Gluing Reference Patches Together for Face Super-Resolution

JI-SOO KIM, (Student Member, IEEE), KEUNSOO KO, (Student Member, IEEE), AND CHANG-SU KIM, (Senior Member, IEEE)

School of Electrical Engineering, Korea University, Seoul 136-701, South Korea

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

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2018R1A2B3003896 and No. NRF-2021R1A4A1031864).

ABSTRACT

Face super-resolution is a domain-specific super-resolution task to generate a high-resolution facial image from a low-resolution one. In this paper, we propose a novel face super-resolution network, called CollageNet, to super-resolve an input image by exploiting a reference image of an identical person at the patch level. First, we extract feature pyramids from input and reference images to exploit multi-scale information hierarchically. Next, we compute the patch-wise similarities between input and reference feature pyramids and select the $K$ most similar reference patches to each input patch. Then, we compose a collaged feature pyramid by gluing those selected patches together. Finally, we obtain a super-resolved image by blending the collaged feature pyramid and the input feature. Experimental results demonstrate that the proposed CollageNet yields state-of-the-art performances.

INDEX TERMS

Face super-resolution, reference based super-resolution, convolutional neural network, and patch matching.

I. INTRODUCTION

The objective of face super-resolution (SR) is to generate a high-resolution (HR) facial image from a low-resolution (LR) one. It can be used in various tasks, such as face recognition [1], face attribute recognition [2], face alignment [3], old photo restoration [4], face re-identification [5], and video surveillance [6], in which an input image is often of low quality.

Recently, with the success of convolutional neural network (CNN) in the field of low level vision [7]–[10], a wide variety of face SR algorithms have been proposed. Especially, recent CNN-based algorithms use big training datasets to overcome the ill-posedness of the SR problem — an LR image has multiple HR counterparts [11]–[14]. Some algorithms attempt to improve face SR performance by employing adversarial losses and/or external data (e.g., facial attributes or reference images) [13], [15], [16]. However, their results often suffer from a lack of details, causing blurry artifacts.

In this paper, we super-resolve a person’s LR facial image by exploiting the same person’s HR reference image. This is useful in some applications. For example, someone may want to zoom in on his or her tiny face in a group photo. In such a case, the person’s high quality photo can be easily collected from the same album and then used to perform the reference-based face SR faithfully. Also, in old film restoration, to restore a facial region in a degraded film, a high-quality photo of the same actor or actress can be used as a reference. Note that the reference-based face SR is beneficial in not only restoring details but also in preserving personal identity. This reference-based face SR was also considered in [15], [16], in which a reference image is aligned to an input image based on warping, assuming that the input and reference images are similar everywhere, including background. However, these methods may yield unreliable results when the assumption is invalid (e.g., different backgrounds). Furthermore, they use only local information. In contrast, we propose CollageNet to use a reference image at the patch level. For each patch in an input image, we search the $K$ most similar patches non-locally in a reference image. We glue those similar patches together to generate a high-quality SR result.

To achieve the reference-based face SR effectively, we design the guided feature extractor (GFE) based on an autoencoder structure. GFE extracts feature pyramids, consisting of hierarchical features at multiple spatial scales, from input and reference images. Also, we develop the patch grouping layer (PGL), in which we calculate the patch-wise similarities and
select the $K$ most similar reference patches to each input patch. Then, we compose a collaged feature pyramid by gluing those selected patches together. Lastly, using the proposed feature pyramid blender (FPB), we obtain an SR facial image by blending the collaged feature pyramid and the input feature. The proposed CollageNet outperforms conventional SR algorithms by exploiting detailed information in reference images.

This work has the following contributions:

- We propose a novel face SR algorithm that uses a reference image at the patch level.
- We develop the three composing blocks of GFE, PGL, and FPB, which are tailored for feature collaging and image restoration tasks.
- The proposed CollageNet provides remarkable face SR performances and outperforms existing algorithms.

II. RELATED WORK

Face SR aims at improving the resolution of a facial image. To super-resolve a facial image, some techniques use external data, such as reference images or facial attributes (e.g., gender, age, and beard style), as well as internal data (i.e., the input image itself).

A. TRADITIONAL FACE SR

Baker and Kanade [17] built Gaussian pyramids of LR input images and HR training images and learned the resolution enhancement function using them. To exploit not only internal but also external data, Wheeler et al. [18] used multi-frame facial images. They assumed that the multiple frames had small variations and could be blended well, but sufficient detail information could not be extracted because each of those frames had an LR. Kolouri and Rohde [19] proposed a transport-based face SR algorithm to learn a nonlinear Lagrangian model for HR facial images.

B. CNN-BASED FACE SR

Recently, various CNN-based face SR algorithms have been developed. Zhu et al. [20] proposed the cascade bi-network containing a common branch and a high-frequency branch. Yu and Porikli [21] adopted the generative adversarial network (GAN) to reconstruct realistic facial images even at a large scale factor ($\times 8$). Chen et al. [22] proposed an end-to-end face SR algorithm using landmark heatmaps and parsing maps. Shi et al. [23] proposed a local enhancement network for patches and developed a recurrent policy network to select an optimal patch sequence based on reinforcement learning. Xin et al. [24] extended the capsule network [25] to encode facial information and yield reliable SR results even from noisy images. Yin et al. [26] considered joint SR and landmark localization of tiny faces. Ma et al. [27] performed face SR through iterative collaboration of image recovery and landmark estimation. Also, Hu et al. [28] extracted 3D priors from an image and used them to yield a faithful hallucination result through spatial attention. Yu et al. [29] developed the attribute-embedded upsampling network using facial attributes.

C. REFERENCE-BASED SR

Some algorithms use an HR reference image as external information for SR of general images. Zhang et al. [30] and Yang et al. [31] matched the most similar pairs of input and reference feature patches. They used those reference patches to restore detailed information. However, content dissimilarity and different camera views hinder their methods from selecting reliable reference images. Unlike general images, for a person’s facial image, a reliable reference image can be selected from the same person’s album. Liu et al. [32] adopted a conditional variational autoencoder to restore details by employing the same person’s HR image as external data. They embedded input and reference images into a joint latent space using an encoder and generated an SR image using a decoder. Li et al. [15] adopted two stages of warping and reconstruction. They warped a reference image to an input image using optical flow, but they used the reference image at the image level without dividing it into patches. Thus, misalignment and background differences may hinder reliable reconstruction. To alleviate the misalignment problem, Li et al. [16] selected the best-aligned reference image...
among multiple candidates. However, if there are not enough candidates, their algorithm may be still affected adversely by the misalignment.

### III. PROPOSED ALGORITHM

Figure 1 is an overview of CollageNet containing the three components of GFE, PGL, and FPB. Let $I_{LR}$, $I_{GT}$, and $I_{Ref}$ denote an input LR image, its ground-truth HR image, and a reference HR image, respectively. CollageNet aims to generate an SR image $I_{SR}$, which is as similar to the ground-truth $I_{GT}$, using the information in both $I_{LR}$ and $I_{Ref}$.

#### A. GUIDED FEATURE EXTRACTOR (GFE)

To extract informative features, reference-based SR methods [30], [31] use a feature pyramid [33], composed of features at multiple spatial scales. To extract the pyramid, they use a pre-trained network, such as VGG19 [34], for classification directly without changing any parameters. However, features from the pre-trained network may lose essential information for restoring an image. To extract a more optimized feature pyramid for SR, we first design GFE based on symmetric autoencoder architecture in Figure 2(a) and then train it using the proposed guidance loss.

In Figure 2(a), GFE has convolution layers with ReLU activations, pooling layers, and upsampling layers. We adopt the VGG19 structure as the encoder and design the decoder by replacing the pooling layers with the upsampling layers. Let $\kappa = 2^M$ denote the scale factor. In other words, given an input $I_{LR}$ of resolution $H \times W$, we attempt to generate a super-resolved $I_{SR}$ of resolution $\kappa H \times \kappa W$. GFE generates a feature pyramid with $M + 1$ levels. Specifically, given an image $I$ of resolution $\kappa H \times \kappa W$, GFE yields the feature pyramid

$$F = \{F_m \mid \text{the spatial resolution } 2^m H \times 2^m W \}_{m=0}^{M}$$

where $m = 0$ and $M$ correspond to the coarsest and finest pyramidal levels, respectively. There are $M$ pooling layers in GFE. Table 1 represents the detailed network structure of GFE at scale factor $\kappa = 8$.

---

**TABLE 1.** Detailed network structure of GFE. Conv($N_m$, $N_m$) denotes a $3 \times 3$ convolution layer with $N_m$ input channels and $N_m$ output channels. Also, ReLU, Pool, and Up denote the ReLU activation, a $2 \times 2$ max pooling layer with stride 2, and an upsampling layer using the bicubic interpolation, respectively.

| Level | Input | Layers | Output |
|-------|-------|--------|--------|
| 3     | $I$   | Conv(3, 64), ReLU | $F_3$  |
| 2     | $F_3$ | Conv(64, 64), ReLU, Pool Conv(64, 128), ReLU | $F_2$  |
| 1     | $F_2$ | Conv(128, 128), ReLU, Pool Conv(128, 256), ReLU | $F_1$  |
| 0     | $F_1$ | Conv(256, 256), ReLU Conv(256, 256), ReLU Conv(256, 256), ReLU Pool Conv(256, 512), ReLU | $F_0$  |
| 0     | $F_0$ | Conv(512, 256), ReLU, Up Conv(256, 256), ReLU Conv(256, 256), ReLU Conv(256, 256), ReLU | $\hat{F}_0$ |
| 1     | $\hat{F}_0$ | Conv(256, 128), ReLU, Up Conv(128, 128), ReLU | $\hat{F}_1$ |
| 2     | $\hat{F}_1$ | Conv(128, 64), ReLU, Up Conv(64, 64), ReLU | $\hat{F}_2$ |
| 3     | $\hat{F}_2$ | Conv(64, 3), ReLU | $I$    |
Algorithm 1 Patch Grouping Layer (PGL)

Input: \{D^{LR}_{m}\}_{m=0}^M, \{D^{Ref\uparrow}_{m}\}_{m=0}^M, and \{P^{Ref}_{m}\}_{m=0}^M
1: for all \(m = 0, 1, \ldots, M\) do
2: for all patch \(P^{LR\uparrow}_{m,i}\) do
3: for all patch \(P^{Ref\uparrow}_{m,j}\) do
4: Compute \(\rho_{i,j}\) via Eq. (4)
5: end for
6: Find the \(K\) most similar patches in \(P^{Ref}_{m}\)
7: Obtain \(P^{G}_{m,i}\) via Eq. (5)
8: Convolve \(P^{G}_{m,i}\) to yield \(P^{Col}_{m,i}\)
9: end for
10: Generate \(F^{Col}_{m}\) from \(\{P^{Col}_{m,i}\}_{i=1}^L\)
11: end for

Output: \(F^{Col}_{m}\) \(\_{m=0}^M\)

The guidance loss is defined as
\[
L_{G} = \left\| I - \hat{I} \right\|_1
\] (2)

which is the \(l_1\)-norm between the input image \(I\) and its reconstruction \(\hat{I}\) obtained by the decoder. This guidance loss is added to the total network loss to train GFE jointly with the other parts of CollageNet in an end-to-end manner.

B. PATCH GROUPING LAYER (PGL)

In the existing reference-based face SR methods \cite{15, 32}, the information in a reference image is exploited at the image level. In contrast, we use reference information at the patch level, as in Zhang et al. \cite{30} and Yang et al. \cite{31}, which transfer textural information based on the similarity between input and reference patches. Whereas the most similar reference patch is used in \cite{30, 31}, the \(K\) most similar patches are employed in the proposed PGL.

The reference HR image \(I^{Ref}\) is assumed to have more detail information than the input image \(I^{LR}\). For each local patch in \(I^{LR}\), we attempt to find the \(K\) most similar patches in \(I^{Ref}\) and exploit their detail information to super-resolve it. However, because \(I^{LR}\) and \(I^{Ref}\) have different resolutions, they are not directly comparable \cite{30, 31}. Hence, we modify \(I^{LR}\) and \(I^{Ref}\) accordingly, specifically, we upsample \(I^{LR}\) to get \(I^{LR\uparrow}\) and also apply downsampling and upsampling sequentially to \(I^{Ref}\) to get \(I^{Ref\uparrow}\). For the upsampling and downsampling, we use the bicubic interpolation, which is used to generate input images from the ground-truth HR images during the training of CollageNet. Note that both \(I^{LR\uparrow}\) and \(I^{Ref\uparrow}\) have the same resolution \(\kappa H \times \kappa W\) as the ground-truth \(I^{GT}\).

As in Eq. (1), we extract the feature pyramid \(P^{LR\uparrow} = \{P^{LR\uparrow}_{m}\}_{m=0}^M\) from the modified input image \(I^{LR\uparrow}\) through GFE. For example, for scale factor \(\kappa = 8\), there are four feature maps \(F^{LR\uparrow}_0\), \(F^{LR\uparrow}_1\), \(F^{LR\uparrow}_2\), and \(F^{LR\uparrow}_3\) from the coarsest to the finest levels. We reorganize these feature maps into overlapping patches of size \(h \times w\). The patch size is \(h = w = 3 \cdot 2^m\) at level \(m\). For example, at the coarsest level \(m = 0\), overlapping patches of size \(3 \times 3\) are employed. Let \(P^{LR\uparrow}_{m}\) be the set of patches from \(F^{LR\uparrow}_{m}\) given by
\[
P^{LR\uparrow}_{m} = \{P^{LR\uparrow}_{m,l}\}_{l=1}^L
\]

where \(C\) is the number of channels, and \(L\) is the number of patches in \(P^{LR\uparrow}_{m}\). Similarly, \(P^{Ref\uparrow}_{m}\) is obtained from \(F^{Ref\uparrow}_{m}\), and \(P^{Ref}_{m}\) is from \(F^{Ref}_{m}\).

These three sets \(P^{LR\uparrow}_{m}\), \(P^{Ref\uparrow}_{m}\), and \(P^{Ref}_{m}\) are fed into PGL in Figure 2(b). For each \(i\)th patch \(P^{LR\uparrow}_{m,i}\) in \(P^{LR\uparrow}_{m}\), we compute the similarity score \(\rho_{i,j}\) with every \(j\)th patch \(P^{Ref\uparrow}_{m,j}\) in \(P^{Ref}_{m}\),
\[
\rho_{i,j} = \left\langle \frac{P^{LR\uparrow}_{m,i}}{\|P^{LR\uparrow}_{m,i}\|_2}, \frac{P^{Ref\uparrow}_{m,j}}{\|P^{Ref\uparrow}_{m,j}\|_2} \right\rangle
\] (4)

where each patch is assumed to be reshaped into a vector. Next, we find the \(K\) most similar patches in \(P^{Ref}_{m}\) using the similarity scores, and \(K = 5\) in this work. Let \(j_k\) be the index of the \(k\)th most similar patch to \(P^{LR\uparrow}_{m,i}\). Then, the weighted patch group \(P^{G}_{m,i}\) is obtained by
\[
P^{G}_{m,i} = \rho_{i,j_k}P^{Ref}_{m,j_k} / \rho_{i,j_1}P^{Ref}_{m,j_1} \cdots / \rho_{i,j_K}P^{Ref}_{m,j_K}
\]

where \(\rho\) denotes the concatenation along the channel dimension. Every patch group \(P^{G}_{m,i}\) in \(\mathbb{R}^{h \times w \times C}\) is processed by a convolution layer with \(C\) filters of size \(1 \times 1 \times C\), resulting in the collaged patch \(P^{Col}_{m,i}\) in \(\mathbb{R}^{h \times w \times C}\). Finally, we generate the collaged feature map \(F^{Col}_{m}\) by overlapping such collaged patches. Note that, for overlapped areas between patches, we perform averaging. This is repeated for all \(M + 1\) levels, resulting in the collaged feature pyramid \(F^{Col}_{m}\) which is used to restore the details of the input image \(I^{LR}\). The overall procedure of PGL is summarized in Algorithm 1.

C. FEATURE PYRAMID BLENDER (FPB)

To achieve faithful face SR, it is important to blend the base information in \(I^{LR}\) and the detail information in \(I^{Ref}\) at each level. The output of the PGL \(F^{Col}_{m}\) is further processed by a feature pyramid blender (FPB) with 

\[\cdot\] : upsampling
\[\cdot\] : downsampling
\[3 \times 3\] Conv
\[\cdot\] : residual block
\[\cdot\] : branch block
\[\cdot\] : channel attention
\[\cdot\] : blending block
\[\cdot\] : FC layer
\[\cdot\] : sigmoid

FIGURE 3. Flow diagram of FPB.
effectively. Hence, we design FPB to mix the LR feature $F_{\text{LR}}^M$ at the finest level $M$ with the collaged feature pyramid $F_{\text{Col}}^M$ across all levels $0 \leq m \leq M$, as shown in Figure 3(a). FPB includes convolution layers, residual blocks [8], up/down-sampling layers, and blending blocks. Each blending block accepts features at multiple levels via up/down-sampling, processes them using convolution layers, concatenates the results, and performs the channel attention [9] to yield a more powerful feature representation and improve the SR performance. Note that there are $M - m + 1$ blending blocks at level $m$. The output of each blending block is filtered subsequently by multiple residual blocks and a single convolution layer. Finally, we match the resolutions of the features from all levels, concatenate them, and use a convolution layer to restore the SR image $\hat{I}^{\text{SR}}$.

Figure 3(c) and (d) show the residual block [8] and the channel attention block [9] in FPB. These components help FPB to blend a collaged feature pyramid and an input feature, yielding a more powerful feature representation. As shown in Figure 3(d), an average-pooled feature $F_{\text{avg}}^m$ and a max-pooled feature $F_{\text{max}}^m$, extracted from the module input $F$, are fed into a shared network to produce $F_{\text{avg}}$ and $F_{\text{max}}$, respectively. The shared network is composed of two fully-connected (FC) layers: the former FC layer reduces the channel dimension by a factor of $\frac{1}{64}$, while the latter one increases it by a factor of 16. Then, $F_{\text{avg}}$ and $F_{\text{max}}$ are merged by the element-wise summation, followed by the sigmoid function, to produce a channel attention map $M_{\text{ch}}$.

### D. LOSS FUNCTIONS

In addition to the reconstruction loss $L_R = \| \hat{I}^{\text{SR}} - I^{\text{GT}} \|_1$ between a super-resolved image $\hat{I}^{\text{SR}}$ and the ground-truth $I^{\text{GT}}$, we use the perceptual loss $L_P$ and the adversarial loss $L_A$ to enhance the visual quality of $\hat{I}^{\text{SR}}$. The perceptual loss is defined as

$$L_P = \| F_{\text{VGG19}}^{\text{SR}} - F_{\text{VGG19}}^{\text{GT}} \|_1 + \| F_{M}^{\text{SR}} - F_{M}^{\text{Col}} \|_1 \tag{6}$$

where $F_{\text{VGG19}}^{\text{SR}}$ and $F_{\text{VGG19}}^{\text{GT}}$ are the output of the VGG19 relu5-1 layer [34] using $I^{\text{SR}}$ and $I^{\text{GT}}$ as input, respectively. In Eq. (6), the first term $\| F_{\text{VGG19}}^{\text{SR}} - F_{\text{VGG19}}^{\text{GT}} \|_1$ helps to yield better perceptual qualities [7], [30]. The second term $\| F_{M}^{\text{SR}} - F_{M}^{\text{Col}} \|_1$ encourages the SR image $\hat{I}^{\text{SR}}$ to have similar texture features to the collaged feature map $F_{M}^{\text{Col}}$ constructed from the reference image $I^{\text{Ref}}$ [31]. It facilitates the transfer of detail information from $I^{\text{Ref}}$ to $\hat{I}^{\text{SR}}$. The adversarial loss $L_A$ is used to enhance subjective image qualities by emphasizing high-frequency information. We adopt the Wasserstein GAN loss for stable training [30]. Also, we employ the guidance loss $L_G$ in Eq. (2) to train GFE effectively.

Finally, we use the overall loss $L$, given by

$$L = w_R L_R + w_P L_P + w_A L_A + w_G L_G \tag{7}$$

where $w_R$, $w_P$, $w_A$, and $w_G$ are the weights for the reconstruction, perceptual, adversarial, and guidance losses, respectively.

### E. IMPLEMENTATION DETAILS

The overall loss in Eq. (7) consists of the four loss functions. However, we found that, at an initial stage of training, it is better to use only the guidance and reconstruction losses. By doing so, GFE is trained more reliably. We train CollageNet for 100 epochs. For the first two epochs, we set $w_R = 0$, $w_G = 0$, and $w_P = w_A = 0$. For the remaining 98 epochs, $w_R = 1$, $w_G = 0.1$, and $w_P = w_A = 0$. For a fair comparison, we slightly modify the training set in [27] so that each pair satisfies the same identity constraint as in Figure 4. The construction process is detailed in the Appendix A. We train CollageNet using CelebA, and test it on CelebA, LFW, and RefSR-Face.

- **CelebA** [2] contains 202,599 facial images of 10,177 people, among which Ma et al. [27] used 168,854 images for training and 1,000 ones for testing. For a fair comparison, we slightly modify the training set in [27] for the reference-based face SR without reducing the number of the testing images. Specifically, in the case

---

*This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3138442, IEEE Access*
of a person with only one image, the person is excluded from the training of CollageNet, resulting in 168,704 pairs for training and 1,000 pairs for testing.

- **LFW** [38] contains 13,233 facial images of 5,749 people, among which we use 3,707 image pairs for testing as done in [39]. In the case of a person with only one image, a modified input image $I_{LR}^w$ is used as the reference image to maintain the number of the testing images in [39].

- **RefSR-Face** [32] contains 560 facial image pairs of 428 people, which are used for testing as done in [32].

For quantitative comparison of face SR results, we measure the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) in the Y channel of the YCbCr space.

### A. COMPARATIVE ASSESSMENT

We assess the proposed algorithm in two modes: 1) the default CollageNet for high perceptual quality and 2) CollageNet-Rec for high PSNR performance. Specifically, CollageNet-Rec is trained by setting $w_p = w_A = 0$ in Eq. (7) to use $L_R$ and $L_G$ only.

#### 1) Comparison with conventional algorithms

Table 2 compares the proposed algorithm with single-image SR methods (SRResNet [13], RDN [35]), reference-based SR methods (SRNTT [30], TTSR [31]), single-image face SR methods (URDGN [21], PFSR [36], FSRNet [22], Attention-FH [23], DIC [27]), and reference-based face SR method (RefSR-VAE [32]).

It is worth pointing out that, for a fair comparison, the proposed algorithm should be compared with only reference-based methods. However, the source codes of the reference-based face SR methods in [15], [16] are not available. Thus, we can compare the proposed algorithm with only one reference-based face SR method [32] and two reference-based SR methods [30], [31]. To compensate for this limited set of comparison methods, we compare the proposed algorithm additionally with URDGN [21], PFSR [36], FSRNet [22], Attention-FH [23], and DIC [27], as done in [30]–[32].

In Table 2, the scores of [13], [21], [22], [27], [35], [36] on CelebA are excerpted from [27], and those of Attention-FH [23] on LFW are from [23]. For the other scores, we either train [13], [30], [31], [35] on CelebA ourselves or use the available models [22], [27], [32], [36] trained on facial images.

CollageNet-Rec outperforms the state-of-the-art TTSR, DIC, and RefSR-VAE on the three datasets and yields a comparable SSIM to Attention-FH on LFW. Moreover, whereas some methods [22], [27], [31], [32], [35], [36] seem to be overfit to specific datasets, CollageNet yields robust performances on all datasets. Figure 5 compares qualitative results on CelebA and LFW at the scale factor $\kappa = 8$. We see that CollageNet provides more faithful results with fewer artifacts especially on detailed patterns within the red rectangles.

#### 2) Various scale factors

We perform comparisons at the scale factors $\kappa = 2, 4, 8$, and 16. To this end, we train RDN, TTSR, and DIC using the available training codes until the performances are saturated. Table 3 compares the average PSNRs on CelebA and LFW. CollageNet-Rec outperforms the conventional algorithms at every scale factor. Figure 6 visualizes qualitative results at $\kappa = 16$. Although the input images provide little information because of their extremely low resolution, both CollageNet-Rec and CollageNet reconstruct the details — e.g., eyes, noses, and teeth — more faithfully than the conventional algorithms. More qualitative comparison results are in the Appendix B.

#### 3) User study

We conducted a user study to compare subjective qualities. Each of 15 participants was asked to assign a score from 1 (bad quality) to 5 (excellent quality) to SR results, as done in [36]. We selected 20 images randomly from CelebA and LFW testing sets and super-resolved them using the bicubic interpolation, TTSR, PFSR, FSRNet, DIC, and RefSR-VAE.
CollageNet. In Table 4, CollageNet yields an outstanding score, which means that it provides perceptually higher qualities than the conventional algorithms do.

**B. ABLATION AND ANALYSIS**

We do ablation studies and analysis on CelebA at $\kappa = 8$.

1) Robustness to reference images

We test the robustness of CollageNet to variations in reference images. To this end, we construct a *negative* pair by selecting a person’s input image and a different person’s reference image from CelebA test set. When such negative pairs are used, CollageNet and CollageNet-Rec yield PSNR scores of 26.15dB and 27.23dB, respectively. As compared with the results in Table 2, in which *positive* pairs are used, the PSNR scores are lowered by about 0.2dB by the negative pairs. It is, however, remarkable that CollageNet-Rec yields a comparable score to the state-of-the-art TTSR [31] and DIC [27] even with such negative pairs.

Figure 7 illustrates how CollageNet reconstructs an image using different references. The bottom row includes similarity maps obtained by overlapping similarity scores in Eq. (4).
In Figure 7(b), similarity scores are high in general, since the same person’s reference is used. In contrast, in Figure 7(c) or (d), a different person’s reference yields relatively low similarity scores, degrading the SR performance. Between Figure 7(c) and (d), the reference in (c) is more similar to the input than that in (d); more faithful reconstruction is achieved in (c). Even with the negative references, the proposed algorithm reconstructs details effectively. The person’s facial attributes, however, are transferred from the different people, which may be undesirable. Notice that the proposed algorithm is designed to use the same person’s reference.

2) Loss functions

Because of limited resources, we perform the training for seven epochs only in the following experiments. First, we analyze the efficacy of each loss term in Eq. (7). Figure 8 and Table 5 show the results of CollageNet trained with various combinations of the four loss terms. By comparing settings I with II, we see that the guidance loss $L_G$ helps to restore details and also improves the PSNR performance. Notice that setting II corresponds to CollageNet-Rec. In setting III, the adversarial loss $L_A$ enhances subjective qualities by generating plausible images. However, it degrades the overall PSNR performance in Table 5 and fails to restore the person’s unique facial attributes, determining his identity, in Figure 8(c). Finally, in setting IV (i.e CollageNet), this problem is alleviated by the perceptual loss $L_P$, which helps to transfer

![Image of the text content](image-url)
FIGURE 7. Comparison of face SR results (top) according to reference images and similarity maps (bottom). The proposed algorithm uses a positive reference in (b) and two negative references in (c) and (d). The results of TTSR and DIC are shown in (e) and (f).

FIGURE 8. Examples of SR results of CollageNet according to the combinations of loss terms, summarized in Table 5.

TABLE 6. Ablation studies of the proposed CollageNet.

| Components | PGL | GFE | FPB | PSNR |
|------------|-----|-----|-----|------|
|            | ✓   | ✓   | ✓   | 26.93|
|            | ✓   | ✓   | ✓   | 26.97|
|            | ✓   | ✓   | ✓   | 26.98|
|            | ✓   | ✓   | ✓   | 27.05|

3) Ablation studies

In Table 6, we provide the average PSNRs of several ablation studies to validate the efficacy of PGL, GFE, and FPB. First, without PGL, we use the single most similar reference patch only, which degrades the performance by about 0.12 dB. This confirms that PGL exploits the information in reference images effectively. Second, the inclusion of GFE, instead of the pre-trained VGG19, improves PSNR by about 0.08 dB. This is because GFE extracts more informative feature pyramids from both input and reference images. Finally, using FPB helps to transfer detail information from reference to input images. It improves PSNR 0.07 dB in comparison with the model, which uses the combination of a concatenation operator and a convolution layer instead of FPB.

4) Number \( K \) of most similar patches

Finally, we discuss the impacts of the number \( K \) of most similar patches in PGL. Table 7 shows that the performance increases up to \( K = 5 \) and decreases thereafter, indicating that the proposed patch grouping is more effective than using the single most similar reference patch as done in [30], [31].

C. REAL-WORLD SCENARIO

To assess the proposed CollageNet in a real-world scenario, we super-resolve LR facial images of Marie Skłodowska Curie and Albert Einstein in the group photo in Figure 9, which was taken in the Solvay Conference, 1927. Note that the photo is intentionally downsampled to compare SR results more easily. We crop their tiny facial regions and super-resolve them by employing TTSR [31], PFSR [36], FSRNet [22], DIC [27], and CollageNet. We collect their HR reference images easily from the Internet. In Figure 9,
CollageNet super-resolves the facial regions more effectively and preserves the personal identity more faithfully than the conventional algorithms do, which indicates that CollageNet is also effective in real applications.

V. CONCLUSIONS
We proposed CollageNet for novel reference-based face SR at the patch level. It consists of GFE, PGL, and FPB. First, GFE obtains feature pyramids from input and reference images. Then, PGL generates a collaged feature pyramid to exploit the multi-scale information of the reference image. Finally, FPB yields an SR image by blending the collaged feature pyramid and the input feature. Experimental results demonstrated that CollageNet significantly outperforms conventional algorithms.

Although CollageNet restores details effectively using reference images, the patch grouping demands high memory and computational complexities. It is a future research issue to design a more efficient grouping scheme, which would facilitate super-resolving facial images of large sizes. Also, we will generalize the proposed algorithm to use multiple reference images to super-resolve an input image.

APPENDIX A CONSTRUCTING IMAGE PAIRS
The proposed CollageNet super-resolves a person’s low-resolution (LR) facial image by exploiting the same person’s high-resolution (HR) reference image. Therefore, we...
construct pairs of input and reference images using existing datasets [2], [38], in which each image has identity information. For an input image, we compute the similarity score, based on SIFT [41], with every image of the same identity information [30]. We then select the image with the highest similarity as the reference image, as illustrated in Figure 10.

APPENDIX B MORE QUALITATIVE RESULTS

Figure 11 and Figure 12 visualize more qualitative comparison results at the scale factors $\kappa = 8$ and $16$ on CelebA.

REFERENCES

[1] E. Zhou, Z. Cao, and Q. Yin, “Naïve-deep face recognition: Touching the limit of LFW benchmark or not?” arXiv:1501.04690, 2015.
[2] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in ICCV, 2015.
[3] 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, 2015.
[4] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, and F. Wen, “Old photo restoration via deep latent space translation,” arXiv:2009.07047, 2020.
[5] A. Dantcheva and J.-L. Dugelay, “Frontal-to-side face re-identification based on hair, skin and clothes patches,” in AVSS, 2011.
[6] M. Grigic, K. Delac, and S. Grigic, “SCface-surveillance cameras face database,” Multimedia Tools and Applications, vol. 51, no. 3, pp. 863–879, 2011.
[7] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in ECCV, 2016.
[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016.
[9] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in CVPR, 2018.
[10] J.-S. Kim, K. Ko, and C.-S. Kim, “Adaptive lattice-aware image de-noising using global and local information,” in ICIP, 2020.
[11] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, 2015.
[12] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in CVPR, 2016.
[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in CVPR, 2017.
[14] K. Ko, Y. J. Koh, S. Chang, and C.-S. Kim, “Light field super-resolution via adaptive feature remixing,” IEEE Trans. Image Process., vol. 30, pp. 4114–4128, 2021.
[15] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, “Learning warped guidance for blind face restoration,” in ECCV, 2018.
[16] 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 CVPR, 2020.

VOLUME 1, 2021

FIGURE 10. Candidate pairs of input (left) and reference (right) images. We select the second pair within the red rectangle, which exhibits the highest score.
FIGURE 11. Comparison of face SR results at $\kappa = 8$: Input images in (a) are super-resolved by TTSR [31] in (b), PFSR [36] in (c), FSRNet [22] in (d), DIC [27] in (e), CollageNet-Rec in (f), and CollageNet in (g). Ground-truth and reference images are in (h) and (i), respectively.
FIGURE 12. Comparison of face SR results at $\kappa = 16$: Input images in (a) are super-resolved by RDN [35] in (b), TTSR [31] in (c), DIC [27] in (d), CollageNet-Rec in (e), and CollageNet in (f). Ground-truth and reference images are in (g) and (h), respectively.