of multiple reconstruction blocks in RPF, starting with the first reconstruction block: we train each reconstruction block after freezing the parameters of the previous blocks. Below we describe the losses for training the nth reconstruction block. Note that GT is also downsampled to the corresponding spatial resolution via bicubic interpolation.

1) WARPING LOSS

To define the warping loss $\mathcal{L}_{\text{warp}}$, we construct a warped reference $\mathbf{r}_{\text{warp}}$ by warping an HR reference $\mathbf{r}$ to GT $\mathbf{y}$, as shown in Figure 8. To this end, we extract facial landmarks from each image using the landmark extractor in [31]. Based on the matching pairs of landmarks, we compute the thin-plate spline warping (TPS) function [32], denoted by $f(\cdot)$, and use it to warp only facial regions in $\mathbf{r}$ to yield $\mathbf{r}_{\text{warp}}$. Here, non-facial regions are removed as stated in Section III-A. Figure 9 shows that the facial configurations of $\mathbf{r}_{\text{warp}}$ are well aligned with those of $\mathbf{y}$. Next, we define the warping loss

$$\mathcal{L}_{\text{warp}} = \left\| \tilde{\mathbf{r}}^{(n)} - \mathbf{r}_{\text{warp}}^{(n)} \right\|_1 \tag{3}$$

where $\tilde{\mathbf{r}}^{(n)}$ is a warped image by the warping module and $\mathbf{r}_{\text{warp}}^{(n)}$ is obtained by downsampling $\mathbf{r}_{\text{warp}}$.

2) SYNTHESIS LOSS

The synthesis loss $\mathcal{L}_{\text{syn}}$ is defined as

$$\mathcal{L}_{\text{syn}} = \mathcal{L}_{\text{rec(s)}} + \alpha \mathcal{L}_{\text{adv(s)}} \tag{4}$$

where $\mathcal{L}_{\text{rec(s)}}$ is a reconstruction loss and $\mathcal{L}_{\text{adv(s)}}$ is an adversarial loss. Here, we set the hyperparmeter α as 0.1. In the synthesis module, we focus on facial regions only to prevent negative impacts of non-facial regions in a reference image. Hence, we downsample the binary face mask $\mathbf{m}$ in Section III-A to $\mathbf{m}^{(n)}$ and then define the reconstruction loss

$$\mathcal{L}_{\text{rec(s)}} = \left\| \mathbf{m}^{(n)} \odot \tilde{\mathbf{x}}^{(n)} - \mathbf{m}^{(n)} \odot \mathbf{y}^{(n)} \right\|_1 \tag{5}$$

to penalize synthesis errors in the facial regions. Also, we adopt the adversarial loss $\mathcal{L}_{\text{adv(s)}}$ to enhance subjective qualities by emphasizing high-frequency information. Specifically, we use the Wasserstein loss [33], [34] as $\mathcal{L}_{\text{adv(s)}}$, where $\mathbf{m}_{\mathbf{y}}^{(n)} \odot \tilde{\mathbf{x}}^{(n)}$ and $\mathbf{m}_{\mathbf{y}}^{(n)} \odot \mathbf{y}^{(n)}$ are used as the input.

3) REFINEMENT LOSS

We define the refinement loss $\mathcal{L}_{\text{ref}}$ as a combination of the reconstruction loss $\mathcal{L}_{\text{rec(r)}}$, perceptual loss $\mathcal{L}_{\text{per}}$, identity loss $\mathcal{L}_{\text{id}}$, and adversarial loss $\mathcal{L}_{\text{adv(r)}}$,

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{rec(r)}} + \mathcal{L}_{\text{per}} + \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{adv(r)}} \tag{6}$$

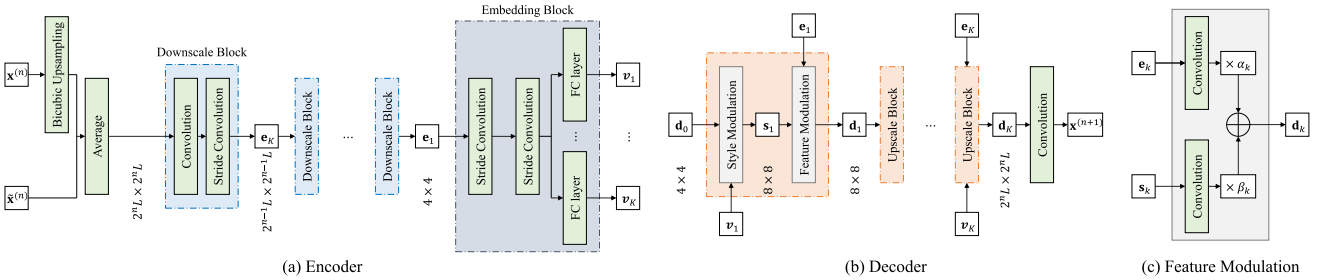


FIGURE 6. The network structure of the refinement module, composed of an encoder and a decoder, in the *n*th reconstruction block.

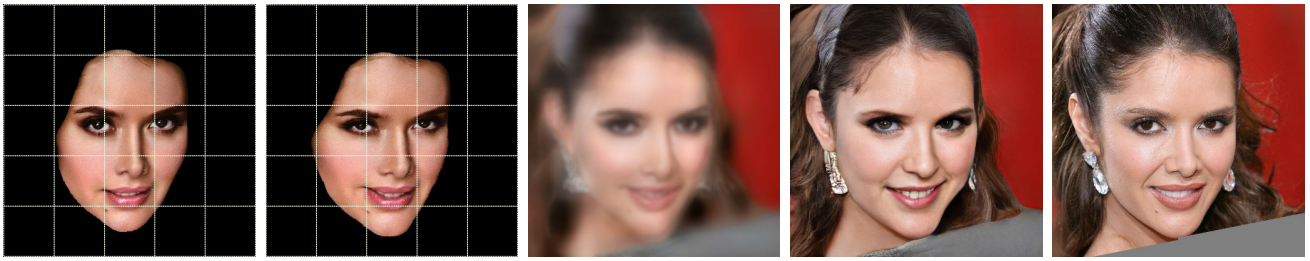


FIGURE 7. Intermediate results of a reconstruction block: The reference image r in (a) is aligned by the warping module, resulting in the aligned reference $\tilde{\mathbf{r}}^{(n)}$ in (b). It is fed into the synthesis module to generate the synthesized image $\tilde{\mathbf{x}}^{(n)}$ in (c). Finally, the refinement module provides the final reconstruction $\mathbf{x}^{(n+1)}$ in (d), which is similar to GT in (e).

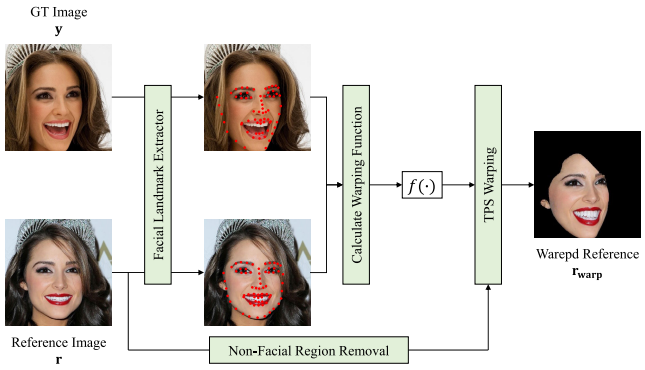


FIGURE 8. For the warping loss $\mathcal{L}_{\text{warp}}$ in training, a warped reference $\mathbf{r}_{\text{warp}}$ is obtained by warping a reference image r to the GT y.

FIGURE 9. Examples of the GT images y, HR reference images r, and warped references $\mathbf{r}_{\text{warp}}$.

where the reconstruction loss is given by

$$\mathcal{L}_{\text{rec(r)}} = \|\mathbf{x}^{(n+1)} - \mathbf{y}^{(n)}\|_1. \tag{7}$$

Also, the perceptual loss

$$\mathcal{L}_{\text{per}} = \|\mathbf{f}_{\mathbf{x}^{(n+1)}} - \mathbf{f}_{\mathbf{y}^{(n)}}\|_2 \tag{8}$$

measures the similarity between deep features $\mathbf{f}_{\mathbf{x}^{(n+1)}}$ and $\mathbf{f}_{\mathbf{y}^{(n)}}$, which are extracted from $\mathbf{x}^{(n+1)}$ and $\mathbf{y}^{(n)}$ through AlexNet [35]. It helps to yield better perceptual qualities [34], [36]. Similarly, the identity loss $\mathcal{L}_{\text{id}}$ is defined as

$$\mathcal{L}_{\text{id}} = \|\boldsymbol{i}_{\mathbf{x}^{(n+1)}} - \boldsymbol{i}_{\mathbf{y}^{(n)}}\|_1 \tag{9}$$

where $\boldsymbol{i}_{\mathbf{x}^{(n+1)}}$ and $\boldsymbol{i}_{\mathbf{y}^{(n)}}$ are extracted from the ArcFace network [37] for face recognition. Since these features are trained to discern different people [2], $\mathcal{L}_{\text{id}}$ helps to preserve the person's identity. Finally, we also adopt the Wasserstein loss for $\mathcal{L}_{\text{adv(r)}}$ but use $\mathbf{x}^{(n+1)}$ and $\mathbf{y}^{(n)}$ as the input.

IV. EXPERIMENTS

A. DATASETS AND METRICS

1) CelebA-HQ [27]

It contains 30,000 facial images of 6,217 people. Among them, for reference-based face SR, we select 28, 278 pairs of original and reference images. Specifically, we sample reference candidates for each original image, satisfying the same identity constraint [13]. We then choose the reference image

among the candidates, whose facial landmarks are the closest to those in the original image. To measure the distances between facial landmarks, we adopt the average L1 distance. Also, we extract those landmarks using the algorithm in [31]. We split the collected pairs into 27,291 pairs for training and 987 pairs for testing.

2) FFHQ [8]

It consists of 70,000 facial images, which are more diverse than CelebA-HQ in terms of ages, ethnicities, accessories, and shooting locations. However, since it provides no identity information, it is used only for training the refinement module.

3) EVALUATION METRICS

For quantitative comparisons of face SR results, we measure the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS) [38], and identity similarity score (ISC) [2], which measures the preservation of identity information quantitatively.

B. IMPLEMENTATION DETAILS

We train the proposed RPF network in two stages. First, we train the refinement modules of all reconstruction blocks on FFHQ. For each refinement module, we use a resized LR image via bicubic interpolation as the input and the resized GT as the output. Thus, in this stage, we do not require reference images. Second, we train the warping and synthesis modules in each reconstruction block using CelebA-HQ. The training is performed sequentially from the coarsest to the finest reconstruction blocks. For efficient training, after fixing the parameters of the previous reconstruction blocks, we train a current reconstruction block. Similarly, within each reconstruction block, after training and fixing the warping module, we train the synthesis module. Each module is trained for five epochs.

In all modules except the refinement modules, we fix the feature dimension (*i.e.*, the number of channels) to $C = 64$. The refinement modules are implemented based on GCFSR [2]. In each convolution layer, we perform zero padding and adopt the ReLU activation. We exclude batch normalization since we use a small mini-batch of size 1. We use the Adam optimizer [40] with a learning rate of 10^{-4}. The training is performed with an RTX 3090 GPU.

C. COMPARISONS

In Table 1, we compare the proposed RPF algorithm with recent state-of-the-art face super-resolution (SR) algorithms, including PULSE [1], mGANprior [39], GLEAN [4], GPEN [3], and GCFSR [2], at large scale factors $\kappa = 32$ and 64. The results were obtained by executing the available codes of these algorithms. Note that RPF meaningfully outperforms the existing algorithms in all experiments and metrics. The performance gap between RPF and the second-best algorithm GCFSR is larger at the larger scale factor $\kappa = 64$, demonstrating the advantages of the proposed algorithm in restoring details when the input is too small. This is attributed to the exploitation of HR reference images in the proposed algorithm.

Figure 10 shows qualitative results for the proposed RPF and state-of-the-art face SR algorithms at $\kappa = 64$. We see that, while the existing algorithms generate detailed facial images, they fail to preserve the identity information. In contrast, the proposed RPF provides more faithful results, preserving the identity information by employing reference images. It is noteworthy that the reference and input images in our experiments have different facial expressions and angles. Despite these differences, RPF effectively transfers the detail information from the reference images to super-resolve the extremely LR input images. These results demonstrate the effectiveness of the proposed RPF algorithm in generating high-quality SR results while preserving identity information.

We then compare the proposed algorithm with CollageNet [13], which is a recent algorithm for reference-based face SR. Since CollageNet is designed for scale factors 2, 4, 8, and 16, we employ CollageNet twice in this test to assess its performance at higher scale factor of 32 and 64. As shown in Table 1, the proposed RPF outperforms CollageNet in LPIPS and ISC metrics, demonstrating the effectiveness of RPF in restoring details and preserving identity information. Note that the traditional metrics PSNR and SSIM are less suitable for assessing SR results at such high scale factors ($\times 32$, $\times 64$). As shown in Figure 10, they may favor smoothed results that are perceptually inferior. Thus, the performance on perceptual metrics, such as LPIPS and ISC, is more crucial in accurately evaluating the quality of high-scale SR results.

In Table 1, we also compare the proposed RPF algorithm with MPRNet [28], a recent image restoration algorithm. Since MPRNet is originally designed for deblurring, we upsample an input image via bicubic interpolation and then feed it to MPRNet to obtain a high-resolution image. We see that the proposed RPF outperforms MPRNet at all scale factors in terms of all metrics.

D. ANALYSIS

Next, we perform several experiments on CelebA-HQ to analyze the proposed RPF algorithm. Here, we fix the scale factor κ to 64.

1) PROGRESSIVE RECONSTRUCTION

Table 2 compares the proposed algorithm with the alternative method that increases the resolution by a factor of 64 at once, instead of the progressive reconstruction using multiple reconstruction blocks. We see that the progressive reconstruction significantly improves the SR results.

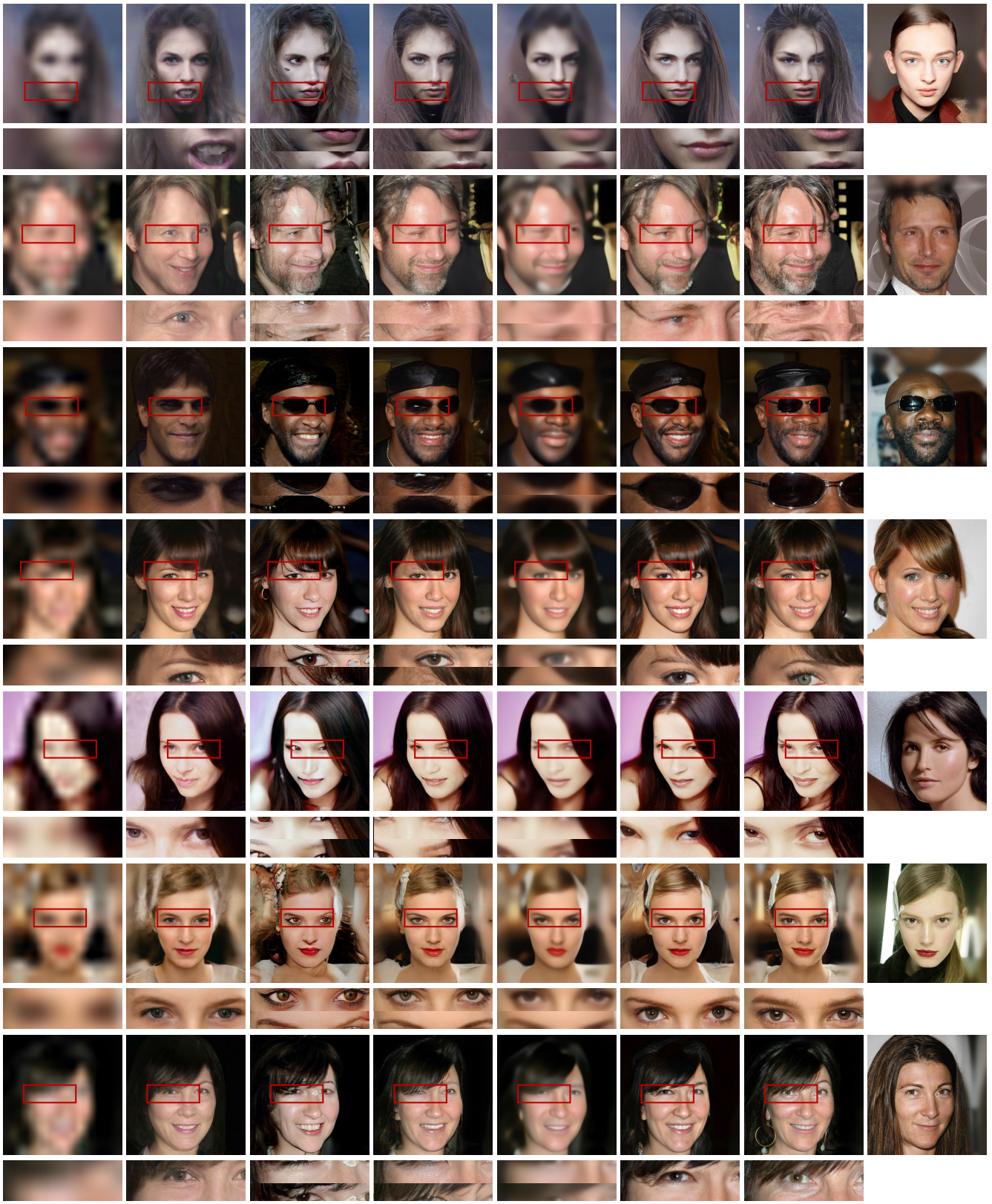


FIGURE 10. Qualitative comparison of face SR results at the scale factor $\kappa = 64$ on CelebA-HQ: Input images in (a) are super-resolved by PULSE [1] in (b), GPEN [3] in (c), GCFSR [2] in (d), CollageNet [13] in (e), and RPF in (f). GT and reference images are in (g) and (h), respectively.

TABLE 1. PSNR, SSIM, LPIPS, and ISC performance comparison on CelebA-HQ at scale factors $\kappa = 32$ and $\kappa = 64$. The best result is boldfaced, and the second best one is underlined.

	×32				×64			
	PSNR↑	SSIM↑	LPIPS↓	ISC↑	PSNR↑	SSIM↑	LPIPS↓	ISC↑
PULSE [1]	19.78	0.5732	0.5432	0.4923	17.53	0.5306	0.6068	0.4068
mGANprior [39]	21.26	0.6117	0.5099	0.5230	18.69	0.5721	0.5530	0.4397
GLEAN [4]	24.34	0.6534	0.3257	0.7750	21.38	0.6016	0.4109	0.6118
GPEN [3]	23.65	0.6417	0.3340	0.7541	20.20	0.5991	0.4306	0.5978
GCFSR [2]	24.64	0.6786	0.3064	0.7837	22.19	0.6345	0.3689	0.6623
CollageNet [13]	**25.71**	**0.7003**	0.5046	0.5114	**24.56**	**0.7926**	0.5696	0.4157
MPRNet [28]	21.97	0.6201	0.5073	0.5301	18.93	0.5846	0.5513	0.4279
RPF	25.03	0.6913	**0.3003**	**0.8063**	23.16	0.6980	**0.3591**	**0.7101**

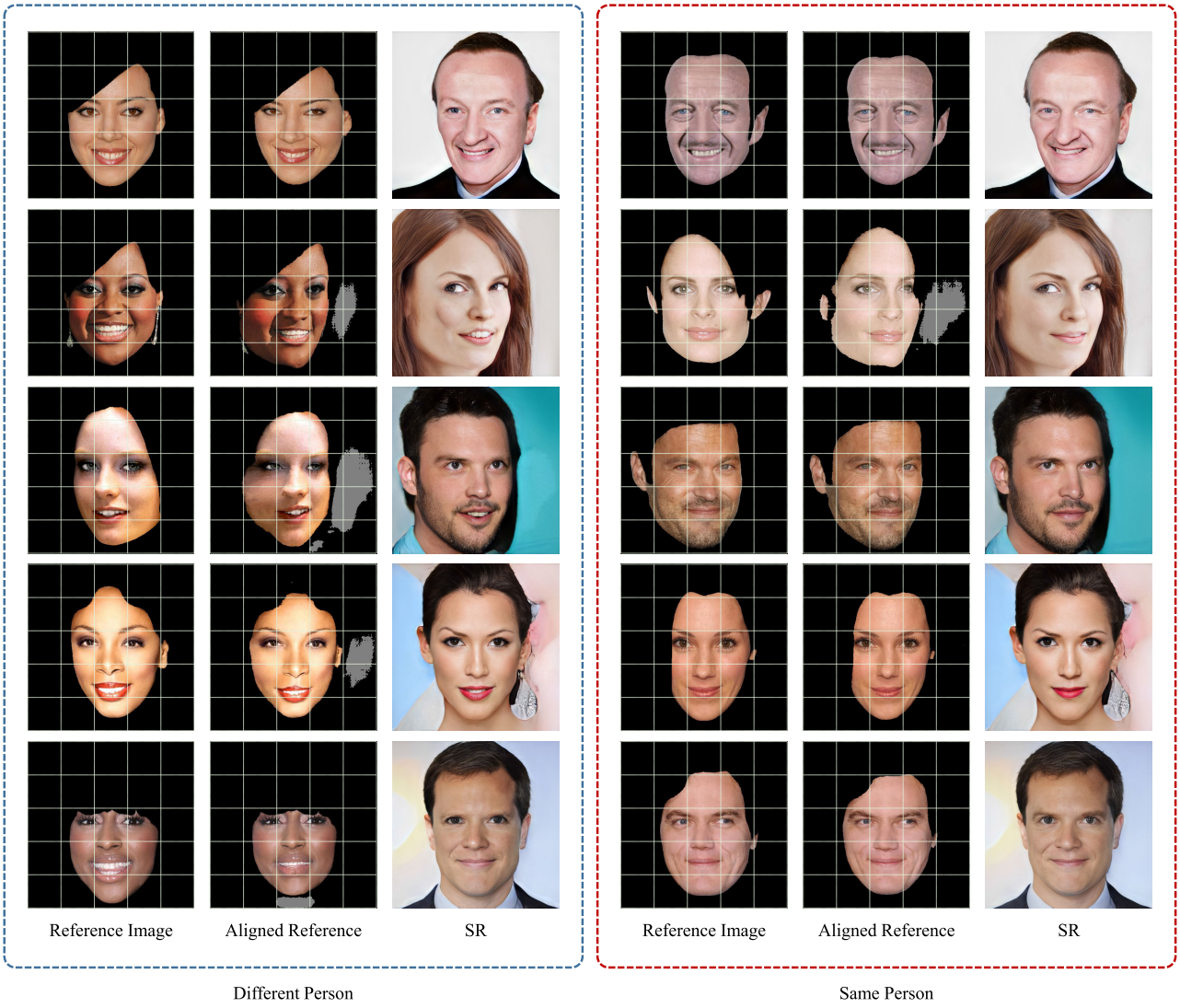
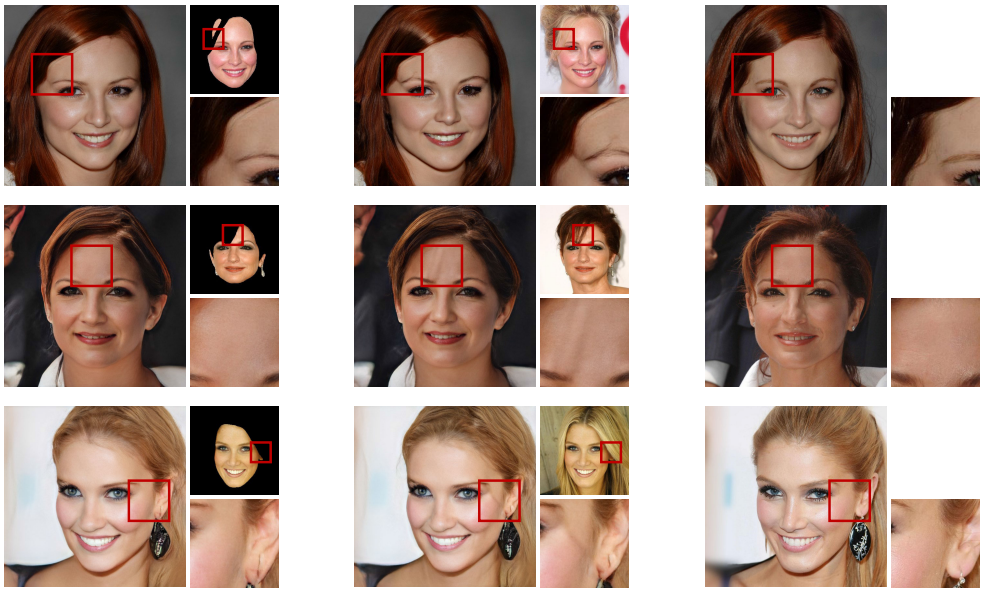


FIGURE 11. SR results using diverse reference images. In the left half, each reference image is selected from a different person's photos. In contrast, in the right half, each reference is the same person's photo.

2) REFERENCE IMAGES

We replace HR reference images with resized input images via bicubic interpolation to analyze the contribution of HR reference images. Table 2 shows that reference images are essential for faithful face SR. Especially, they increase the ISC score considerably, indicating that the proposed RPF



(a) Original Setting (b) w/o Non-Face Region Removal (c) GT

FIGURE 12. SR images are obtained by the proposed RPF algorithm in (a) and by the ablated method without the non-facial region removal process in (b). GT images are shown in (c).

transfers the identity information in reference images to input images effectively.

3) DIVERSE REFERENCE IMAGES

Figure 11 shows face SR results using diverse reference images from the same people or even from different people. We see that the proposed RPF generates quite different SR results according to the references, which indicates that RPF uses the information in reference images effectively. However, despite of those differences, the proposed RPF generates reliable SR images.

4) NON-FACIAL REGION REMOVAL

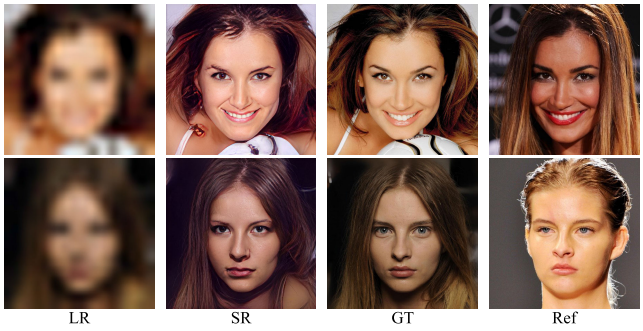
We leave out the non-facial region removal for HR reference images. This degrades the PSNR score severely due to visual artifacts in non-facial regions, as illustrated in Figure 12.

5) REAL-WORLD DEGRADATION MODEL

As done in [41] and [42], we evaluate the robustness of the proposed RPF algorithm using the real-world degradation model in [41], which generates diverse LR images using blur kernels and noise. Figure 13 shows that the proposed RPF super-resolves these LR images as well.

TABLE 2. PSNR, LPIPS, and ISC performance comparison of ablated methods on CelebA-HQ at the scale factor $\kappa = 64$.

	PSNR↑	LPIPS↓	ISC↑
w/o Progressive Reconstruction	22.58	0.3659	0.6649
w/o Reference Image	22.90	0.3623	0.6631
w/o Non-Facial Region Removal	22.39	0.3678	0.6820
RPF	**23.16**	**0.3591**	**0.7101**



LR SR GT Ref

FIGURE 13. The proposed RPF algorithm super-resolves degraded LR images by exploiting reference images. The real-world degradation model in [41] is used in this test.

V. CONCLUSION

We proposed a novel algorithm, called RPF, for reference-based face super-resolution without losing details and identity. The algorithm consists of three key components: warping, synthesis, and refinement modules. The warping module aligns a high-resolution reference image with a low-resolution input image. The synthesis module utilizes facial information from the HR reference image to restore both details and identity in the LR input. Lastly, the refinement module further enhances the facial regions of the synthesized image and generates realistic details in non-facial areas. Experimental results demonstrated that the proposed PRF outperforms existing algorithms quantitatively and yields high-quality SR results by leveraging the information contained in HR reference images.

The proposed RPF algorithm has a limitation that it is designed to use only one reference image and thus ignores

other potential sources of improvement. Further research could explore the use of multiple reference images to overcome this limitation. Also, it is worth noting that the proposed RPF is specifically designed for facial SR, so it is evaluated solely on facial images. Therefore, another future research issue is to generalize RPF for the SR of general images.

REFERENCES

[1] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2437–2445.

[2] J. He, W. Shi, K. Chen, L. Fu, and C. Dong, "GCFSR: A generative and controllable face super resolution method without facial and GAN priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1889–1898.

[3] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 672–681.

[4] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14245–14254.

[5] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of LFW benchmark or not?" 2015, *arXiv:1501.04690*.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[7] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[10] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *Proc. ECCV*, 2018, pp. 272–289.

[11] X. Li, W. Li, D. Ren, H. Zhang, M. Wang, and W. Zuo, "Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2706–2715.

[12] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Reference based face super-resolution," *IEEE Access*, vol. 7, pp. 129112–129126, 2019.

[13] J.-S. Kim, K. Ko, and C.-S. Kim, "Gluing reference patches together for face super-resolution," *IEEE Access*, vol. 9, pp. 169321–169334, 2021.

[14] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. ECCV*, 2016, pp. 614–630.

[15] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. ECCV*, 2016, pp. 318–333.

[16] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1689–1697.

[17] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2492–2501.

[18] J. Xin, N. Wang, X. Jiang, J. Li, X. Gao, and Z. Li, "Facial attribute capsules for noise face super resolution," in *Proc. AAAI*, 2020, pp. 12476–12483.

[19] Y. Yin, J. Robinson, Y. Zhang, and Y. Fu, "Joint super-resolution and alignment of tiny faces," in *Proc. AAAI*, 2020, pp. 12693–12700.

[20] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5569–5578.

[21] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 908–917.

[22] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, "Face super-resolution guided by 3D facial priors," in *Proc. ECCV*, 2020, pp. 763–780.

[23] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. ECCV*, 2018, pp. 217–233.

[24] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. ICANN*, 2011, pp. 44–51.

[25] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9168–9178.

[26] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, 2018, pp. 325–341.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018, pp. 1–9.

[28] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[30] G. Wolberg, *Digital Image Warping*. IEEE Computer Society Press, 1990.

[31] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[32] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 1–11.

[34] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7982–7991.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–9.

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[37] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[39] J. Gu, Y. Shen, and B. Zhou, "Image processing using multi-code GAN prior," in *Proc. CVPR*, 2020, pp. 3012–3021.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[41] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.

[42] R. Zhou and S. Susstrunk, "Kernel modeling super-resolution on real low-resolution images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2433–2443.

JI-SOO KIM (Student Member, IEEE) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2018, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include image processing and machine learning.

KEUNSOO KO (Student Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2017 and 2023, respectively. He is currently a Postdoctoral Fellow with the Basic Research Laboratory for Order Learning, Korea University. His research interests include image processing, computer vision, and machine learning.

HANUL KIM (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2014 and 2020, respectively. From 2020 to 2021, he was a Senior Engineer with Qualcomm AI Research. In July 2021, he joined the Department of Applied Artificial Intelligence, Seoul National University of Science and Technology, as an Assistant Professor. His research interests include computer vision and machine learning.

CHANG-SU KIM (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Seoul National University (SNU). From 2000 to 2001, he was a Visiting Scholar with the Signal and Image Processing Institute, University of Southern California, Los Angeles. From 2001 to 2003, he coordinated the 3-D Data Compression Group, National Research Laboratory for 3-D Visual Information Processing, SNU. From 2003 to 2005, he was an Assistant Professor with the Department of Information Engineering, The Chinese University of Hong Kong. In September 2005, he joined the School of Electrical Engineering, Korea University, where he is currently a Professor. He has published more than 320 journals and conference papers. His research interests include image processing, computer vision, and machine learning. He was a member of the Multimedia Systems and Application Technical Committee (MSATC) of the IEEE Circuits and Systems Society. In 2009, he received the IEEK/IEEE Joint Award for Young IT Engineer of the Year. In 2014, he received the Best Paper Award from *Journal of Visual Communication and Image Representation* (JVCI). During his Ph.D. study, he received the Distinguished Dissertation Award, in 2000. He served as an editorial board member for *JVCI* and an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON MULTIMEDIA. He is a Senior Area Editor of *JVCI*. He was an APSIPA Distinguished Lecturer, from 2017 to 2018.