Learning Spatial Attention for Face Super-Resolution

Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, *Senior, IEEE*, and Kwan-Yee K. Wong, *Senior, IEEE*

***Abstract*—General image super-resolution techniques have difficulties in recovering detailed face structures when applying to low resolution face images. Recent deep learning based methods tailored for face images have achieved improved performance by jointly trained with additional task such as face parsing and landmark prediction. However, multi-task learning requires extra manually labeled data. Besides, most of the existing works can only generate relatively low resolution face images (*e.g.*, 128×128), and their applications are therefore limited. In this paper, we introduce a novel SPatial Attention Residual Network (SPARNet) built on our newly proposed Face Attention Units (FAUs) for face super-resolution. Specifically, we introduce a spatial attention mechanism to the vanilla residual blocks. This enables the convolutional layers to adaptively bootstrap features related to the key face structures and pay less attention to those less feature-rich regions. This makes the training more effective and efficient as the key face structures only account for a very small portion of the face image. Visualization of the attention maps shows that our spatial attention network can capture the key face structures well even for very low resolution faces (*e.g.*, 16×16). Quantitative comparisons on various kinds of metrics (including PSNR, SSIM, identity similarity, and landmark detection) demonstrate the superiority of our method over current state-of-the-arts. We further extend SPARNet with multi-scale discriminators, named as SPARNetHD, to produce high resolution results (*i.e.*, 512×512). We show that SPARNetHD trained with synthetic data cannot only produce high quality and high resolution outputs for synthetically degraded face images, but also show good generalization ability to real world low quality face images. Codes are available at https://github.com/chaofengc/Face-SPARNet.**

***Index Terms*—Face Super-Resolution, Spatial Attention, Generative Adversarial Networks**

I. Introduction

Face super-resolution (SR), also known as face hallucination, refers to generating high resolution (HR) face images from the corresponding low resolution (LR) inputs. Since there exist many low resolution face images (*e.g.*, faces in surveillance videos) and face analysis algorithms (*e.g.*, face recognition) often perform poorly on such images, there is a growing interest in face SR.

Different from general image SR, face SR places its emphasis on the recovery of the key face structures (*i.e.*, shapes of face components and face outline). These structures only account for a very small portion of the image, but are often more difficult to recover as they exhibit larger pixel variations.

Chaofeng Chen and Kwan-Yee K. Wong are with the Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong. (E-mail: [cfchen, kykwong]@cs.hku.hk)

Dihong Gong, Hao Wang and Zhifeng Li are with Tencent AI Lab, Shenzhen 518000, China. (E-mail: gongdihong@gmail.com, hawelwang@tencent.com, michaelzfli@tencent.com)

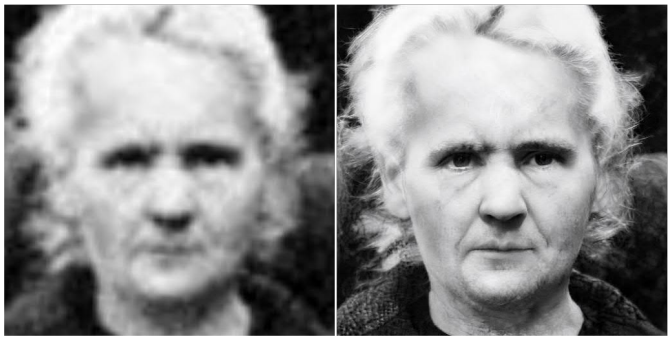


Fig. 1: Super-resolution result produced by SPARNetHD for an old photo of Marie Curie from Solvay conference 1927. Please zoom in to see details.

Training a deep neural network with the commonly used mean square error (MSE) loss, which weights pixels equally, is not very effective in recovering these "sparse" structures. Previous works [1], [2], [3] proposed incorporating additional task, such as face parsing and landmark detection, to assist the training of face SR networks. [1], [3] also used predicted face priors to help face SR. Although joint training with these additional tasks helps enhance the importance of the key face structures, there are two major drawbacks, namely (1) extra effort is needed to label the data for the additional task, and (2) predicting face prior from LR inputs is itself also a difficult problem.

On the other hand, if we subdivide a face image into many small regions and consider each region as an individual sample, the unbalanced distribution between regions containing key face structures (referred to as hard regions) and those not containing key face structures (referred to as easy regions) will resemble the imbalance between foreground and background samples in object detection. This suggests that we may adopt some techniques similar to *bootstrapping* or *online hard example mining* (OHEM) [4] in object detection to solve our face SR problem.

In this paper, we introduce a carefully designed Face Attention Unit (FAU) to construct a SPatial Attention Residual Network (SPARNet) for face SR. The key idea is to bootstrap features related to the key face structures using a 2D spatial attention map. Instead of hard selection, the spatial attention map assigns a score between 0 and 1 to each spatial location of the feature map. This allows learning the prediction of the spatial attention map through gradient descent. Spatial attention maps in different FAUs of the network can learn to focus on different face structures. For example, attention

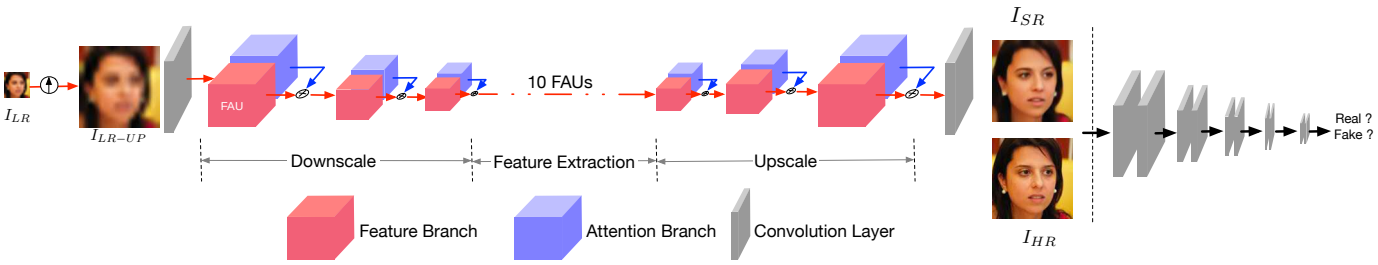


Fig. 2: Architecture of the proposed SPatial Attention Residual Network (SPARNet).

maps in deeper layers focus more on coarse structures such as eyes and mouth, while those in shallower layers focus more on detailed textures such as hairs. Considering that most of the existing face SR methods can only produce 128×128 outputs, we further extend SPARNet, referred to as SPARNetHD, to generate high resolution outputs (*i.e.*, 512×512). Specifically, we enlarge the output resolution of SPARNet from 128×128 to 512×512 and adopt a multi-scale discriminator loss similar to Pix2PixHD [5] to generate more realistic textures. Experiments show that SPARNetHD trained with synthetic LR data is pretty robust with natural LR inputs, while models without the proposed spatial attention mechanism produce undesirable artifacts. We show an example result of SPARNetHD on an old photo in Fig. 1. We can see that SPARNetHD can restore key face components very well and also generate high resolution and realistic textures.

The key contributions of this paper can be summarized as follows:

1) We propose an efficient framework named SPARNet for face super-resolution. Without relying on any extra supervisions (*e.g.*, face parsing maps and landmarks), it achieves state-of-the-art performance on various kinds of metrics, including PSNR, SSIM, identity similarity, and landmark detection.
2) We show that the proposed FAU, the basic building block of SPARNet, can bootstrap the key face structures (*i.e.*, face components and face outline) and significantly improve the performance of face super-resolution.
3) By repeating FAUs in SPARNet, the spatial attention maps in different FAUs can learn to focus on different face structures and further improve the performance of SPARNet.
4) We introduce SPARNetHD to generate high resolution face images (*i.e.*, 512×512), and the model trained with synthetic data works equally well on natural LR images.

II. Related Works

In this section, we briefly review the literature on face super-resolution and recent attention neural networks.

A. Face Super-Resolution

Face hallucination was pioneered by Baker and Kanade [6], who showed that it is possible to perform high magnification SR for images of a specific category (*e.g.*, face and text). Since then many methods were proposed to improve the performance of face hallucination. They can be roughly classified into sub-space based methods [7], [8], [9], [10], [11] and component based methods [12], [13], [14]. Sub-space based methods usually rely on Principle Component Analysis (PCA), which requires precisely aligned faces. Component based methods require detecting facial landmarks, which is difficult for LR face images. Both kinds of methods fail to produce satisfactory results for face SR with a high upscale factor.

Recently, deep convolutional neural networks (CNNs) have brought remarkable progress to face SR. Zhu *et al.* [15] proposed a cascaded two-branch network to optimize face hallucination and dense correspondence field estimation in a unified framework. Yu *et al.* exploited generative adversarial networks (GAN) [16] to directly super-resolve the LR inputs. They further improved their model to handle unaligned faces [17], noisy faces [18], and faces with different attributes [19]. Instead of directly inferring HR face images, Huang *et al.* proposed to predict wavelet coefficients from LR images to reconstruct HR images. Latest works employed extra face prior supervisions, such as face parsing maps [1], landmark heatmaps [2], [3], and identity information [20], to train their networks. Kim *et al.* [21] proposed a facial attention loss which focuses the network on the landmark region. Ma *et al.* [22] introduced an iterative method which predicts SR results and landmarks iteratively. Although extra supervisions help to improve the performance, there are two shortcomings, namely (1) extra effort is needed to label the data, and (2) it is an indirect way to direct the SR network towards the key face structures. Different from these methods, our SPARNet learns to predict spatial attention maps to bootstrap key face structures and other feature-rich regions.

B. Attention Networks

Attention mechanism has been widely used in high level vision tasks, such as image classification [23], [24], [25], [26], image captioning [27], [28], and visual question answering [29], [30]. The key idea is to reweight features using a score map to emphasize important features and suppress less useful ones [24]. Wang *et al.* [25] introduced a trunk-and-mask attention mechanism to a residual network for image classification. He *et al.* [24] proposed the Squeeze-and-Excitation network which employs a channel-wise attention mechanism and demonstrates significant performance improvements. Woo *et al.* proposed a convolutional block attention module (CBAM) which sequentially infers attention maps along the channel and spatial dimensions separately. Attention mechanism has also

been employed in image generation tasks recently. Zhang *et al*. [31] combined channel attention with a very deep residual network for image SR. Cao *et al*. [32] proposed an attention aware face SR framework based on reinforcement learning, which sequentially attends to, crops out, and super-resolves a patch.

Although closely related, our work is different from [31] and [32] in several aspects. First, attention mechanisms in high level tasks often employ pooling for extracting semantic information. For face SR, however, this may result the loss of important low and middle level features such as edges and shapes. In contrast, our attention mechanism is designed to take advantages of multi-scale features. Second, unlike [31] which utilizes channel attention, our work considers spatial attention which facilitates region based attention of the key face components. Third, unlike [32] which takes a patch based approach, we generate spatial attention maps for the entire face. This allows our network to have a global view of the face structures and benefit from the contextual information.

C. High Resolution Image Generation with GAN.

Recently, GAN has demonstrated to be very effective in generating HR images. Wang *et al*. [5] proposed Pix2PixHD and used a multi-scale generator and discriminator architecture for HR image generation. Karras *et al*. [33], [34] proposed a new progressive training methodology for GAN to generate HR face images from random input vectors, and improved the quality of synthesized faces in [35]. Park *et al*. [36] proposed the spatially-adaptive normalization layer for generating photo-realistic images given an input semantic layout. We follow the idea of Pix2PixHD and use a multi-scale discriminator to generate high resolution results. Different from Pix2PixHD, our SPARNetHD does not require progressive training from LR to HR images, and our results are much better than Pix2PixHD with the help of the proposed spatial attention mechanism.

III. SPARNet for Face SR

A. Overview

As shown in Fig. 2, our SPARNet consists of three modules, namely the downscale module, the feature extraction module, and the upscale module. Each of these modules is composed of a stack of FAUs (see III-B for details). Let $I_{LR} \in \mathbb{R}^{3\times H'\times W'}$, $I_{SR} \in \mathbb{R}^{3\times H\times W}$, and $I_{HR} \in \mathbb{R}^{3\times H\times W}$ denote the LR face image (*i.e*., input), the super-resolved image (*i.e*., output), and the ground truth HR image respectively. I_{LR} is first upsampled to the same spatial dimension as I_{HR} through bicubic interpolation, denoted as $I_{LR-UP} \in \mathbb{R}^{3\times H\times W}$, which is then fed to SPARNet to produce I_{SR}. Given a training set with N pairs of LR-HR images, $\{I^i_{LR}, I^i_{HR}\}_{i=1}^N$, we optimize SPARNet by minimizing the pixel-level L_2 loss given by

$$\mathcal{L}_{pix}(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{F}_{SPAR}(I^i_{LR-UP}, \Theta) - I^i_{HR}\|_2^2, \quad (1)$$

where $\mathcal{F}_{SPAR}$ and Θ denote SPARNet and its network parameters respectively.

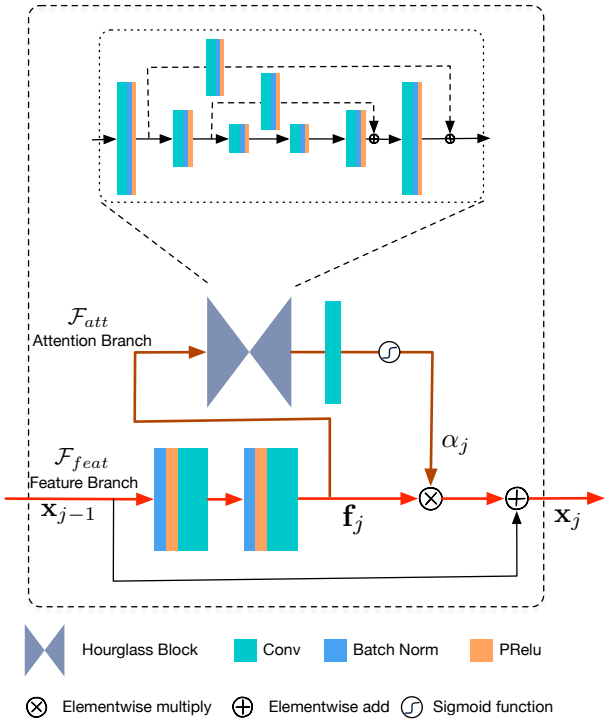


Fig. 3: Face Attention Unit

B. Face Attention Unit

Based on the observation that some face parts (*e.g*., key face components like eyes, eyebrows, nose, and mouth) are more important than the others (*e.g*., shading of the cheek) in face SR, we propose a spatial attention mechanism to make our network focus more on the important and informative features. The key question is how to produce the attention map and how to integrate it with the convolutional layers. First, we believe the spatial attention mechanism should not only have a high level view of the face but should also focus on low level structures. Note that a high level view helps the network to learn how faces look, while a low level view makes the network learn local details better. Hence, it would be desirable for the spatial attention mechanism to be able to learn from multi-scale features. Second, residual blocks have demonstrated great success in both general SR task [31], [37], [38] and face SR task [1], [2]. It should therefore be beneficial to integrate spatial attention mechanism with residual blocks.

Based on the above discussion, we propose a Face Attention Unit (FAU) which extends the original residual block by introducing a spatial attention branch (see Fig. 3). By stacking FAUs together, important features for face SR are continuously enhanced. Denote the feature input of the j-th indexed FAU as $\mathbf{x}_{j-1} \in \mathbb{R}^{C_{j-1}\times H_{j-1}\times W_{j-1}}$, the attention map $\boldsymbol{\alpha}_j \in \mathbb{R}^{1\times H_j\times W_j}$ is computed as

$$\mathbf{f}_j = \mathcal{F}_{feat}(\mathbf{x}_{j-1}), \quad (2)$$

$$\boldsymbol{\alpha}_j = \sigma(\mathcal{F}_{att}(\mathbf{f}_j)), \quad (3)$$

where $\mathbf{f}_j \in \mathbb{R}^{C_j\times H_j\times W_j}$ is the output of the feature branch $\mathcal{F}_{feat}$, $\mathcal{F}_{att}$ denotes the attention branch, and σ is the sigmoid

function. Finally, the output of the j-th indexed FAU is given by

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \boldsymbol{\alpha}_j \otimes \mathbf{f}_j, \tag{4}$$

where "$\otimes$" denotes element-wise multiplication. Details of $\mathcal{F}_{feat}$ and $\mathcal{F}_{att}$ are given in the next two paragraphs.

Attention Branch As discussed above, the attention branch should extract multi-scale features. We adopt the hourglass block followed by an extra Conv layer to generate the attention map. The hourglass block is known to be capable of capturing information at multiple scales [39]. It has also shown great performance in face analysis tasks, such as face alignment [40] and face parsing [1]. The kernel size and filter number for all Conv layers in the hourglass block are 3×3 and 64 respectively.

Feature Branch After experimenting with several variants of residual blocks [38], [41], we finally choose the pre-activation Residual Unit with PReLU [42] as our feature branch. Although previous general image SR work [38] argued that networks without batch normalization perform better, we find that pre-activation structure shows slightly better performance with batch normalization. We use PReLU as the activation function after batch normalization to avoid "dead features" caused by zero gradients in ReLU, and it shows more stable performance [43]. For the residual blocks in downscale and upscale module, we slightly modify the original residual branch using scale Conv (see Fig. 4), and (4) becomes

$$\mathbf{x}_j = \mathcal{F}_{scale}(\mathbf{x}_{j-1}) + \boldsymbol{\alpha}_j \otimes \mathbf{f}_j, \tag{5}$$

where $\mathcal{F}_{scale}$ denotes the scale Conv layer. Downscale Conv is a normal Conv layer with stride 2, and upscale Conv is a nearest-neighbor upsampling layer with a normal Conv layer which helps avoid checkerboard artifacts [44].

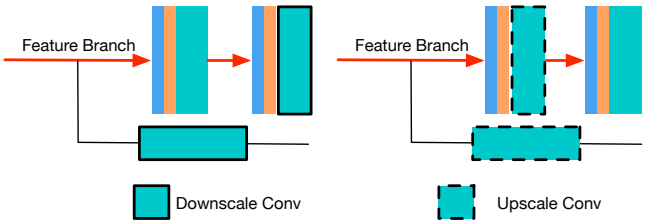


Fig. 4: Scale Residual Block

C. Multi-scale Discriminator Network

We extend our SPARNet to SPARNetHD to generate high resolution and more realistic SR images. SPARNetHD increases the channel number of SPARNet and adopt multi-scale discriminators similar to Pix2PixHD [5]. We refer to the discriminators as D_1, D_2, and D_3, which are used to discriminate SR images at three different scales, namely 512×512, 256×256, and 128×128, respectively, from the ground truth downsampled to the same resolutions. Using multiple discriminators at different scales can help to improve the quality of the SR images.

The loss functions used in training SPARNetHD are composed of the following four components:

(1) *Pixel loss*. We use L1-norm as the pixel level loss between I_{SR} and I_{HR}. It mainly helps to constrain the low level information in the outputs especially color, and is defined as

$$\mathcal{L}^h_{pix} = \frac{1}{N}\sum_{i=1}^{N} \|I^i_{SR} - I^i_{HR}\|_1, \tag{6}$$

$$I^i_{SR} = G(I^i_{LR-UP}), \tag{7}$$

where G denotes the SPARNetHD generator.

(2) *Adversarial loss*. It is the critical loss which helps to make the outputs sharper and generate more realistic textures such as hair. The loss functions of the generator and discriminator are formulated as

$$\mathcal{L}^h_{GAN\_G} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3} -D_k(I^i_{SR}), \tag{8}$$

$$\begin{aligned}\mathcal{L}^h_{GAN\_D} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3}[&\max(0, 1 - D_k(I^i_{HR})) \\ &+ \max(0, 1 + D_k(I^i_{SR}))],\end{aligned} \tag{9}$$

where the outputs of D_k are scalars which indicates whether the input images are real (≥ 1) or fake (≤ -1).

(3) *Feature matching loss*. This is the feature space loss of the discriminators [5]. It helps to stabilize the training of GAN. Let $\mathbf{f}^l_{D_k}$ be the feature map of the l-th layer in D_k, L_k be the total number of layers in D_k, and M^l_k be the number of elements in $\mathbf{f}^l_{D_k}$. The feature matching loss is then formulated as

$$\mathcal{L}^h_{fm} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3}\sum_{l=1}^{L_k}\frac{1}{M^l_k}\|\mathbf{f}^l_{D_k}(I^i_{SR}) - \mathbf{f}^l_{D_k}(I^i_{HR})\|_1, \tag{10}$$

(4) *Perceptual loss*. Different from feature matching loss, perceptual loss [37] is the feature space loss of the pretrained VGG19 network [45]. It helps to constrain the high level semantics in the outputs. We follow the notation of Eq. 10 and denote the perceptual loss as

$$\mathcal{L}^h_{pcp} = \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_{VGG}}\frac{1}{M^l_{VGG}}\|\mathbf{f}^l_{VGG}(I^i_{SR}) - \mathbf{f}^l_{VGG}(I^i_{HR})\|_1. \tag{11}$$

Finally, the loss functions are defined as

$$\mathcal{L}^h_G = \lambda_{pix}\mathcal{L}^h_{pix} + \lambda_{adv}\mathcal{L}^h_{GAN\_G} + \lambda_{fm}\mathcal{L}^h_{fm} + \lambda_{pcp}\mathcal{L}^h_{pcp}, \tag{12}$$

$$\mathcal{L}^h_D = \mathcal{L}^h_{GAN\_D}, \tag{13}$$

where $\mathcal{L}^h_G$ and $\mathcal{L}^h_D$ are minimized iteratively to train G and D, and $\lambda_{pix}, \lambda_{adv}, \lambda_{fm}$, and λ_{pcp} are the weights for each loss item respectively.

The reason why we use more complicated loss functions to train SPARNetHD is to generate results with high perceptual quality. As pointed out by [46], distortion and perceptual quality are at odds with each other. Methods trained with only $\mathcal{L}_{pix}$ in Eq. 1 always lead to better PSNR and SSIM scores, but over-smoothed results with bad perceptual quality. Therefore, we train SPARNet with $\mathcal{L}_{pix}$ to make fair comparison on distortion metrics (*i.e.*, PSNR and SSIM scores) with previous works.

D. Training Details

For SPARNet, we set the batch size as 64, and fix the learning rate at 2×10^{-4}. We use Adam [47] to optimize the model with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. For SPARNetHD, we empirically set $\lambda_{pix} = 100$, $\lambda_{adv} = 1$, $\lambda_{fm} = 10$, and $\lambda_{pcp} = 1$. The learning rates of G and D are 1×10^{-4} and 4×10^{-4} respectively. We use Adam to optimize both G and D with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The batch size is set to 2. Our models are implemented in PyTorch and run on a Tesla K40 GPU.

IV. Experiments

In this section, we conduct experiments for SPARNet and SPARNetHD respectively. First, we analyze the effectiveness of the proposed spatial attention mechanism in SPARNet, and compare SPARNet with previous face SR methods and general SR methods with different evaluation metrics using the same training data. Second, we evaluate the performance of SPARNetHD on real LR faces trained on synthetic datasets.

A. Analysis of SPARNet

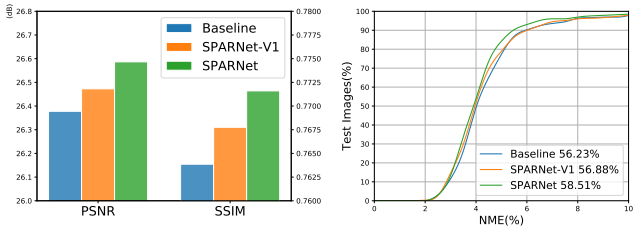
1) Datasets and Evaluation Metrics: **Training Data** We use CelebA [48] to train SPARNet. We first detect faces using MTCNN [49], and crop out the face regions roughly without any pre-alignment operation. Next, we select images larger than 128×128, resize them to 128×128 through bicubic interpolation, and use them as the HR training set. The LR training set is obtained by downsampling the HR images to 16×16. This results in roughly 179K image pairs. To avoid overfitting, we carry out data augmentation by random horizontal flipping, image rescaling (between 1.0 and 1.3), and image rotation (90°, 180° and 270°).

Testing Data Following previous works [1], [15], [50], we use the test set of Helen [51] for evaluation of image quality and landmark detection. For identity similarity evaluation, we use the same test set as in [20], which is specifically designed to preserve identity in face SR. This dataset contains randomly selected images of $1,000$ identities from UMD-Face [52].

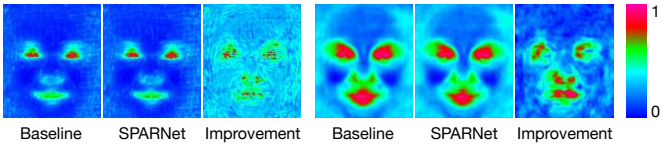
Following the practice of image super-resolution, the Peak Signal-to-Noise Ratio (PSNR) and Structure SIMilarity (SSIM) index calculated on the luminance channel are used as the primary quantitative evaluation metrics. Since two most important applications of face SR are alignment and identification of LR faces, we adopt two further metrics, namely landmark detection and identity similarity, to evaluate face SR performance.

Landmark Detection As mentioned before, structure recovery is very important in face SR. Similar to [1], [2], we use landmark detection accuracy for evaluation. Specifically, we use a popular landmark detection model, FAN[1] [40], which is pre-trained on HR faces, to detect landmarks on SR and HR faces. Following [40], we use the area under the curve (AUC) of the normalized mean error (NME) to quantify the performance.

[1]https://github.com/1adrianb/face-alignment



(a) Image Quality (b) Landmark Detection



(c) Visualization of the PSNR (left) and SSIM (right) error maps. The heatmaps for Baseline and SPARNet are the average 2D PSNR/SSIM error maps (larger value indicates larger error), and the improvement heatmaps show the error reduction (larger value indicates more error reduction).

Fig. 5: Quantitative comparison between Baseline, SPARNet-V1 and SPARNet on the Helen test set.

Identity Similarity Identity similarity measures how well identity information is preserved in a super-resolved face. Same as [20], we first extract identity feature vectors for the SR and HR faces using a pre-trained SphereFace model[2] [53], and then compute the identity similarity as the cosine distance between the two feature vectors.

2) Ablation Study: We compare the following three variants of our model:

- Baseline: residual SR network without any spatial attention branches.
- SPARNet-VN: To evaluate the effectiveness of using multiple FAU blocks, we keep the feature branch unchanged and vary the numbers of attention branch used in SPARNet [3]. Considering that decoder parts are supposed to be more sensitive to spatial attention, we gradually increase the number of attention branches from the backend. We denote models with N spatial attention branches as SPARNet-VN, where $N \in \{1, 2, 4, 8, 16\}$.
- SPARNet-SM: To demonstrate that it is crucial for the attention mechanism to learn from multi-scale features, we vary the numbers of downsample/upsample blocks to change the smallest size (scale) of feature maps in the bottleneck. We denote SPARNet with $M \times M$ size of feature maps in the bottleneck of attention branch as SPARNet-SM, where $M \in \{2, 4, 8, 16\}$ and smaller M indicates more scales of features are used.
- SPARNet: the full model used in this work, *i.e.*, SPARNet-V16-S4.

Effectiveness of Spatial Attention To evaluate the effectiveness of the proposed spatial attention mechanism, we compare

[2]https://github.com/clcarwin/sphereface_pytorch

[3]Reducing the numbers of entire FAU would make the network shallower which would definitely make the results worse, therefore we only vary the number of attention branch

the results of SPARNet-V1 and SPARNet with that of the Baseline model. Fig 5(a) shows that, with the proposed spatial attention mechanism, we can generate results with better image quality in terms of both PSNR and SSIM scores. This is expected as the spatial attention mechanism enables our network to focus on and better recover the key face structures. To visualize the improvement brought by the attention mechanism, we align the faces to a fixed template based on five key facial landmarks (*i.e.* two eye centers, nose, and two mouth corners), calculate 2D PSNR and SSIM error maps[4] for each image, and average the error maps over the whole test set. It can be observed from Fig. 5(c) that pixels corresponding to key face structures are most difficult to recover (*i.e.*, with larger errors), and as expected, most of the improvement happens around them. Furthermore, we compare the landmark detection results on the super-resolved images. Better recovery of key face structures should lead to higher landmark detection accuracy. It can be seen in Fig. 5(b) that SPARNet achieves better landmark detection accuracy (58.51% AUC) than the Baseline (56.23% AUC). This demonstrates the benefit of the spatial attention mechanism. Fig. 6 shows some SR results on the Helen test set. It can be observed that SPARNet can produce sharper and clearer face structures than the Baseline, especially for the eyes.

Number of FAUs By comparing the results of SPARNet-V1 with that of SPARNet in Fig. 5, we can see that SPARNet, which is composed entirely of FAUs, performs better than SPARNet-V1, which has only one FAU at the end. Although the attention map of SPARNet-V1 can focus on the key face structures, some noises also show up on the face region and cause some distraction (see Fig. 6 bottom left). This suggests that bootstrapping the key face structures using a single FAU may be sub-optimal. On the other hand, the sequence of FAUs in SPARNet allow the attention maps to gradually focus on different key face structures at different stages. For instances, one can see the attention map of the second last FAU focuses cleanly on the outline of the key face structures, and that of the last FAU focuses on the general face region (see Fig. 6 bottom right). In Fig. 7(a), we show more results comparison of SPARNet-VN. It can be observed that the evaluation PSNR scores of SPARNet-VN show a positive correlation with N, which demonstrates the effectiveness of stacking multiple FAUs.

Multiscale features in FAU The results of SPARNet-SM are shown in Fig. 7(b). We can see that models with more scales (*i.e.*, smaller M) generally perform better, but the improvement declines when M becomes smaller. The performance of SPARNet-S4 is similar to SPARNet-S2. Therefore, we set $M = 4$ for SPARNet to balance the performance and computation cost.

B. Comparisons of SPARNet against Other Methods

We compare SPARNet with state-of-the-art general image SR method RCAN [31], latest attention network for image classification CBAM [26], and other face SR methods, including URDGN [54], Wavelet-SRNet [50], Attention-FH [32], FSRNet [1], SICNN [20], PFSRNet [21] and DICNet [22]. When comparing with RCAN and CBAM, we replace our FAU with the corresponding attention module (borrowed from public released codes) and keep network depth the same for a fair comparison. *All models are trained using the same dataset except for FSRNet and SICNN.* Note that FSRNet only provides test code and SICNN relies on a much larger pre-aligned training set.

[4]The PSNR error is calculated by mean square error, and the SSIM error is calculated by subtracting the SSIM score from 1.

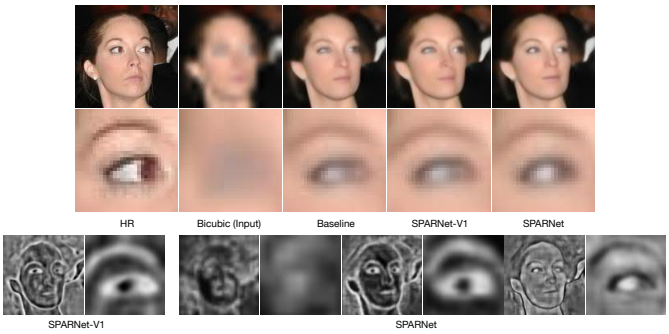
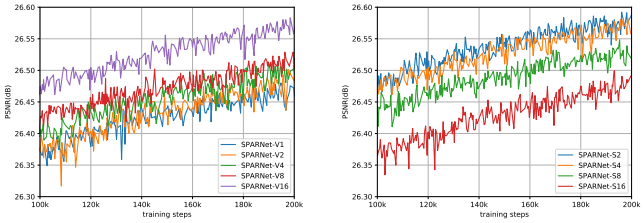


Fig. 6: Results on the Helen test set. Top: HR photo and SR results; bottom: spatial attention map in FAU of SPARNet-V1 (left) and last three FAUs of SPARNet (right).



(a) Evaluation PSNR score of SPARNet-VN during training.

(b) Evaluation PSNR score of SPARNet-SM during training.

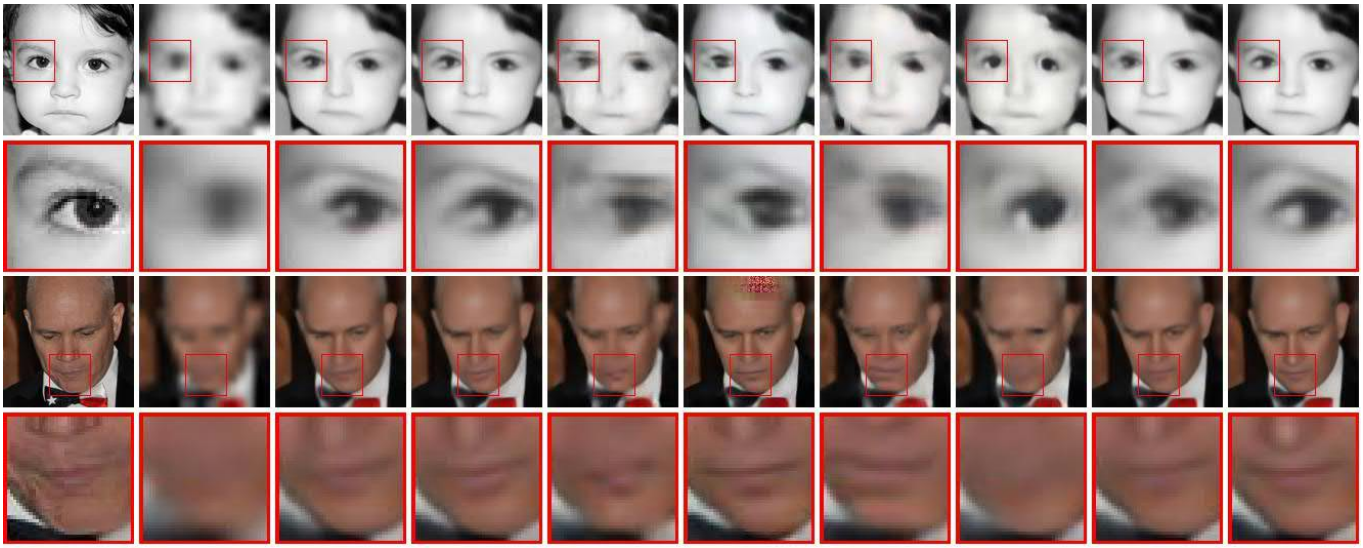
Fig. 7: Ablation study of SPARNet-VN and SPARNet-SM on Helen testset. We set $M = 4$ when training SPARNet-VN, and $N = 16$ for SPARNet-SM.

1) Overall Results: Quantitative comparison with other state-of-the-art methods on the Helen test set is shown in Table I. SPARNet clearly outperforms both generic SR methods and face SR methods in terms of both PSNR and SSIM scores. With the spatial attention mechanism, SPARNet can better recover important face structures. One can observe from Fig. 8 that while most of the other methods fail to recover the eyes and nose, and the shapes are blurry, SPARNet can generate sharper results with shapes close to the ground truth HR images.

2) Detailed Comparisons: **Attention Mechanisms** Among all the methods under comparison, only RCAN, CBAM, and Attention-FH have an attention mechanism. RCAN is designed for generic SR task. It applies a channel attention mechanism in the feature space. Although channel attention has been shown to be beneficial to general SR task, spatial structures are more important in very low resolution face super-resolution.

TABLE I: Quantitative comparison on the Helen (first 3 rows) and UMD (last row) test sets with a 16×16 input size and an upscale factor of $8\times$. The AUCs are calculated under a threshold of 10%. The results of FSRNet* are generated using the test model provided by its authors.

Method	Bicubic	RCAN	CBAM	URDGN	Wavelet	Att-FH	FSRNet*	Baseline	SPARNet
PSNR	23.52	26.40	**26.46**	25.17	26.42	25.10	24.97	26.38	**26.59**
SSIM	0.6408	0.7648	0.7666	0.7140	**0.7711**	0.7188	0.7091	0.7639	**0.7716**
AUC($< 10\%$)	4.42%	56.53%	56.83%	41.78%	**58.44%**	46.63%	40.42%	56.23%	**58.51%**
Identity	0.1851	0.5373	**0.5378**	0.3981	0.5147	0.4731	0.4765	0.5302	**0.5546**



(a) HR (b) Bicubic (c) RCAN [31] (d) CBAM [26] (e) URDGN [54] (f) Wavelet [50] (g) Att-FH [32] (h) FSRNet [3] (i) Baseline (j) SPARNet

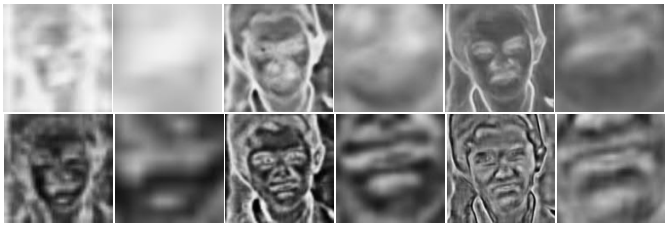
Fig. 8: Qualitative comparison with state-of-the-art methods. The resolution of the input is 16×16 and the upscale factor is $8\times$.

This explains why RCAN does not show big improvement over the Baseline in terms of PSNR score (26.40 vs 26.38). Attention-FH adopts a patch level attention and cannot recover details (see Fig 8(g)). CBAM also employs a spatial attention. However, it is manipulated to extract semantic information through pooling operations which cause loss of shape and edge details. We compare the SR results and spatial attention maps generated by SPARNet and CBAM in Fig 9. We can see that the attention maps produced by CBAM are blurry and not able to outline important details such as eyes and mouth, whereas the attention maps of SPARNet, thanks to the multi-scale feature based approach, provide more detailed structure information. Accordingly, the mouth and nose generated by SPARNet are sharper and clearer than that of CBAM.

Comparison with FSRNet [1] Since FSRNet only provides test code, the results of FSRNet reported in Table I are not obtained using the same training set as the others. To make a fair comparison, we retrain and test our model on the same dataset used by FSRNet (18K images from CelebA), and show the results in Table II. Although FSRNet uses face parsing map as an additional supervision, SPARNet still outperforms FSRNet in terms of PSNR and SSIM scores. This demonstrates the superiority of the proposed spatial attention mechanism.

Comparison with Wavelet-SRNet [50] Wavelet-SRNet shows closest performance to SPARNet in terms of SSIM score.

(a) SR results of CBAM and SPARNet.



(b) Attention maps of CBAM (top) and SPARNet (bottom).

Fig. 9: SR results and attention maps of CBAM and SPARNet.

TABLE II: Comparison between FSRNet and SPARNet on CelebA.

Method	FSRNet	SPARNet
PSNR/SSIM	26.31/0.7522	**26.68/0.7741**

Nonetheless, SPARNet requires much less parameters and is more computational efficient and flexible than Wavelet-SRNet. The performance and efficiency comparison between Wavelet-

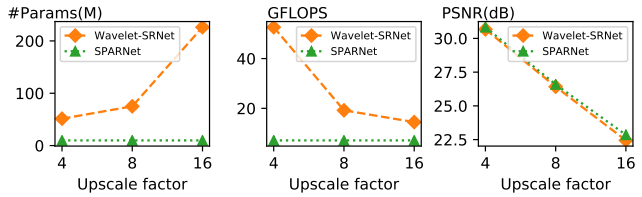


Fig. 10: Performance and efficiency comparison between Wavelet-SRNet and SPARNet under different upscale factors.

SRNet and SPARNet is summarized in Fig. 10. Wavelet-SRNet has 4^n parallel subnets with an upscale factor of 2^n. This implies the network parameters increase quadratically with the upscale factor. Besides, it requires different input size for each upscale factor. Because the output size is usually fixed, a smaller upscale factor needs a larger size of input when training the network. Since a larger input means more convolution operations and requires higher computational power, the GFLOPs of Wavelet-SRNet increase dramatically when the upscale factor is small. In contrast, SPARNet uses the same architecture for different upscale factors, and achieves better performance without increasing its computational complexity.

Comparison with landmark based methods We compare SPARNet with two recent landmark based methods PFSRNet[5] [21] and DICNet[6] [22]. To get the best results of the compared method, we directly use their public models instead of retraining them since both of them are trained on CelebA, the same as SPARNet. We evaluate the performance on the Helen test dataset provided by DICNet because our test images are cropped in a different way and DICNet performs bad on it. The quantitative results and visual examples are shown in Table III and Fig. 11 respectively. As we can observe from Table III, SPARNet shows the best PSNR and SSIM scores in the test dataset. The visual results in Fig. 11 indicate that SPARNet can recover the key face components, especially the eyes, better than DICNet even without any extra landmark information. We hypothesize that this is because it is too difficult to detect accurate landmarks for low resolution face images, and the multi-stage iterative process in DICNet amplifies the error. As for PFSRNet, we found it performs badly when test faces are not aligned. Since detecting landmarks from 16×16 LR face is by itself a difficult task, we find the practical value of PFSRNet is limited compared with DICNet and our SPARNet. We also show the computation complexity in Table III. Compared with state-of-the-art method DICNet, our method produces better results with much less parameters. Moreover, our SPARNet is faster than DICNet which is an iterative framework. Compared with PFSRNet, our method achieves much better performance with similar computation complexity.

Comparison with Different Upscale Factors We also trained models with $4\times$, and $16\times$ upscale factors. The $4\times$ model was trained from scratch, whereas the $16\times$ model was initialized with the pre-trained $8\times$ model. We show the results of $4\times$ and $16\times$ upscale factors in Table IV. It can be seen that SPARNet achieves state-of-the-art results for both $4\times$ and $16\times$ upscale factors, especially for the $16\times$ upscale factor.

[5]https://github.com/DeokyunKim/Progressive-Face-Super-Resolution
[6]https://github.com/Maclory/Deep-Iterative-Collaboration

TABLE III: Quantitative comparison on the Helen test dataset provided by DICNet [22].

Methods	PFSRNet [21]	DICNet [22]	SPARNet
PSNR	24.13	26.73	**26.97**
SSIM	0.6688	0.7955	**0.8026**
Params (M)	8.97	22.80	9.86
Time (s)	0.045	0.065	0.051

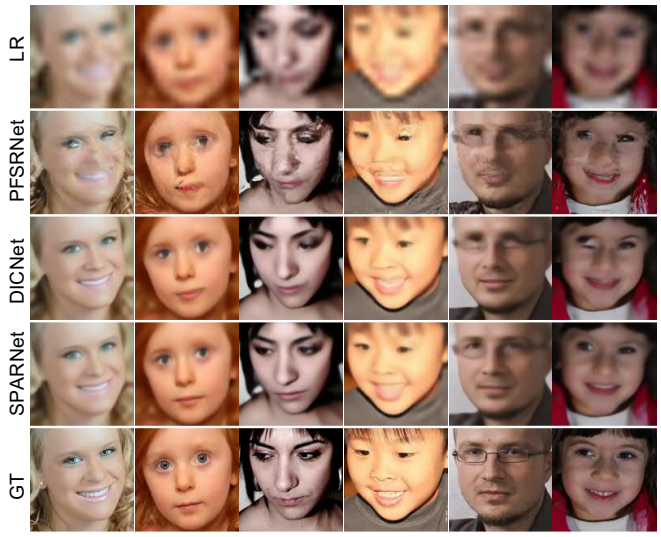


Fig. 11: Visual comparison between PFSRNet, DICNet, and SPARNet on Helen test dataset provided by DICNet. All results are generated by public models trained on CelebA. PFSRNet generates bad results for unaligned test faces. SPARNet generates better key face components than DICNet especially in the eyes.

TABLE IV: Image quality comparison on Helen with upscale factors $4\times$ and $16\times$

	Scale	Bicubic	RCAN	CBAM	Wavelet	SPARNet
PSNR	$4\times$	27.43	30.73	**30.76**	30.67	**30.83**
SSIM		0.8049	0.8857	0.8861	**0.8888**	**0.8872**
PSNR	$16\times$	20.22	20.71	**20.75**	22.44	**22.85**
SSIM		0.5180	0.6360	**0.6375**	0.6314	**0.6426**

3) Landmark Detection Results: As mentioned previously, the proposed spatial attention mechanism helps recover important face structures. Landmark detection can be exploited to evaluate such an ability. From Table I, we can see that SPARNet achieves the best performance in landmark detection. Wavelet-SRNet's performance is similar to SPARNet, but SPARNet is more efficient (refer to Fig. 10) and shows much better performance (in terms of PSNR/SSIM scores) for the $16\times$ upscale factor (refer to Table IV).

4) Identity Similarity Results: We first evaluate the models in Table I on the UMD-Face test set. Note that these models are trained on CelebA without pre-alignment. SPARNet achieves the best performance. We also notice that GAN based methods such as URDGN are worse than non-GAN based methods. This is because GAN based methods target at producing

realistic textures but they do not care whether the generated textures are consistent with the HR images. Therefore, they may generate textures which look sharper but disturb the identity information. To further explore how well SPARNet can preserve identity information, we retrain our model using the same dataset as SICNN to allow a fair comparison. We denote this model as SPARNet*. From Table V, we can see that SPARNet* shows better performance than SICNN even without using any explicit identity supervision. This proves that the proposed spatial attention mechanism is beneficial to preserving identity information in face SR.

TABLE V: Comparison of identity similarity between SICNN and SPARNet*. (SPARNet* is trained with the same dataset as SICNN.)

Method	SICNN [20]	SPARNet*
Identity Similarity	0.5978	**0.6272**

C. Evaluation of SPARNetHD

Different from SPARNet which aims to hallucinate very low resolution faces (*i.e.*, 16×16) that are difficult to recognize, the extended SPARNetHD tries to deal with real world low resolution faces. Typical application scenarios include old photos and unclear faces shot by low-end devices. These LR faces usually do not have fixed upscale factors and are usually noisy and blurry. Hence, we need a different degradation model for data synthesis. We also need high resolution datasets in order to get high resolution outputs.

1) Degradation Model: According to previous works [55], [56] and common practice in single image SR framework, we generate the LR image I_l^s from the HR image I_h using the following degradation model:

$$I_l^s = ((I_h * \mathbf{k}_\varrho) \downarrow_s + \mathbf{n}_\delta)_{JPEG_q}, \tag{14}$$

where $*$ represents the convolution operation between I_h and a blur kernel $\mathbf{k}_\varrho$ with parameter ϱ. $\downarrow_s$ is the downsampling operation with a scale factor s. $\mathbf{n}_\delta$ denotes the additive white Gaussian noise (AWGN) with a noise level δ. $(\cdot)_{JPEG_q}$ indicates the JPEG compression operation with quality factor q. The hyper parameters ϱ, s, δ, q are randomly selected for each HR image I_h, and I_l^s is generated online. We set the hyper parameters as

- $\mathbf{k}_\varrho$ is the blur kernel. We randomly choose one of the following four kernels: Gaussian Blur ($3 \leq \varrho \leq 15$), Average Blur ($3 \leq \varrho \leq 15$), Median Blur ($3 \leq \varrho \leq 15$), Motion Blur ($5 \leq \varrho \leq 25$);
- $\downarrow_s$ is the downsample operation. The scale factor s is randomly selected in $[\frac{16}{512}, \frac{128}{512}]$;
- $\mathbf{n}_\delta$ is addictive white gaussian noise (AWGN) with $0 \leq \delta \leq 0.1 \times 255$;
- $JPEG_q$ is the JPEG operation. The compression level is randomly chosen from $[60, 85]$, in which higher means stronger compression level.

After obtaining I_l^s, $I_{LR-UP} = (I_l^s) \uparrow_s$ serves as the LR input of SPARNetHD.

2) Datasets and Evaluation Metrics: **Training data** We adopt the FFHQ [34] dataset for training. This dataset consists of $70,000$ high-quality images at a size of 1024×1024 crawled from the internet. All images are automatically cropped and aligned. We resize the images to 512×512 with bilinear downsampling as the ground-truth HR images, and synthesize the LR inputs online with Eq. 14.

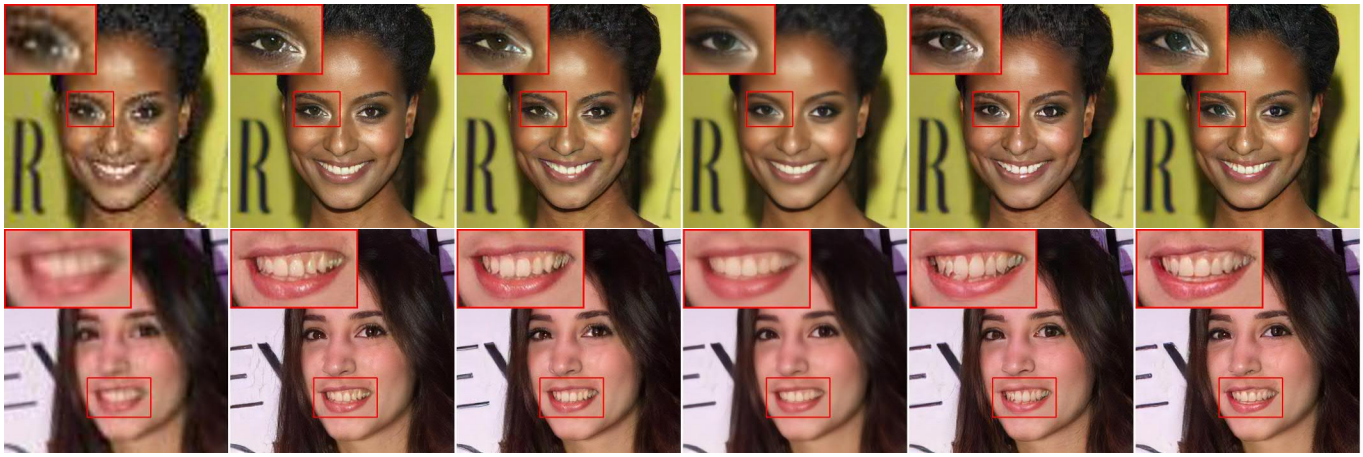
Synthetic Test Data We use the test partition of CelebAHQ dataset [33] as synthetic test data. CelebAHQ contains $30,000$ HR faces in total which are split into a training set ($24,183$ images), a validation set ($2,993$ images), and a test set ($2,824$ images). We generate the LR dataset with Eq. 14, denoted as CelebAHQ-Test.

Natural Test Data We collect $1,020$ faces smaller than 48×48 from CelebA. We also collect some old photos from the internet for testing. All images are cropped and aligned in the same manner as FFHQ, and then resized to 512×512 using bicubic upsampling. We merge all these images together and create a new dataset containing $1,051$ natural LR faces, denoted as CelebA-TestN.

Evaluation Metrics For CelebAHQ-Test, we use LPIPS (Learned Perceptual Image Patch Similarity) [57] as our evaluation metric because it can better represent the visual quality of HR images than PSNR and SSIM. For CelebA-TestN, since there are no ground truth HR images, we use FID (Fréchet Inception Distance) [58] to measure the similarity between the SR results and reference datasets containing HR face images. We use the HR version of CelebAHQ-Test as the reference dataset.

3) Ablation Study of Loss Functions: To verify the effectiveness of different loss terms in Eq. 12, we conduct an ablation study by removing each of them separately. Same as SPARNetHD, all models are trained for 100k iterations which take about 3 days on a single GPU. The comparison results are shown in Table VI and Fig. 12. We can observe that $\mathcal{L}_{pix}$ has the least influence on the final performance because it mainly affects the subtle low level details. Comparing the results of second column and third column, the skin color of SPARNetHD (w/o $\mathcal{L}_{pix}$) is mixed with slight background green color compared with full SPARNetHD in the first row, and teeth shape is less natural w/o $\mathcal{L}_{pix}$ in the second row (please zoom in to see the details). From the 4th column, we can see that the network fails to generate sharp edges and realistic details without adversarial loss, and these greatly degrade the quantitative performance. In the 5th column, we can observe that the generated details are less realistic and there are some noise-like artifacts near the hair in the first row. This is mainly because the GAN network trained without $\mathcal{L}_{fm}$ is less stable. In the last column, we can see obvious shape distortions in the eyes and lips since $\mathcal{L}_{pcp}$ mainly helps to constrain mid-level and high-level semantics.

4) Comparison with Other Methods: We conduct experiments to compare SPARNetHD with other methods on CelebAHQ-Test and CelebA-TestN. Since very few previous face SR methods can produce high resolution outputs, we mainly compare SPARNetHD with general SR methods and GAN based image-to-image translation methods which provide public codes and can be used to generate high resolution



LR Input SPARNetHD w/o $\mathcal{L}_{pix}$ w/o $\mathcal{L}_{GAN}$ w/o $\mathcal{L}_{fm}$ w/o $\mathcal{L}_{pcp}$

Fig. 12: Ablation study of losses used in SPARNetHD on the synthetic dataset CelebAHQ-Test (first row) and real dataseet CelebA-TestN (second row). Better zoom in to see the details.

(a) Synthetic LR (b) ESRGAN (c) Pix2PixHD (d) SFTGAN (e) Ours (f) GT

Fig. 13: Visual comparison between results produced by SPARNetHD and other methods on the synthetic dataset CelebAHQ-Test. Better zoom in to see the details.

outputs. Specifically, we compare SPARNetHD with state-of-the-art SR method ESRGAN [59], parsing map based SR method SFTGAN [60], high resolution image translation method Pix2PixHD [5], and blind face restoration method based on a reference image GFRNet [55]. We retrain ESRGAN, SFTGAN and Pix2PixHD. We use existing face parsing model to generate face parsing map for SFTGAN. As for GFRNet, we only carry out a visual comparison with the results reported in the original paper because it needs reference images for training and testing.

Table VI gives a quantitative comparison on CelebAHQ-Test and CelebA-TestN. We can observe that SPARNetHD outperforms the other methods on both datasets. Pix2PixHD performs better than ESRGAN because of the multi-scale discriminators. The fact that the results of SPARNetHD are much better than that of Pix2PixHD demonstrates the effectiveness of the proposed attention mechanism, which is the key difference between SPARNetHD and Pix2PixHD. SFTGAN shows slightly better performance than Pix2PixHD with the help of face parsing maps. This shows the advantage of taking parsing map as additional supervision. Nevertheless, the proposed SPARNetHD outperforms SFTGAN by a large margin. This is because parsing map can only provide coarse semantic level guidance, while the proposed spatial attention

TABLE VI: Quantitative comparison on CelebAHQ-Test and CelebA-TestN. *Note for w/o $\mathcal{L}_{GAN}{}^*$: because L_{fm} depends on the discriminator, we also remove it when $\mathcal{L}_{GAN}$ is not used.*

Methods	CelebAHQ-Test (LPIPS↓)	CelebA-TestN (FID↓)
ESRGAN [59]	0.49	60.67
SFTGAN [60]	0.36	37.76
Pix2PixHD [5]	0.36	43.1
SPARNetHD	**0.29**	**27.16**
w/o $\mathcal{L}_{pix}$	0.31	30.63
w/o $\mathcal{L}_{GAN}{}^*$	0.38	39.75
w/o $\mathcal{L}_{fm}$	0.32	33.52
w/o $\mathcal{L}_{pcp}$	0.31	32.97

(a) Real LR (b) ESRGAN (c) Pix2PixHD (d) SFTGAN (e) Ours

Fig. 14: Visual comparison between results produced by SPARNetHD and other methods on real LR dataset CelebA-TestN. Better zoom in to see the details.

mechanism cannot only provide coarse semantic guidance but also give low level texture guidance.

The qualitative results in Fig. 13 and Fig 14 are consistent with the observations in Table VI. Fig. 13 shows the results on the synthetic test dataset CelebAHQ-Test. We can see that ESRGAN fails to generate realistic faces. While the results of Pix2PixHD look much better, undesirable artifacts show up in the right eyes. SFTGAN does not have such problem, but it cannot generate detailed textures, especially in the hair and teeth. In contrast, SPARNetHD can restore key face components as well as the texture details in the hair and teeth. Similar phenomenon can also be observed in Fig. 14, which shows the results on natural LR faces. While the competitive methods generate many artifacts for real LR faces, our results are much more robust and natural. This illustrates the good generalization ability of SPARNetHD.

We also compare SPARNetHD with a recent blind face restoration method GFRNet. GFRNet takes a LR image and a HR image of the same person as input to restore the LR image. We show some qualitative results on natural LR faces in Fig. 15. It can be observed that SPARNetHD generates much better texture details than GFRNet even without utilizing any additional information in the training stage.

(a) Real LR (b) GFRNet (c) Ours

Fig. 15: Visual comparison between results produced by SPARNetHD and GFRNet. Better zoom in to see the details.

V. Conclusion

We propose a SPatial Attention Residual Networks (SPARNet) for very low resolution face super-resolution. SPARNet is composed by stacking Face Attention Units (FAUs), which extend vanilla residual block with a spatial attention branch. The spatial attention mechanism allows the network to pay less attention on the less feature-rich regions. This makes the training of SPARNet more effective and efficient. Extensive experiments with various kinds of metrics illustrate the advantages of SPARNet over current state-of-the-arts. We further extend SPARNet to SPARNetHD with more channel numbers and multi-scale discriminator networks. SPARNetHD trained on synthetic datasets is able to generate realistic and high resolution outputs (*i.e.*, 512×512) for LR face images. Quantitative and qualitative comparisons with other methods indicate that the proposed spatial attention mechanism is beneficial to restore texture details of LR face images. We also demonstrate that SPARNetHD can generalize well to real world LR faces, making it highly practical and applicable.

Acknowledgments

The authors would like to thank Tencent AI Lab for supporting this research. The work of Kwan-Yee K. Wong was supported by a grant from the Research Grant Council of the Hong Kong (SAR), China, under the Project HKU 17203119.

REFERENCES

[1] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 5, 6, 7

[2] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 5

[3] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. HARTLEY, "Face super-resolution guided by facial component heatmaps," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 217–233. 1, 2, 7

[4] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[5] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4, 10

[6] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 9, pp. 1167–1183, 2002. 2

[7] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 3, pp. 425–434, 2005. 2

[8] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision (IJCV)*, vol. 75, no. 1, pp. 115–134, 2007. 2

[9] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognition (PR)*, vol. 43, no. 6, pp. 2224–2236, 2010. 2

[10] C. Liu, H.-Y. Shum, and C.-S. Zhang, "A two-step approach to hallucinating faces: global parametric model and local nonparametric model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 2

[11] W. Liu, D. Lin, and X. Tang, "Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2005, pp. 478–484. 2

[12] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 236–249. 2

[13] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1099–1106. 2

[14] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 4537–4543. 2

[15] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *European Conference on Computer Vision (ECCV)*, 2016. 2, 5

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680. 2

[17] X. Yu and F. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks." in *AAAI*, vol. 2, 2017, p. 3. 2

[18] ——, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3760–3768. 2

[19] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[20] K. Zhang, Z. ZHANG, C.-W. Cheng, W. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 183–198. 2, 5, 6, 9

[21] K. Deokyun, K. Minseon, K. Gihyun, and K. Dae-Shik, "Progressive face super-resolution via attention to facial landmark," in *Proceedings of the 30th British Machine Vision Conference (BMVC)*, 2019. 2, 6, 8

[22] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6, 8

[23] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, 2014, pp. 2204–2212. 2

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6450–6458. 2

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 2018. 2, 6, 7

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057. 2

[28] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6298–6306. 2

[29] I. Schwartz, A. Schwing, and T. Hazan, "High-order attention models for visual question answering," in *Advances in Neural Information Processing Systems*, 2017, pp. 3664–3674. 2

[30] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4187–4195. 2

[31] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conference on Computer Vision (ECCV)*, 2018. 3, 6, 7

[32] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1656–1664. 3, 6, 7

[33] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. 3, 9

[34] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410. 3, 9

[35] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *arXiv preprint arXiv:1912.04958*, 2019. 3

[36] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[37] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and S. Wenzhe, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4

[38] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 3, 4

[39] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 483–499. 4

[40] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision (ICCV)*, 2017. 4, 5

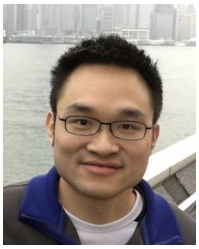
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *arXiv preprint arXiv:1603.05027*, 2016. 4

[42] ——, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034. 4

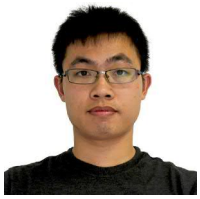
[43] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 391–407. 4

[44] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard 4

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 4
[46] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," 2018, pp. 6228–6237. 4
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5
[48] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision (ICCV)*, 2015. 5
[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016. 5
[50] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1689–1697. 5, 6, 7
[51] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2013. 5
[52] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "Umdfaces: An annotated face dataset for training deep networks," *arXiv preprint arXiv:1611.01484v2*, 2016. 5
[53] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5
[54] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 318–333. 6, 7
[55] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *European Conference on Computer Vision (ECCV)*, September 2018. 9, 10
[56] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 251–260. 9
[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018. 9
[58] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637. 9
[59] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 10
[60] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 10



Chaofeng Chen is currently a Ph.D. student in the Department of Computer Science, The University of Hong Kong. He received his B.E. degree from Huazhong University of Science and Technology in 2015. His research interests are centered around computer vision and deep learning.



Dihong Gong received the Ph.D. degree of computer science from the University of Florida, in 2018. He then joined the Tencent AI Lab as a senior research scientist, with research interest primarily focused on face related technologies, including face detection, recognition, liveness examination, etc.

Hao Wang is currently a senior researcher with Tencent AI Lab. He received the M.Eng. degree from the University of Science and Technology of China in 2016. His research interests include computer vision, deep learning, and face recognition.

Zhifeng Li (M'06-SM'11) is currently a top-tier principal researcher with Tencent. He received the Ph. D. degree from the Chinese University of Hong Kong in 2006. After that, He was a postdoctoral fellow at the Chinese University of Hong Kong and Michigan State University for several years. Before joining Tencent, he was a full professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include deep learning, computer vision and pattern recognition, and face detection and recognition. He is currently serving on the Editorial Boards of Neurocomputing and IEEE Transactions on Circuits and Systems for Video Technology. He is a fellow of British Computer Society (FBCS).

Kwan-Yee K. Wong received the BEng degree (Hons.) in computer engineering from The Chinese University of Hong Kong, in 1998, and the MPhil and PhD degrees in computer vision (information engineering) from the University of Cambridge, in 2000 and 2001, respectively. Since 2001, he has been with the Department of Computer Science at The University of Hong Kong, where he is currently an associate professor. His research interests are in computer vision and machine intelligence. He is currently an editorial board member of International Journal of Computer Vision (IJCV).