FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors

Yu Chen[1*] Ying Tai[2*] Xiaoming Liu[3] Chunhua Shen[4] Jian Yang[1]
[1]Nanjing University of Science and Technology [2]Youtu Lab, Tencent
[3]Michigan State University [4]University of Adelaide

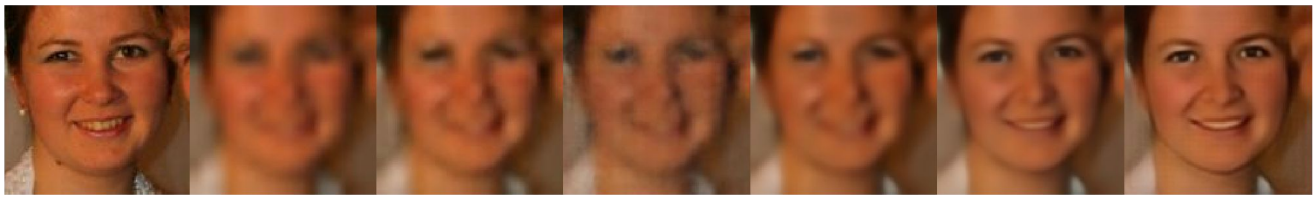

Figure 1: Visual results of different super-resolution methods on scale factor 8.

Abstract

Face Super-Resolution (SR) is a domain-specific super-resolution problem. The specific facial prior knowledge could be leveraged for better super-resolving face images. We present a novel deep end-to-end trainable Face Super-Resolution Network (FSRNet), which makes full use of the geometry prior, i.e., facial landmark heatmaps and parsing maps, to super-resolve very low-resolution (LR) face images without well-aligned requirement. Specifically, we first construct a coarse SR network to recover a coarse high-resolution (HR) image. Then, the coarse HR image is sent to two branches: a fine SR encoder and a prior information estimation network, which extracts the image features, and estimates landmark heatmaps/parsing maps respectively. Both image features and prior information are sent to a fine SR decoder to recover the HR image. To further generate realistic faces, we propose the Face Super-Resolution Generative Adversarial Network (FSRGAN) to incorporate the adversarial loss into FSRNet. Moreover, we introduce two related tasks, face alignment and parsing, as the new evaluation metrics for face SR, which address the inconsistency of classic metrics w.r.t. visual perception. Extensive benchmark experiments show that FSRNet and FSRGAN significantly outperforms state of the arts for very LR face SR, both quantitatively and qualitatively. Code will be made available upon publication.

1. Introduction

Face Super-Resolution (SR), a.k.a. face hallucination, aims to generate a High-Resolution (HR) face image from a Low-Resolution (LR) input. It is a fundamental problem in face analysis, which can greatly facilitate face-related tasks, e.g., face alignment [16, 36], face parsing [23], and face recognition [34, 40], since most existing techniques would degrade substantially when given very LR face images.

As a special case of general image SR, there exists face-specific prior knowledge in face images, which can be pivotal for face SR and is unavailable for general image SR [22, 32, 33]. For example, facial correspondence field could help recover accurate face shape [45], and facial components reveal rich facial details [31, 39]. However, as compared in Tab. 1, the previous face SR methods that utilize facial priors all adopt multi-stage, rather than end-to-end, training strategies, which is inconvenient and complicated.

Based on deep Convolutional Neural Network (CNN), in this work, we propose a novel *end-to-end trainable Face Super-Resolution Network (FSRNet)*, which estimates facial landmark heatmaps and parsing maps during training, and then uses these prior information to better super-resolve very LR face images. It is a consensus that end-to-end training is desirable for CNN [16], which has been validated in many areas, e.g., speech recognition [8] and image recognition [20]. Unlike previous Face SR methods that estimate local solutions in separate stages, our end-to-end framework learns the global solution directly, which is more convenient and elegant. To be specific, since it is non-trivial to estimate facial landmarks and parsing maps directly from LR inputs, we first construct a coarse SR network to recover a coarse HR image. Then, the coarse HR image is sent to a fine SR

*indicates equal contributions. This work was partially done when Yu Chen was visiting University of Adelaide.

Method	VDSR [17] (CVPR'16)	SRResNet [22] (CVPR'17)	StructuredFH [39] (CVPR'13)	CBN [45] (ECCV'16)	URDGN [41] (ECCV'16)	AttentionFH [2] (CVPR'17)	LCGE [31] (IJCAI'17)	FSRNet (ours)
Facial Prior KNWL	×	×	Components	Dense corres. field	×	×	Components	Landmark/parsing maps
Deep Model	√	√	×	√	√	√	√	√
End-to-End	√	√	×	×	√	√	×	√
Unaligned	√	√	×	√	×	×	×	√
Scale Factor	$2/3/4$	$2/4$	4	$2/3/4$	8	$4/8$	4	8

Table 1: Comparisons with previous state-of-the-art super-resolution methods, where VDSR and SRResNet are generic image SR methods, and StructuredFH, CBN, URDGN, AttentionFH and LCGE are face SR methods.

network, where a *fine SR encoder* and a *prior estimation network* share the coarse HR image as the input, followed by a *fine SR decoder*. The fine SR encoder extracts the image features, while the prior estimation network estimates landmark heatmaps and parsing maps jointly, via multi-task learning. After that, the image features and facial prior knowledge are fed into a fine SR decoder to recover the final HR face. The coarse and fine SR networks constitute our basic FSRNet, which already significantly outperforms the state of the arts (Fig. 1). To further generate realistic HR faces, *Face Super-Resolution Generative Adversarial Network (FSRGAN)* is introduced to incorporate the adversarial loss into the basic FSRNet. As in Fig. 1, FSRGAN recovers more realistic textures than FSRNet, and clearly shows superiority over the others.

It's a consensus that Generative Adversarial Network (GAN)-based models recover visually plausible images but may suffer from low Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) or other quantitative metrics, while Mean Squared Error (MSE)-based deep models recover smooth images but with high PSNR/SSIM. To quantitatively show the superiority of GAN-based model, in [22], the authors asked 26 users to conduct a mean opinion score testing. However, such a testing is not objective and difficult to follow for fair comparison. To address this problem, we introduce two related face analysis tasks, *face alignment and parsing*, as the new evaluation metrics for face SR, which are demonstrated to be suitable for both MSE and GAN-based models.

In summary, the main contributions of this work include:

• To the best of our knowledge, this is the *first* deep face super-resolution network utilizing *facial geometry prior* in a convenient and elegant *end-to-end training* manner.

• Two kinds of facial geometry priors: *facial landmark heatmaps* and *parsing maps* are introduced simultaneously.

• The proposed FSRNet achieves state-of-the-art performance when hallucinating *unaligned* and *very low-resolution* (16×16 pixels) face images by an upscaling factor of 8, and the extended FSRGAN further generates more realistic face images.

• Face alignment and parsing are adopted as the *novel evaluation metrics* for face super-resolution, which are further demonstrated to resolve the inconsistency of classic metrics w.r.t. the visual perception.

2. Related Work

We review the prior works from two perspectives, and contrast with the most relevant papers in Tab. 1.

Facial Prior Knowledge There are many face SR methods that use facial prior knowledge to better super-resolve LR faces. Early techniques assume that faces are in a controlled setting with small variations. Baker and Kanade [1] proposed to learn a prior on the spatial distribution of the image gradient for frontal face images. Wang et al. [37] implemented the mapping between LR and HR faces by an eigen transformation. Kolouri et al. [18] learnt a nonlinear Lagrangian model for HR face images, and enhanced the degraded image by finding the model parameters that could best fit the given LR data. Yang et al. [39] incorporated the face priors by using the mapping between specific facial components. However, the matchings between components are based on the landmark detection results that are difficult to estimate when the down-sampling factor is large.

Recently, deep convolutional neural networks have been successfully applied to the face SR task. Zhu et al. [45] super-resolved very LR and unaligned faces in a task-alternating cascaded framework. In their framework, face hallucination and dense correspondence field estimation are optimized alternatively. Besides, Song et al. [31] proposed a two-stage method, which first generated facial components by CNNs and then synthesized fine-grained facial structures through a component enhancement method. Different from the above methods that conduct face SR in multiple steps, our FSRNet fully leverages facial landmark heatmaps and parsing maps in an end-to-end training manner.

End-to-end Training End-to-end training is widely used in general image SR. Tai et al. [32] proposed Deep Recursive Residual Network (DRRN) to address the problems of model parameters and accuracy, which recursively learns the residual unit in a multi-path model. The authors also proposed a deep end-to-end persistent memory network to address the long-term dependency problem in CNN for image restoration [33]. Moreover, Ledig et al. [22] proposed Super-Resolution Generative Adversarial Network (SRGAN) for photo-realistic image SR using a perceptual loss function that consists of an adversarial loss and a content loss.

There are also many face SR methods adopting the end-

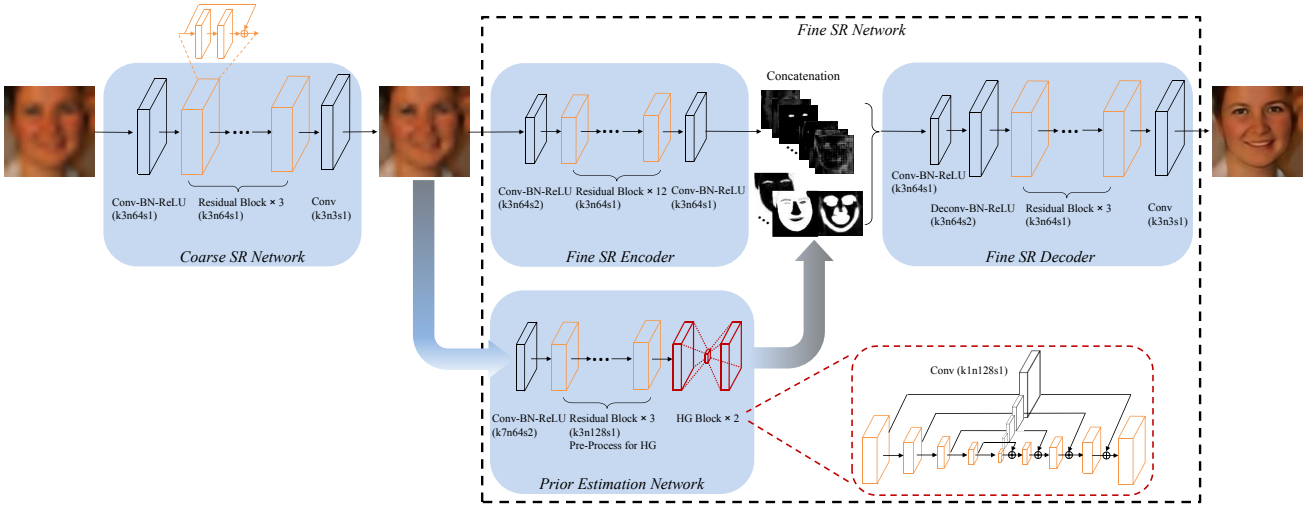

Figure 2: Network structure of the proposed FSRNet. 'Conv-BN-ReLU' indicates a convolutional layer, followed by Batch Normalization (BN) [12] and ReLU [26]. 'k3n64s1' indicates the kernel size to be 3×3, the feature map number to be 64 and the stride to be 1. We estimate a landmark by a heatmap , but for convenience of display, we show all landmarks in one heatmap here.

to-end training strategy. Yu et al. [41] investigated GAN [7] to create perceptually realistic HR face images. The authors further proposed transformative discriminative autoencoder to super-resolve unaligned, noisy and tiny LR face images [42]. More recently, Cao et al. [2] proposed an attention-aware face hallucination framework, which resorts to deep reinforcement learning for sequentially discovering attended patches and then performing the facial part enhancement by fully exploiting the global image interdependency. Different from the above methods that only rely on the power of deep models, our FSRNet is not only an end-to-end trainable Neural Network, but also combines the rich information from the facial prior knowledge.

3. Face Super-Resolution Network

3.1. Overview of FSRNet

Our basic FSRNet $\mathbf{F}$ consists of four parts: *coarse SR network*, *fine SR encoder*, *prior estimation network* and finally a *fine SR decoder*. Denote $\mathbf{x}$ as the low-resolution input image, $\mathbf{y}$ and $\mathbf{p}$ as the recovered high-resolution image and estimated prior information by FSRNet.

Since the very low-resolution input image may be too indistinct for prior estimation, we first construct the coarse SR network to recover a coarse SR image,

$$\mathbf{y}_c = \mathcal{C}(\mathbf{x}), \tag{1}$$

where $\mathcal{C}$ denotes the mapping from a LR image $\mathbf{x}$ to a coarse SR image $\mathbf{y}_c$ by the coarse SR network. Then, $\mathbf{y}_c$ is sent to the prior estimation network $\mathcal{P}$ and fine SR encoder $\mathcal{F}$ respectively,

$$\mathbf{p} = \mathcal{P}(\mathbf{y}_c),\ \mathbf{f} = \mathcal{F}(\mathbf{y}_c), \tag{2}$$

where $\mathbf{f}$ is the features extracted by $\mathcal{F}$. After encoding, the SR decoder $\mathcal{D}$ is utilized to recover the SR image by *concatenating* the image feature $\mathbf{f}$ and prior information $\mathbf{p}$,

$$\mathbf{y} = \mathcal{D}(\mathbf{f}, \mathbf{p}), \tag{3}$$

Given a training set $\{\mathbf{x}^{(i)}, \tilde{\mathbf{y}}^{(i)}, \tilde{\mathbf{p}}^{(i)}\}_{i=1}^{N}$, where N is the number of training images, $\tilde{\mathbf{y}}^{(i)}$ is the ground-truth high-resolution image of the low-resolution image $\mathbf{x}^{(i)}$ and $\tilde{\mathbf{p}}^{(i)}$ is the corresponding ground-truth prior information, the loss function of our FSRNet is

$$\begin{aligned}\mathcal{L}_{\mathbf{F}}(\Theta) = &\frac{1}{2N}\sum_{i=1}^{N}\{\|\tilde{\mathbf{y}}^{(i)} - \mathbf{y}_c^{(i)}\|^2 + \|\tilde{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\|^2 \\ &+ \lambda\|\tilde{\mathbf{p}}^{(i)} - \mathbf{p}^{(i)}\|^2\},\end{aligned} \tag{4}$$

where Θ denotes the parameter set, λ is the weight of prior loss, and $\mathbf{y}^{(i)}, \mathbf{p}^{(i)}$ are the recovered HR image and estimated prior information of the i-th image respectively.

3.2. Details inside FSRNet

We now present the details of our FSRNet, which consists of a coarse and a fine SR network, where the fine SR network contains three parts: a prior estimation network, a fine SR encoder and a fine SR decoder.

3.2.1 Coarse SR network

First, we use a coarse SR network to roughly recover a coarse HR image. The motivation is that it is non-trivial to estimate facial landmark positions and parsing maps directly from a LR input image. Using the coarse SR network may help to ease the difficulties for estimating the priors. The architecture of the coarse SR network is shown in

Fig. 2. It starts with a 3×3 convolution followed by 3 *residual blocks* [10]. Then another 3×3 convolutional layer is used to reconstruct the coarse HR image.

3.2.2 Fine SR Network

In the following fine SR network, the coarse HR image is sent to two branches, prior estimation network and fine encoder network, to estimate facial priors and extract features, respectively. Then the decoder jointly uses results of both branches to recover the fine HR image.

Prior Estimation Network Any real-world object has distinct distributions in its shape and texture, including face. Comparing facial shape with texture, we choose to model and leverage the shape prior for two considerations. First, when reducing the resolution from high to low, the shape information is better preserved compared to the texture, and hence is more likely to be extracted to facilitate super-resolution. Second, it is much easier to represent shape prior than texture prior. For example, face parsing estimates the segmentations of different face components, and landmarks provide the accurate locations of facial keypoints. Both represent facial shapes, while parsing carries more granularity. In contrast, it is not clear how to represent the higher-dimensional texture prior for a specific face.

Inspired by the recent success of stacked heatmap regression in human pose estimation [3, 27], we adopt the Hour-Glass (HG) structure to estimate facial landmark heatmaps and parsing maps in our prior estimation network. Since both priors represent the 2D face *shape*, in our prior estimation network, *the features are all shared between these two tasks*, except the last layer. The detailed structure of prior estimation network is shown in Fig. 2. To effectively consolidate features across scales and preserve spatial information in different scales, the hourglass block uses a skip connection mechanism between symmetrical layers. An 1×1 convolution layer follows to post-process the obtained features. Finally, the shared hourglass feature is connected to two separate 1×1 convolution layers to generate the landmark heatmaps and the parsing maps.

Fine SR Encoder For fine SR encoder, inspired by the success of ResNet [10] in SR [22, 32], we utilize the residual blocks for feature extraction. Considering the computation cost, the size of our prior features is down-sampled to 64×64. To make the feature size consistent, the fine SR encoder starts with a 3×3 convolutional layer of stride 2 to down-sample the feature map to 64×64. Then the ResNet structure is utilized to extract image features.

Fine SR Decoder The fine SR decoder jointly uses the features and priors to recover the final fine HR image. First, the prior feature $\mathbf{p}$ and image feature $\mathbf{f}$ are concatenated as the input of the decoder. Then a 3×3 convolutional layer reduces the number of feature maps to 64. A 4×4 deconvolutional layer is utilized to up-sample the feature map to size 128×128. Then 3 residual blocks are used to decode the features. Finally, a 3×3 convolutional layer is used to recover the fine HR image.

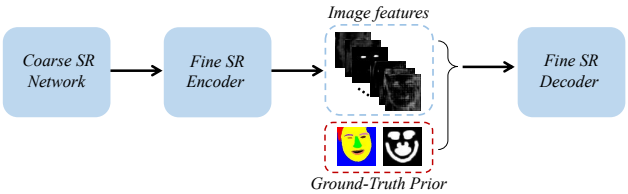

Figure 3: Structure of "upper-bound" model. The ground-truth priors are directly concatenated with image features. Removing priors in the red box and increasing the number of image features by the number of channels in prior induce to the baseline model.

3.3. FSRGAN

As we know, GAN has shown great power in super-resolution [22], which can generate photo-realistic images with superior visual effect than MSE-based deep models. The key idea is to use a discriminative network to distinguish the super-resolved images and the real high-resolution images, and to train the SR network to deceive the discriminator.

To generate realistic high-resolution faces, our model utilizes GAN in the conditional manner [13]. The objective function of the adversarial network $\mathbf{C}$ is expressed as:

$$\mathcal{L}_{\mathbf{C}}(\mathbf{F}, \mathbf{C}) = \mathbb{E}[\log \mathbf{C}(\tilde{\mathbf{y}}, \mathbf{x})] + \mathbb{E}[\log(1 - \mathbf{C}(\mathbf{F}(\mathbf{x}), \mathbf{x})], \quad (5)$$

where $\mathbf{C}$ outputs the probability of the input been real and $\mathbb{E}$ is the expectation of the probability distribution. Apart from the adversarial loss $\mathcal{L}_{\mathbf{C}}$, we further introduce a perceptual loss [15] using high-level feature maps (i.e., features from 'relu5_3' layer) of the pre-trained VGG-16 network [30] to help assess perceptually relevant characteristics,

$$\mathcal{L}_{\mathbf{P}} = \|\phi(\mathbf{y}) - \phi(\tilde{\mathbf{y}})\|^2, \quad (6)$$

where ϕ denotes the *fixed* pre-trained VGG model, and maps the images $\mathbf{y}/\tilde{\mathbf{y}}$ to the feature space. In this way, the final objective function of FSRGAN is:

$$\arg \min_{\mathbf{F}} \max_{\mathbf{C}} \mathcal{L}_{\mathbf{F}}(\Theta) + \gamma_{\mathbf{C}}\, \mathcal{L}_{\mathbf{C}}(\mathbf{F}, \mathbf{C}) + \gamma_{\mathbf{P}}\, \mathcal{L}_{\mathbf{P}}, \quad (7)$$

where $\gamma_{\mathbf{C}}$ and $\gamma_{\mathbf{P}}$ are the weights of GAN and perceptual loss, respectively.

4. Prior Knowledge for Face Super-Resolution

In this section, we would like to answer two questions: (1) Is facial prior knowledge really useful for face super-resolution? (2) How much improvement does different facial prior knowledge bring? To answer these questions, we conduct several tests on the 2,330-image Helen

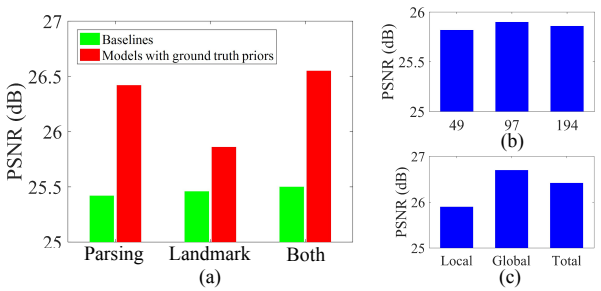

Figure 4: Effects of facial prior knowledge. (a) Comparisons between baselines and models with ground truth priors. The upper bound performance of landmark priors with different numbers of landmarks (b), and parsing priors with different types of parsing maps (c).

dataset [21]. The last 50 images are used for testing and the others are for training. We perform data augmentation on the training images. Specifically, we rotate the original images by $90°$, $180°$, $270°$ and flip them horizontally. This results in 7 additional augmented images for each original one. Besides, each image in Helen dataset has a ground truth label of 194 landmarks and 11 parsing maps.

Effects of Facial Prior Knowledge First, we demonstrate that facial prior knowledge is *significant* for face super-resolution, even without any advanced processing steps. We remove the prior estimation network and construct a single-branch baseline network. Based on the baseline network, we introduce the ground truth facial prior information (i.e., landmark heatmaps and parsing maps) to the "concatenation" layer to construct a new network, as shown in Fig. 3. For fair comparison, we keep the feature map number of "concatenation" layer the same between two networks, which means the results can contrast the effects of the facial prior knowledge. Fig. 4 presents the performance of 3 kinds of settings, including setting with or without parsing maps, landmark heatmaps, or both maps, respectively. As we can see, the models using prior information significantly outperform the corresponding baseline models with the PSNR improvement of 0.4 dB after using landmark heatmaps, 1.0 dB after using parsing maps, and 1.05 dB after using both priors, respectively. These huge improvements on PSNR clearly signify the *positive* effects of facial prior knowledge to face SR.

Upper Bound Improvements from Priors Next, we focus on specific prior information, and study the upper bound improvements that different priors bring. Specifically, for facial landmarks, we introduce 3 sets of landmarks, i.e., 49, 97 and 194 landmarks, respectively. For parsing maps, we introduce the global and local parsing maps, respectively. The global parsing map is shown in Figs. 5(b-c), while Fig. 5(d) shows the local parsing maps containing different facial components. The results of different priors are shown in Fig. 4. We can observe that: (1) Parsing priors contain richer information for face SR and bring much more

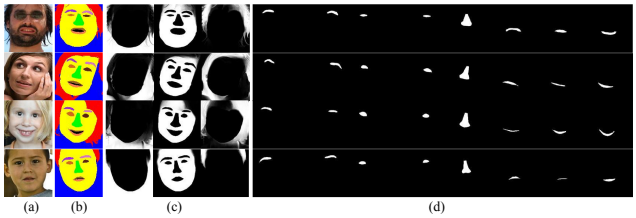

Figure 5: Parsing maps of Helen images. (a) Original image. (b) Color visualization map generated by 11 ground truth parsing maps [23]. It is used as part of the global parsing map. (c) Global parsing maps from the ground truth. (d) Local parsing maps from the ground truth, containing left eyebrow, right eyebrow, left eye, right eye, nose, upper lip, inner mouth, and lower lip, respectively.

Figure 6: Training examples of CelebA (top) and Helen (bottom).

improvements than the landmark prior. (2) Global parsing maps are more useful than local parsing maps. (3) More landmark heatmaps have minor improvements than the version using 49 landmarks.

The above results and analysis demonstrate the effects of both facial priors, and show the upper bound performance that we achieve if the priors are predicted *perfectly*. Since we use the recent popular facial alignment/parsing framework as the prior estimation network, the powerful learning ability enables the network to leverage the priors as much as possible, and hence can benefit the face SR. Apart from the benefit to PSNR, introducing facial prior may bring other advantages, such as more precise recovery of the *face shape*, as reflected by less errors on face alignment and parsing. More details are presented in the next section.

5. Experiments

5.1. Implementation Details

Datasets We conduct extensive experiments on 2 datasets: Helen [21] and celebA [25]. Experimental setting on Helen dataset is described in Sec. 4. For celebA dataset, we use the first $18,000$ images for training, and the following 100 images for evaluation. It should be noted that celebA only has a ground truth of 5 landmarks. We further use a recent alignment model [4] to estimate the 68 landmarks and adopt GFC [23] to estimate the parsing maps as the ground truth.

Training Setting We coarsely crop the training images according to their face regions and resize to 128×128 without any pre-alignment operation. Example training images from celebA and Helen are shown in Fig. 6. For testing, any popular face detector [9] can be used to obtain the cropped

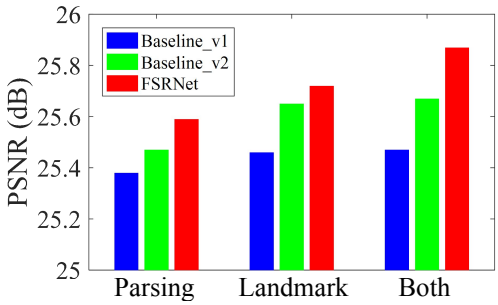

Figure 7: Ablation study on effects of estimated priors.

image as the input. Same as [22], color images are used for training. The input low-resolution images are firstly enlarged by bicubic interpolation, and hence have the same size as the output high-resolution images. For implementation, we train our model with the Torch7 toolbox [5]. The model is trained using the RMSprop algorithm with an initial learning rate of 2.5×10^{-4}, and the mini-batch size of 14. We empirically set $\lambda = 1$, $\gamma_{\mathbf{C}} = 10^{-3}$ and $\gamma_{\mathbf{P}} = 10^{-1}$ for both datasets. Training a basic FSRNet on Helen dataset takes $\sim$6 hours on 1 Titan X GPU.

5.2. Ablation Study

Effects of Estimated Priors We conduct ablation study on the effects of our prior estimation network. Since our SR branch has the similar network structure as SRResNet [22], we clearly show how the performance improves with different kinds of facial priors based on the performance of SRResNet. In this test, we *estimate* the facial priors through the prior estimation network instead of using the ground truth conducted in Sec. 4. Same as the tests conducted in Fig. 4 (a), we conduct 3 experiments to estimate the landmark heatmaps, parsing maps, or both maps, respectively. In each experiment, we further compare our basic FSRNet with two other network structures. Specifically, by removing the prior estimation network from our basic FSRNet, the remaining parts constitute the first network, named 'Baseline_v1', which has the similar structure and hence similar performance as SRResNet. The second network, named 'Baseline_v2', has the same structure as our basic FSRNet except that there is no supervision on the prior estimation network.

Fig. 7 shows the results of different network structures. It can be seen that: (1) The second networks always outperform the first networks. The reason may be even without the supervision, the second branch learns additional features that provide more high-frequency signals to help SR. (2) Compared to the second networks, the supervision on prior knowledge further improves the performance, which indicates the estimated facial priors indeed have positive effects on face super-resolution. (3) The model using both priors achieves the best performance, which indicates richer prior information brings more improvement. (4) The best performance reaches 25.85 dB, which is lower than the performance (i.e., 26.55 dB) when using ground truth. That means our estimated priors are not perfect and a better prior estimation network may result in higher model performance.

CFAN	CFSS	SDM	DeepAlign	FSRNet_S1	FSRNet_S2
9.45	7.26	7.88	6.50	9.44	7.04

Figure 8: Landmark estimations by FSRNet on CelebA. The first row shows the results of the first stacked HG (FSRNet_S1) and the second row is of the second HG (FSRNet_S2). Please zoom in to see the improvements. In the bottom, *NRMSEs of the first four methods are achieved by testing directly on the ground-truth HR images.*

Effects of Hourglass Numbers As discussed in Sec. 4, a powerful prior estimation network may lead to accurate prior estimation. Here, we study the effect of the hourglass number h in the prior estimation network. Specifically, we test $h = 1/2/4$, and the PSNR results are 25.69, 25.87, and 25.95 dB, respectively. Since using more hourglasses leads to a deeper structure, the learning ability of the prior estimation network grows, and hence better performance. To intuitively show the adjustments in stacking more hourglasses, we show the landmark estimations of the first and second stacked hourglass in Fig 8. It can be observed that the estimation is obviously improved in the second stacking.

5.3. Comparisons with State-of-the-Art Methods

We compare FSRNet with state-of-the-art SR methods, including generic SR methods like SRResNet [22], VDSR [17] and SRCNN [6]; and facial SR methods like GLN [35] and URDGN [41]. For fair comparison, we use the released codes of the above models and train all models with the same training set. For URDGN [41], we only train the generator to report PSNR/SSIMs, but the entire GAN network for qualitative comparisons.

Face Super-Resolution First, we compare FSRNet with the state of the arts quantitatively. Tab. 2 summarizes quantitative results on the two datasets. Our FSRNet significantly outperforms state of the arts in both PSNR and SSIM. Not suprisingly, FSRGAN achieves low PSNR/SSIMs. Besides, we also present FSRNet_aug, which sends multiple augmented test images during inference and then fuse the outputs to report the results. This simple yet effective trick brings significant improvements.

Qualitative comparisons of FSRNet/FSRGAN with prior

Dataset	Bicubic	SRCNN	VDSR	SRResNet	GLN	URDGN	FSRNet	FSRNet_aug	FSRGAN
Helen	23.69/0.6592	23.97/0.6779	24.61/0.6980	25.30/0.7297	24.11/0.6922	24.22/0.6909	25.87/0.7602	26.21/0.7720	25.10/0.7234
celebA	23.75/0.6423	24.26/0.6634	24.83/0.6878	25.82/0.7369	24.55/0.6867	24.63/0.6851	26.31/0.7522	26.60/0.7628	25.20/0.7023

Table 2: Benchmark super-resolution, with PSNR/SSIMs for scale factor 8. Red/blue color indicate the best/second best performance.

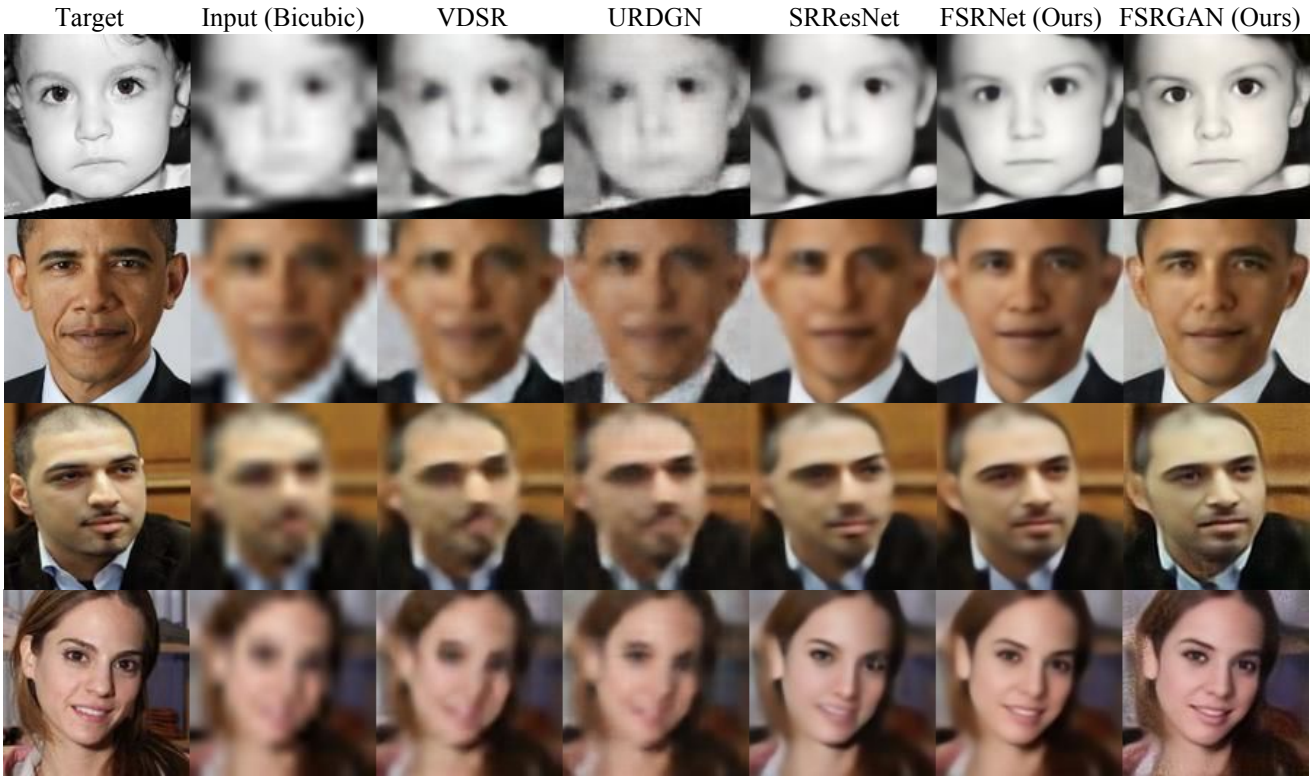

Figure 9: Qualitative comparisons. Top two examples come from Helen and others are from celebA. Please zoom in to see the differences.

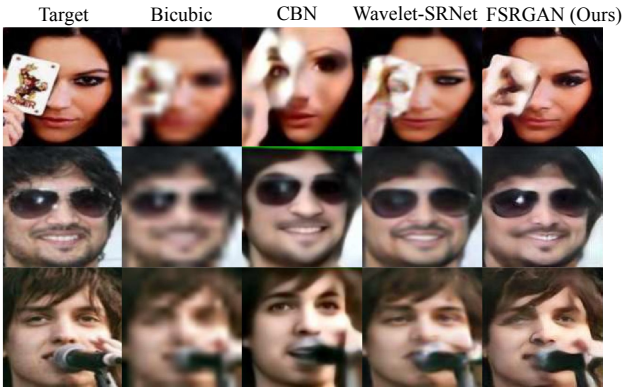

Figure 10: Comparisons with CBN and Wavelet-SRNet.

works are illustrated in Fig. 9. Benefiting from the facial prior knowledge, our method produces relatively sharper edges and shapes, while other methods may give more blurry results. Moreover, FSRGAN further recovers sharper facial textures than FSRNet.

We next compare FSRGAN with two recent face SR methods: Wavelet-SRNet [11] and CBN [45]. We follow the same experimental setting on handling occluded face as [11] and directly import the 16×16 test examples from [11] for super-resolving 128×128 HR images. As shown in Fig. 10, FSRGAN achieves relatively sharper shapes (e.g., nose in all cases) than the state of the arts.

Face Alignment Apart from the evaluation of PSNR/SSIM, we introduce face alignment as a novel evaluation metric for face super-resolution, since accurate face recovery should lead to accurate shape/geometry, and hence accurate landmark points. We adopt a popular alignment model CFAN [43] to estimate the landmarks of different recovered images. Fig. 11 shows the recovered images of SRResNet and our FSRNet, including the results from coarse SR net and final output. Tab. 3 presents the Normalized Root Mean Squared Error (NRMSE) results, which is a popular metric in face alignment and small NRMSE indicates better alignment performance. From the results we can see that: (1) It is difficult for the state-of-the-art alignment model to estimate landmarks directly from very low-resolution images. The estimated landmarks of the bicubic image exhibit large errors around mouth, eyes or other components. In FSRNet, the coarse SR net can ease the alignment difficulty to some extent, which leads to better NRMSE than the input bicubic image. (2) Compared

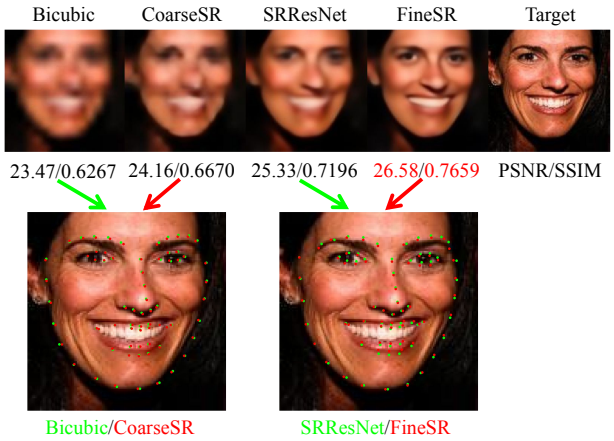

Figure 11: Qualitative comparison of face alignment.

Bicubic	CoarseSR	SRResNet	FineSR	FSRGAN	Target
5.87	5.42	4.87	4.18	3.97	3.32

Table 3: Comparisons of alignment (NRMSE) on Helen dataset.

to SRResNet, our final output provides visually superior estimation on mouth, eyes and shape, and also achieves a large margin of 0.7 quantitatively. That demonstrates the effectiveness of using landmark priors for training.

On the other hand, we also compare the landmarks directly estimated by FSRNet as a by-product, with other methods [19,38,43,44] using their released codes, as shown in the bottom of Fig. 8. It should be noted that *our method starts with the LR images while others are tested directly on the ground-truth* $8\times$ *HR images*. Despite the disadvantage in the input image resolution, our method outperforms most recent methods and is competitive with the state of the art.

Face Parsing We also introduce face parsing as another evaluation metric for face super-resolution. Although our prior estimation network can predict the parsing maps from the LR inputs, for fair comparison, we adopt a recent model GFC [23] to generate the facial parsing maps for the recovered images of all methods, including the bicubic inputs, our coarse SR net, SRResNet, our fine SR net, and targets, respectively. PSNR, SSIM, and Mean Squared Error (MSE) metrics are reported in Tab. 4. As we can see, the coarse SR net also has positive effects on face parsing, and our FSRNet outperforms SRResNet in all of the three evaluations. Fig. 12 presents the estimated parsing maps by [23], the parsing maps from our final HR images recover complete and accurate components, while SRResNet may generate wrong shapes or even lose components (e.g., mouth).

Here, we adopt two side tasks, face alignment and parsing, as the new evaluation metrics for face super resolution. They can subjectively evaluate the quality of geometry in the recovered images, which is compementary to the classic PSNR/SSIM metrics that focus more on photometric quality. Further, Tab. 3 and 4 show that FSRGAN outperforms FSRNet on both metrics, which is consistent

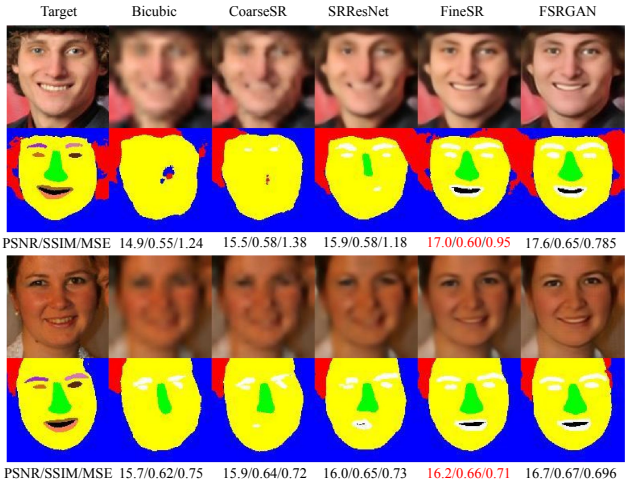

Figure 12: Qualitative comparison of face parsing.

Methods	Bicubic	CoarseSR	SRResNet	FineSR	FSRGAN
PSNR	14.47	14.91	15.32	15.89	16.11
SSIM	0.570	0.585	0.603	0.622	0.629
MSE	1.170	1.100	1.047	0.976	0.934

Table 4: Comparisons of face parsing on Helen dataset.

with the superior visual quality in Fig. 9. This consistency actually addresses one issue in GAN-based super-resolution methods, which has superior visual quality, but lower PSNR/SSIM. This also shows that GAN-based methods can better recover the facial geometry, in addition to perceived visual quality.

Time Complexity Unlike CBN that needs multiple steps and trains multiple models for face hallucination, our FSRNet is faster and more convenient to use, which only needs *one forward process* for inference and costs 0.012s on Titan X GPU, for a 128×128 image. For comparison, CBN has four cascades and totally consumes 3.84s [45], while the traditional face SR requires more time, e.g., [24] needs 8 minutes and [14] needs $15 - 20$ minutes.

6. Conclusions

In this paper, a novel deep end-to-end trainable Face Super-Resolution Network (FSRNet) is proposed for face super-resolution. The key component of FSRNet is the prior estimation network, which not only helps to improve the photometric recovery in terms of PSNR/SSIM, but also provides a solution for accurate geometry estimation directly from very LR images, as shown in the results of facial landmarks/parsing maps. Extensive experimental results show that our FSRNet achieves superior performance than the state of the arts on unaligned face images, both quantitatively and qualitatively. Following the main idea of this work, future research can be expanded in various aspects, including designing a better prior estimation network, e.g., learning the fine SR network iteratively, and investigating other useful facial priors, e.g., texture.

References

[1] S. Baker and T. Kanade. Hallucinating faces. In *FG*, 2000.

[2] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li. Attention-aware face hallucination via deep reinforcement learning. In *CVPR*, 2017.

[3] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and Y. Jian. Adversarial PoseNet: A Structure-aware Convolutional Network for Human Pose Estimation. In *ICCV*, 2017.

[4] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial learning of structure-aware fully convolutional networks for landmark localization. *arXiv:1711.00253*, 2017.

[5] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *NIPS Workshop*, 2011.

[6] C. Dong, C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. on PAMI*, 38(2):295–307, 2016.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[8] A. Graves, A. rahman Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013.

[9] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu. Scale-aware face detection. In *CVPR*, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, 2017.

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[14] Y. Jin and C. Bouganis. Robust multi-image based blind face hallucination. In *CVPR*, 2015.

[15] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[16] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single cnn. In *ICCV*, 2017.

[17] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016.

[18] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *CVPR*, 2015.

[19] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPRW*, 2017.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[21] V. Le, J. Brandt, Z. Lin, L. Boudev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012.

[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, 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.

[23] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *CVPR*, 2017.

[24] C. Liu, H. Shum, and W. Freeman. Face hallucination: Theory and practice. *IJCV*, 2007.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[26] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499, 2016.

[28] N. Pinto, Z. Stone, T. Zickler, and D. Cox. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *CVPRW*, 2011.

[29] J. Salvador and E. Perez-Pellitero. Naive bayes super-resolution forest. In *ICCV*, 2015.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[31] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang. Learning to hallucinate face images via component generation and enhancement. In *IJCAI*, 2017.

[32] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017.

[33] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, 2017.

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[35] O. Tuzel, Y. Taguchi, and J. R. Hershey. Global-local face upsampling network. *arXiv:1603.07235*, 2016.

[36] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015.

[37] X. Wang and X. Tang. Hallucinating face by eigentransformation. *IEEE TSMC, Part C*, 35(3):425–434, 2005.

[38] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. pages 532–539, 2013.

[39] C.-Y. Yang, S. Liu, and M.-H. Yang. Structured face hallucination. In *CVPR*, 2013.

[40] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Trans. on PAMI*, 39(1):156–171, 2017.

[41] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *ECCV*, 2016.

[42] X. Yu and F. Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In *CVPR*, 2017.

[43] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. pages 1–16. Springer, 2014.

[44] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. pages 4998–5006, 2015.

[45] S. Zhu, S. Liu, C. Loy, and X. Tang. Deep cascaded bi-network for face hallucination. In *ECCV*, 2016.

7. Appendix

This supplementary material provides additional details of the following:
(1) The network structure of the discriminator in FSRGAN.
(2) More qualitative comparisons with CBN [45] and several other state-of-the-art face SR methods [14, 29, 39].
(3) More visual examples to show robustness of our FSRNet and FSRGAN on different facial variations.

7.1. Structure of Discriminator in FSRGAN

We follow the previous work [13] to use the "PatchGAN" structure in our discriminator, which down-samples the 128×128 input images to 8×8 feature maps, as shown in Fig. 13. Each pixel in the feature map corresponds to a 16×16 patch in the original image, and the discriminator predicts the identity (fake or real) of each patch in the input.

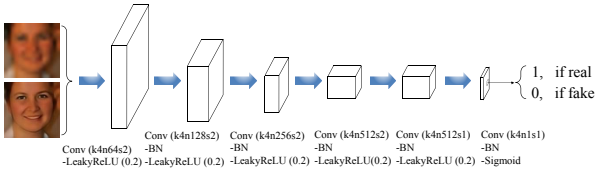

Figure 13: Structure of the discriminator network. The input is the concatenation of the low-resolution image with the recovered high-resolution image (fake) or the ground-truth one (real).

7.2. More Comparisons with State of the Arts

Next, we present more qualitative comparisons with state-of-the-art face SR methods [14, 29, 39, 45]. We follow the same experimental setting as CBN [45], which uses the entire dataset celebA for training and test on dataset PubFig83 [28]. Here, we train our FSRNet/FSRGAN on the scale factor of $4\times$, with the first $201,599$ images or training and the last $1,000$ images for validation. During testing, same as [14, 45], we blur HR faces with $\sigma = 0.4$ to evaluate the robustness of our model on low-resolution and unknown gaussian blur simultaneously. Results are shown in Fig. 15. Compared with CBN, our models have 3 advantages: (1) Our FSRNet looks more similar to the target image than CBN, and FSRGAN recovers competitive results to the target images. (2) There exists border effects in the recovered images of CBN, which is not a problem to our models. (3) CBN needs several steps and models to recover the HR image, which is slow and inconvenient. Our method only needs one forward process during inference, which is fast and convenient.

In Fig. 14, we further present our results on recovering the exact three failure cases shown in CBN [45]. Our models recover sharper and more accurate results than CBN in all three cases. In the first example, CBN exists ghosting effect, while ours are more robust to facial misalignment and pose variations. In the second example, CBN recovers incorrect gaze direction, while ours show the correct direction. In the last example, over-synthesis of eyes is shown in CBN; ours also show this trend but are still better than CBN.

Figure 14: Qualitative results on the exact three representative failure cases of CBN [45]. Please zoom in to see the differences.

7.3. Robustness to Facial Variations

In our paper, we have shown the ability of our models on handling occluded faces. Next, based on the model trained by $201,599$ images from celebA, we present more visual examples from the other $1,000$ validation images on two scale factors, $4\times$ and $8\times$, to show the robustness of our models on more facial variations. Specifically, Fig. 16 shows the robustness of our models to misalignment; Fig. 17 shows the robustness to pose; Fig. 18 shows the robustness to expression; and Fig. 19 shows the robustness to occlusions. The extensive examples demonstrate the robustness of our models to different facial variations. Last but not least, our models can recover extremely good results, which are indeed similar to the ground truth HR images, when handling scale factor to be $4\times$.

Figure 15: Qualitative results on dataset PubFig83. The test samples presented are imported directly from [45].

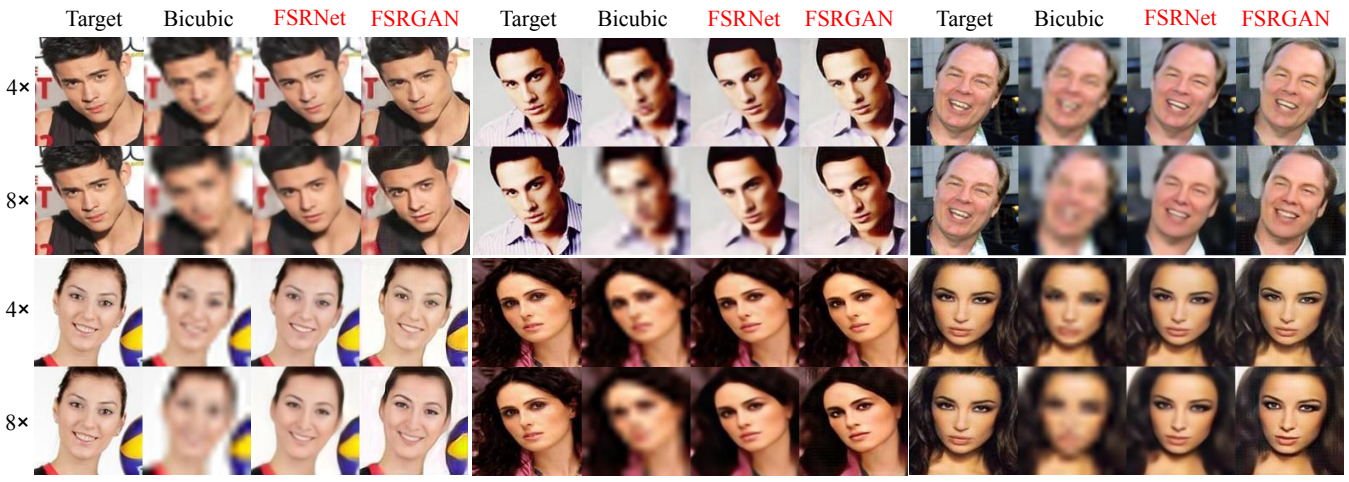

Figure 16: Robustness of FSRNet/FSRGAN to misalignment. Please zoom in for better view.

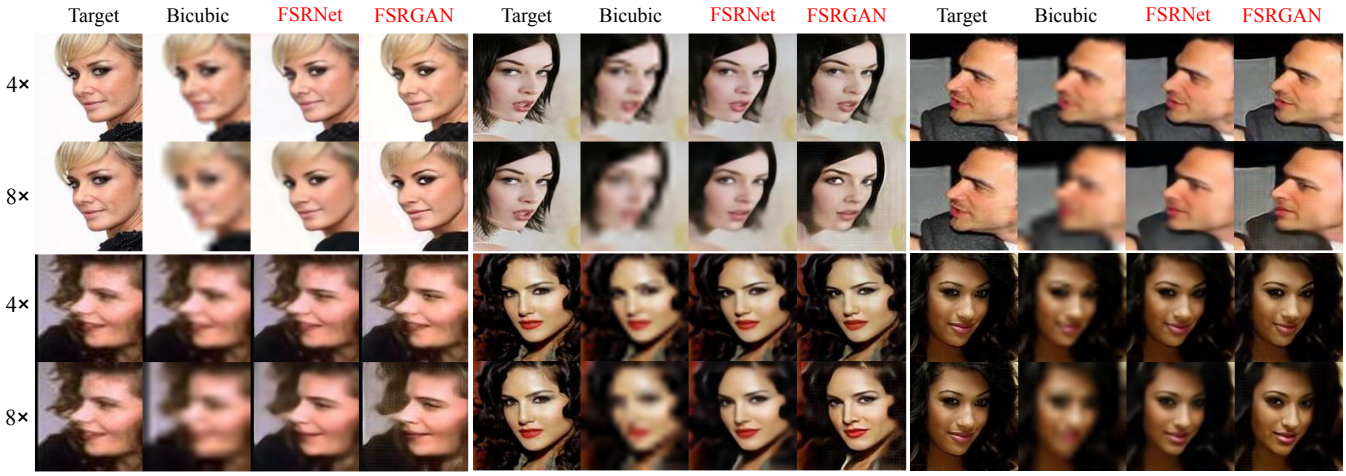

Figure 17: Robustness of FSRNet/FSRGAN to pose. Please zoom in for better view.

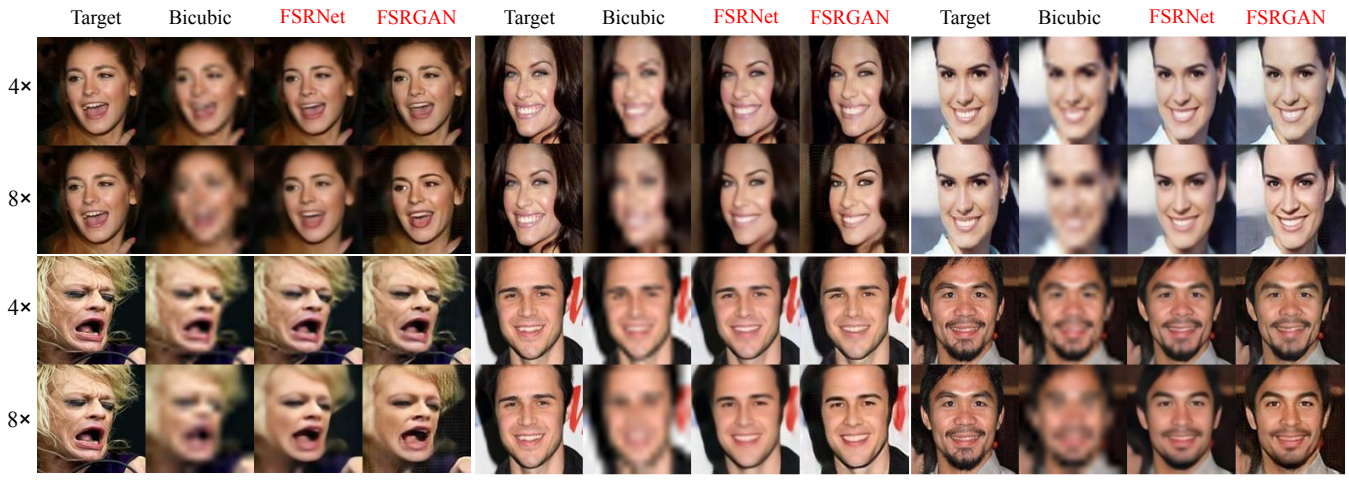

Figure 18: Robustness of FSRNet/FSRGAN to expression. Please zoom in for better view.

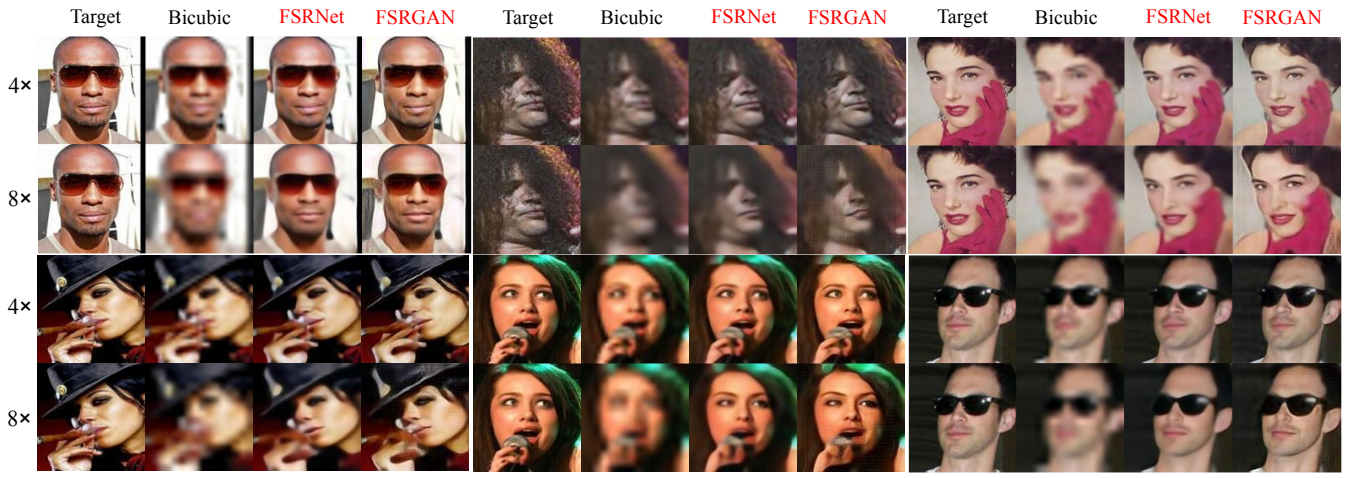

Figure 19: Robustness of FSRNet/FSRGAN to occlusion. Please zoom in for better view.