Face Super-Resolution Guided by 3D Facial Priors

Xiaobin Hu[1,2], Wenqi Ren[2*], John LaMaster[1], Xiaochun Cao[2,6], Xiaoming Li[3], Zechao Li[4], Bjoern Menze[1†], and Wei Liu[5†]

[1] Informatics, Technische Universität München, Germany [2] SKLOIS, IIE, CAS
[3] Harbin Institute of Technology [4] NJUST [5] Tencent AI Lab
[6] Peng Cheng Laboratory, Cyberspace Security Research Center, China

Abstract. State-of-the-art face super-resolution methods employ deep convolutional neural networks to learn a mapping between low- and high-resolution facial patterns by exploring local appearance knowledge. However, most of these methods do not well exploit facial structures and identity information, and struggle to deal with facial images that exhibit large pose variations. In this paper, we propose a novel face super-resolution method that explicitly incorporates 3D facial priors which grasp the sharp facial structures. Our work is the first to explore 3D morphable knowledge based on the fusion of parametric descriptions of face attributes (e.g., identity, facial expression, texture, illumination, and face pose). Furthermore, the priors can easily be incorporated into any network and are extremely efficient in improving the performance and accelerating the convergence speed. Firstly, a 3D face rendering branch is set up to obtain 3D priors of salient facial structures and identity knowledge. Secondly, the Spatial Attention Module is used to better exploit this hierarchical information (i.e., intensity similarity, 3D facial structure, and identity content) for the super-resolution problem. Extensive experiments demonstrate that the proposed 3D priors achieve superior face super-resolution results over the state-of-the-arts.

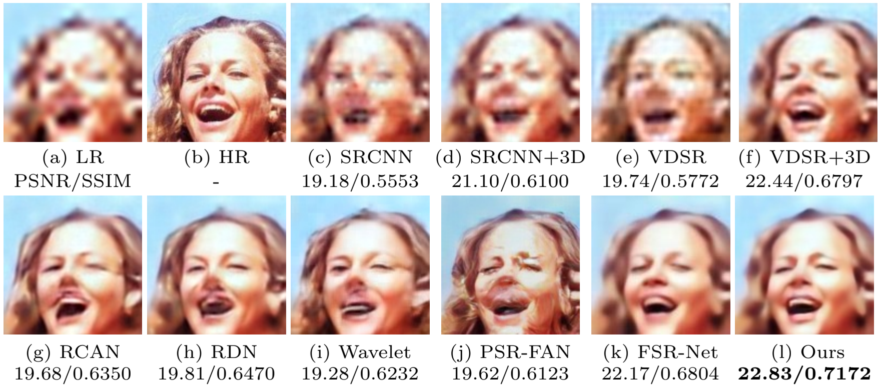
Keywords: face super-resolution, 3D facial priors, facial structures and identity knowledge.

1 Introduction

Face images provide crucial clues for human observation as well as computer analysis [20,45]. However, the performance of most face image tasks, such as face recognition and facial emotion detection [11,32], degrades dramatically when the resolution of a facial image is relatively low. Consequently, face super-resolution, also known as face hallucination, was coined to restore a high-resolution face image from its low-resolution counterpart.

$*$ indicates the corresponding author.

† these authors contributed equally to this work.



(a) LR PSNR/SSIM
(b) HR -
(c) SRCNN 19.18/0.5553
(d) SRCNN+3D 21.10/0.6100
(e) VDSR 19.74/0.5772
(f) VDSR+3D 22.44/0.6797
(g) RCAN 19.68/0.6350
(h) RDN 19.81/0.6470
(i) Wavelet 19.28/0.6232
(j) PSR-FAN 19.62/0.6123
(k) FSR-Net 22.17/0.6804
(l) Ours **22.83/0.7172**

Fig. 1. Visual comparison with state-of-the-art face hallucination methods (×8). (a) 16×16 LR input. (b) 128×128 HR ground-truth. (c) Super-Resolution Convolutional Neural Network (SRCNN) [7]. (d) SRCNN incorporating our 3D facial priors. (e) Very Deep Super-Resolution Network (VDSR) [17]. (f) VDSR incorporating our 3D facial priors. (g) Very Deep Residual Channel Attention Network (RCAN) [42]. (h) Residual Dense Network (RDN) [43]. (i) Wavelet-based CNN for Multi-scale Face Super-Resolution (Wavelet-SRNet) [14]. (j) Progressive Face Super-Resolution using the facial landmark (PSR-FAN) [16]. (k) End-to-End Learning Face Super-Resolution with Facial Priors (FSRNet) [4]. (l) Our proposed method by embedding the 3D facial priors into the Spatial Attention Module (SAM3D).

Although a great influx of deep learning methods [3,5,9,24,36–39,44,46,47] have been successfully applied in face Super-Resolution (SR) problems, super-resolving arbitrary facial images, especially at high magnification factors, is still an open and challenging problem due to the ill-posed nature of the SR problem and the difficulty in learning and integrating strong priors into a face hallucination model. Some researches [4,10,16,28,35,41] on exploiting the face priors to assist neural networks in capturing more facial details have been proposed. A face hallucination model incorporating identity priors was presented in [10]. However, the identity prior was extracted only from the multi-scale up-sampling results in the training procedure and therefore cannot provide extra priors to guide the network. Yu et al. [35] employed facial component heatmaps to encourage the upsampling stream to generate super-resolved faces with higher-quality details, especially for large pose variations. Kim et al. [16] proposed a face alignment network (FAN) for landmark heatmap extraction to boost the performance of face SR. Chen et al. [4] utilized the heatmaps and parsing maps for face SR problems. Although these 2D priors provide global component regions, these methods cannot learn the 3D reconstruction of detailed edges, illumination, and expression priors. In addition, all of these aforementioned face SR approaches ignore facial structure and identity recovery.

In contrast to the aforementioned approaches, we propose a novel face super resolution method by exploiting 3D facial priors to grasp sharp face structures and identity knowledge. Firstly, a deep 3D face reconstruction branch is set up to explicitly obtain 3D face render priors which facilitate the face super-resolution branch. Specifically, the 3D facial priors contain rich hierarchical features, such as low-level (e.g., sharp edge and illumination) and perception level (e.g., identity) information. Then, a spatial attention module is employed to adaptively integrate the 3D facial prior into the network, in which we employ a spatial feature transform (SFT) [34] to generate affine transformation parameters for spatial feature modulation. Afterwards, it encourages the network to learn the spatial inter-dependencies of features between 3D facial priors and input images after adding the attention module into the network. As shown in Figure 1, by embedding the 3D rendered face priors, our algorithm generates clearer and sharper facial structures without any ghosting artifacts compared with other 2D prior-based methods.

The main contributions of this paper are:

- A novel face SR model is proposed by explicitly exploiting facial structure in the form of facial prior estimation. The estimated 3D facial prior provides not only spatial information of facial components but also their 3D visibility information, which is ignored by the pixel-level content and 2D priors (e.g., landmark heatmaps and parsing maps).
- To well adapt to the 3D reconstruction of low-resolution face images, we present a new skin-aware loss function projecting the constructed 3D coefficients onto the rendered images. In addition, we use a feature fusion-based network to better extract and integrate the face rendered priors by employing a spatial attention module.
- Our proposed 3D facial prior has a high flexibility because its modular structure allows for easy plug-in of any SR methods (e.g., SRCNN and VDSR). We qualitatively and quantitatively evaluate the proposed algorithm on multi-scale face super-resolution, especially at very low input resolutions. The proposed network achieves better SR criteria and superior visual quality compared to state-of-the-art face SR methods.

2 Related Work

Face hallucination relates closely to the natural image super-resolution problem. In this section, we discuss recent research on super-resolution and face hallucination to illustrate the necessary context for our work.

Super-Resolution Neural Networks. Recently, neural networks have demonstrated a remarkable capability to improve SR results. Since the pioneering network [7] demonstrates the effectiveness of CNN to learn the mapping between LR and HR pairs, a lot of CNN architectures have been proposed for SR [8,12,18,19,30,31]. Most of the existing high-performance SR networks have residual blocks [17] to go deeper in the network architecture, and achieve better performance. EDSR [22] improved the performance by removing unnecessary batch normalization layers in residual blocks. A residual dense network (RDN) [43] was proposed to exploit the hierarchical features from all the convolutional layers. Zhang et al. [42] proposed the very deep residual channel attention networks (RCAN) to discard abundant low-frequency information which hinders the representational ability of CNNs. Wang et al. [34] used a spatial feature transform layer to introduce the semantic prior as an additional input of the SR network. Huang et al. [14] presented a wavelet-based CNN approach that can ultra-resolve a very low-resolution face image in a unified framework. Lian et al. [21] proposed a Feature-Guided Super-Resolution Generative Adversarial Network (FG-SRGAN) for unpaired image super-resolution. However, these networks require a lot of time to train the massive parameters to obtain good results. In our work, we largely decrease the training parameters, but still achieve superior performance in the SR criteria (SSIM and PSNR) and visible quality.

Facial Prior Knowledge. Exploiting facial priors in face hallucination, such as spatial configuration of facial components [29], is the key factor that differentiates it from generic super-resolution tasks. There are some face SR methods that use facial prior knowledge to super-resolve LR faces. Wang and Tang [33] learned subspaces from LR and HR face images, and then reconstructed an HR output from the PCA coefficients of the LR input. Liu et al. [23] set up a Markov Random Field (MRF) to reduce ghosting artifacts because of the misalignments in LR images. However, these methods are prone to generating severe artifacts, especially with large pose variations and misalignments in LR images. Yu and Porikli [38] interweaved multiple spatial transformer networks [15] with the deconvolutional layers to handle unaligned LR faces. Dahl et al. [5] leveraged the framework of PixelCNN [26] to super-resolve very low-resolution faces. Zhu et al. [47] presented a cascade bi-network, dubbed CBN, to localize LR facial components first and then upsample the facial components; however, CBN may produce ghosting faces when localization errors occur. Recently, Yu et al. [35] used a multi-task convolutional neural network (CNN) to incorporate structural information of faces. Grm et al. [10] built a face recognition model that acts as identity priors for the super-resolution network during training. Yu et al. [4] constructed an end-to-end SR network to incorporate the facial landmark heatmaps and parsing maps. Kim et al. [16] proposed a compressed version of the face alignment network (FAN) to obtain landmark heatmaps for the SR network in a progressive method. However, existing face SR algorithms only employ 2D priors without considering high-dimensional information (3D). In this paper, we exploit the 3D face reconstruction branch to extract the 3D facial structure, detailed edges, illumination, and identity priors to guide face image super-resolution.

3D Face Reconstruction. The 3D shapes of facial images can be restored from unconstrained 2D images by the 3D face reconstruction. In this paper, we employ the 3D Morphable Model (3DMM) [1,2,6] based on the fusion of parametric descriptions of face attributes (e.g., gender, identity, and distinctiveness) to reconstruct the 3D facial priors. The 3D reconstructed face will inherit the facial features and present the clear and sharp facial components.

Closest to ours is the work of Ren et al. [28] which utilizes the 3D priors in the task of face video deblurring. Our method differs in several important ways. First, instead of simple priors concatenation, we employ the Spatial Feature

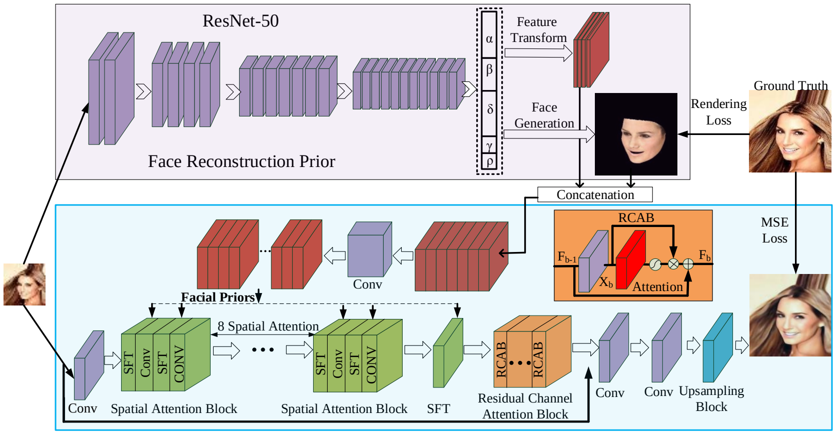


Fig. 2. The proposed face super-resolution architecture. Our model consists of two branches: the top block is a ResNet-50 Network to extract the 3D facial coefficients and restore a sharp face rendered structure. The bottom block is dedicated to face super-resolution guided by the facial coefficients and rendered sharp face structures which are concatenated by the Spatial Feature Transform (SFT) layer.

Transform Block to incorporate the 3D priors in the intermediate layer by adaptively adjusting the modulation parameter pair. Specifically, the outputs of the SFT layer are adaptively controlled by the modulation parameter pair by applying an affine transformation spatially to each intermediate feature map. Second, the attention mechanism is embeded into the network as a guide to bias the allocation of most informative components and the interdependency between the 3D priors and input.

3 The Proposed Method

The proposed face super-resolution framework presented in Figure 2 consists of two branches: the 3D rendering network to extract the facial prior and the spatial attention module aiming to exploit the prior for the face super-resolution problem. Given a low-resolution face image, we first use the 3D rendering branch to extract the 3D face coefficients. Then a high-resolution rendered image is generated using the 3D coefficients and regarded as the high-resolution facial prior which facilitates the face super-resolving process in the spatial attention module.

3.1 Motivations and Advantages of 3D Facial Priors

Existing face SR algorithms only employ 2D priors without considering high dimensional information (3D). The 3D morphable facial priors are the main novelty of this work and are completely different from recently related 2D prior works (*e.g.,* the parsing maps and facial landmark heatmaps by FSRNet [4] and the landmark heatmap extraction by FAN [16]). The 3D coefficients contain

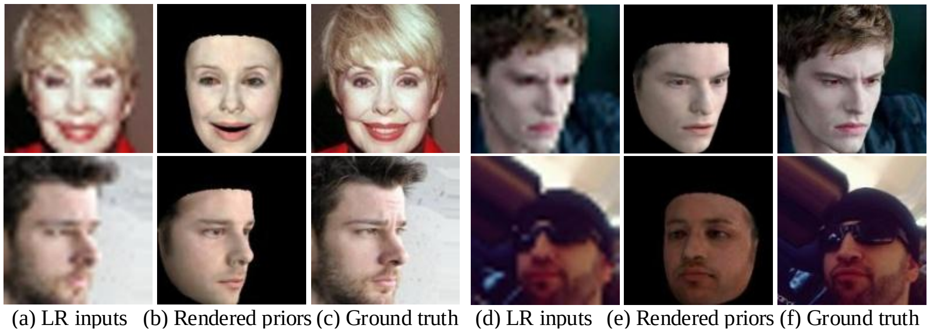


Fig. 3. The rendered priors from our method. (a) and (d) low-resolution inputs. (b) and (e) our rendered face structures. (c) and (f) ground-truths. As shown, the reconstructed facial structures provide clear spatial locations and sharp visualization of facial components even for large pose variations (e.g., left and right facial pose positions) and partial occlusions.

abundant hierarchical knowledge, such as identity, facial expression, texture, illumination, and face pose. Furthermore, in contrast with the 2D landmark-based priors whose attentions only lie at the distinct points of facial landmarks that may lead to the facial distortions and artifacts, our 3D priors are explicit and visible, and can generate the realistic and robust HR results, greatly reducing artifacts even for large pose variations and partial occlusions.

Given low-resolution face images, the generated 3D rendered reconstructions are shown in Figure 3. The rendered face predictions contain the clear spatial knowledge and sharp visual quality of facial components which are close to the ground-truth, even in images containing large pose variations as shown in the second row of Figure 3. Therefore, we concatenate the reconstructed face image as an additional feature in the super-resolution network. The face expression, identity, texture, the element-concatenation of illumination, and face pose are transformed into four feature maps and fed into the spatial feature transform block of the super-resolution network.

For real-world applications of the 3D face morphable model, there are typical problems to overcome, including large pose variations and partial occlusions. As shown in the supplementary material, the morphable model can generate realistic reconstructions of large pose variations, which contain faithful visual quality of facial components. The 3D model is also robust and accurately restores the rendered faces partially occluded by glasses, hair, etc. In comparison with other SR algorithms which are blind to unknown degradation types, our 3D model can robustly generate the 3D morphable priors to guide the SR branch to grasp the clear spatial knowledge and facial components, even for complicated real-world applications. Furthermore, our 3D priors can be plugged into any network and largely improve the performance of existing SR networks (*e.g.,* SRCNN and VDSR demonstrated in Section 5).

3.2 Formulation of 3D Facial Priors

It is still a challenge for state-of-the-art edge prediction methods to acquire very sharp facial structures from low-resolution images. Therefore, a 3DMM-based model is proposed to localize the precise facial structure by generating the 3D facial images which are constructed by the 3D coefficient vector. In addition, there exist large face pose variations, such as in-plane and out-of-plane rotations. A large amount of data is needed to learn the representative features varying with the facial poses. To address this problem, an inspiration came from the idea that the 3DMM coefficients can analytically model the pose variations with a simple mathematical derivation [2,6] and do not require a large training set. As such, we utilize a face rendering network based on ResNet-50 to regress a face coefficient vector. The output of the ResNet-50 is the representative feature vector of $\boldsymbol{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\rho}) \in \mathbb{R}^{239}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}, \boldsymbol{\beta} \in \mathbb{R}^{64}, \boldsymbol{\delta} \in \mathbb{R}^{80}, \boldsymbol{\gamma} \in \mathbb{R}^{9}$, and $\boldsymbol{\rho} \in \mathbb{R}^{6}$ represent the identity, facial expression, texture, illumination, and face pose [6], respectively.

According to the Morphable model [1], we transform the face coefficients to a 3D shape $\mathbf{S}$ and texture $\mathbf{T}$ of the face image as

$$\mathbf{S} = \mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \overline{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \tag{1}$$

and

$$\mathbf{T} = \mathbf{T}(\boldsymbol{\delta}) = \overline{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta}, \tag{2}$$

where $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ are the average values of face shape and texture, respectively. $\mathbf{B}_t$, $\mathbf{B}_{id}$, and $\mathbf{B}_{exp}$ denote the base vectors of texture, identity, and expression calculated by the PCA method. We set up the illumination model by assuming a Lambertian surface for faces, and estimate the scene illumination with Spherical Harmonics (SH) [27] to derive the illumination coefficient $\boldsymbol{\gamma} \in \mathbb{R}^9$. The 3D face pose $\boldsymbol{\rho} \in \mathbb{R}^6$ is represented by rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$.

To stabilize the rendered faces, a modified L_2 loss function for the 3D face reconstruction is presented based on a paired training set

$$\ell_r = \frac{1}{L} \sum_{j=1}^{L} \frac{\sum_{i \in M} A^i \left\| I_j^i - R_j^i(B(\boldsymbol{x})) \right\|_2}{\sum_{i \in M} A^i}, \tag{3}$$

where j is the paired image index, L is the total number of training pairs, i and M denote the pixel index and face region, respectively, I represents the sharp image, and A is a skin color based attention mask obtained by training a Bayes classifier with Gaussian Mixture Models [6]. In addition, x represents the LR (input) images, $B(x)$ denotes the regressed coefficients obtained by the ResNet-50 with input x as input, and finally R denotes the image rendered with the 3D coefficients $B(x)$. Rendering is the process to project the constructed 3D face onto the 2D image plane with the regressed pose and illumination. We use a ResNet-50 network to regress these coefficients by modifying the last fully-connected layer to 239 neurons (the same number of the coefficient parameters).

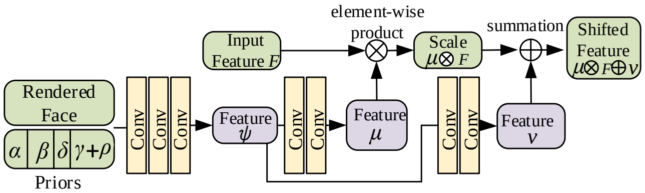


Fig. 4. The structure of the SFT layer. The rendered faces and feature vectors are regarded as the guidance for face super-resolution.

Coefficient Feature Transformation. Our 3D face priors consist of two parts: one directly from the rendered face region (*i.e.*, the RGB input), and the other from the feature transformation of the coefficient parameters. The coefficient parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\rho}$ represent the identity, facial expression, texture, illumination, and face pose priors, respectively. The coefficient feature transformation procedure is described as follows: firstly, the coefficients of identity, expression, texture, and the element-concatenation of illumination and face pose $(\boldsymbol{\gamma}+\boldsymbol{\rho})$ are reshaped to four matrices by setting extra elements to zeros. Afterwards, these four matrices are expanded to the same size as the LR images (16×16 or 32×32) by zero-padding, and then scaled to the interval [0,1]. Finally, the coefficient features are concatenated with the priors of the rendered face images.

3.3 Spatial Attention Module

To exploit the 3D face rendered priors, we propose a Spatial Attention Module (SAM) to grasp the precise locations of face components and the facial identity. The proposed SAM consists of three parts: a spatial feature transform block, a residual channel attention block, and an upscale block.

Spatial Feature Transform Block. The 3D face priors (rendered faces and coefficient features) are imported into the spatial attention transform block [34] after a convolutional layer. The structure of the spatial feature transform layer is shown in Figure 4. The SFT layer learns a mapping function Θ that provides a modulation parameter pair (μ, ν) according to the priors ψ, such as segmentation probability. Here, the 3D face priors are taken as the input. The outputs of the SFT layer are adaptively controlled by the modulation parameter pair by applying an affine transformation spatially to each intermediate feature map. Specifically, the intermediate transformation parameters (μ, ν) are derived from the priors ψ by the mapping function:

$$(\mu, \nu) = \Theta(\psi), \tag{4}$$

The intermediate feature maps are modified by scaling and shifting feature maps according to the transformation parameters:

$$\boldsymbol{SFT}(\boldsymbol{F}|\boldsymbol{\mu}, \boldsymbol{\nu}) = \boldsymbol{\mu} \otimes \boldsymbol{F} + \boldsymbol{\nu}, \tag{5}$$

where $\boldsymbol{F}$ denotes the feature maps, and $\otimes$ indicates element-wise multiplication. At this step, the SFT layer implements the spatial-wise transformation.

Residual Channel Attention Block. An attention mechanism can be viewed as a guide to bias the allocation of available processing resources towards the most informative components of the input [13]. Consequently, the channel mechanism is presented to explore the most informative components and the inter-dependency between the channels. Inspired by the residual channel network [42], the attention mechanism is composed of a series of residual channel attention blocks (RCAB) shown in Figure 2. For the b-th block, the output $\boldsymbol{F_b}$ of RCAB is obtained by:

$$\boldsymbol{F_b} = \boldsymbol{F_{b-1}} + C_b(\boldsymbol{X_b}) \cdot \boldsymbol{X_b}, \tag{6}$$

where C_b denotes the channel attention function. $\boldsymbol{F_{b-1}}$ is the block's input, and $\boldsymbol{X_b}$ is calculated by two stacked convolutional layers. The upscale block is progressive deconvolutional layers (also known as transposed convolution).

4 Experimental Results

To evaluate the performances of the proposed face super-resolution network, we qualitatively and quantitatively compare our algorithm against nine start-of-the-art super-resolution and face hallucination methods including: the Very Deep Super Resolution Network (VDSR) [17], the Very Deep Residual Channel Attention Network (RCAN) [42], the Residual Dense Network (RDN) [43], the Super-Resolution Convolutional Neural Network (SRCNN) [7], the Transformative Discriminative Autoencoder (TDAE) [38], the Wavelet-based CNN for Multi-scale Face Super Resolution (Wavelet-SRNet) [14], the deep end-to-end trainable face SR network (FSRNet) [4], face SR generative adversarial network (FSRGAN) [4] incorporating the 2D facial landmark heatmaps and parsing maps, and the progressive face Super Resolution network via face alignment network (PSR-FAN) [16] using 2D landmark heatmap priors. We use the open-source implementations from the authors and train all the networks on the same dataset for a fair comparison. For simplicity, we refer to the proposed network as Spatial Attention Module guided by 3D priors, or SAM3D. In addition, to demonstrate the plug-in characteristic of the proposed 3D facial priors, we propose two models of SRCNN+3D and VDSR+3D by embedding the 3D facial prior as an extra input channel to the basic backbone of SRCNN [7] and VDSR [17]. The implementation code will be made available to the public. More analyses and results can be found in the supplementary material.

4.1 Datasets and Implementation Details

CelebA [25] and Menpo [40] datasets are used to verify the performance of the algorithm. The training phase uses 162,080 images from the CelebA dataset. In the testing phase, 40,519 images from the CelebA test set are used along with the large-pose-variation test set from the Menpo dataset. The every facial pose test set of Menpo (left, right and semi-frontal) contains 1000 images, respectively. We follow the protocols of existing face SR methods (e.g., [16], [4], [35], [36]) to

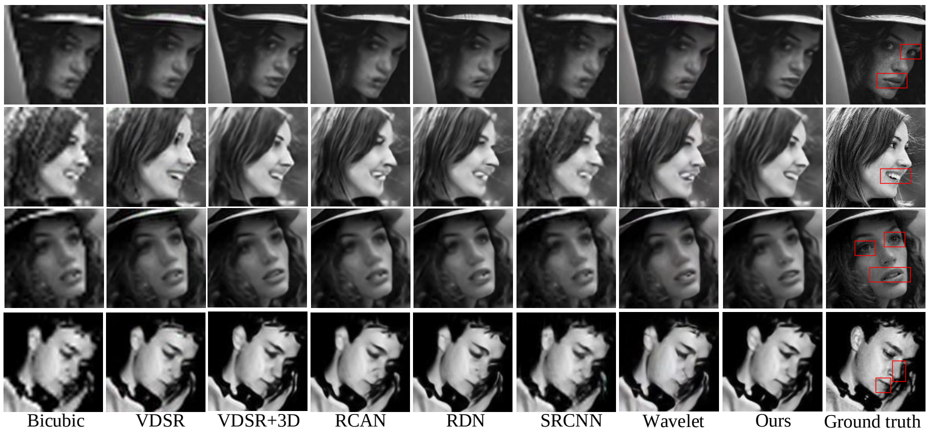


Fig. 5. Comparison of state-of-the-art methods: magnification factors ×4 and the input resolution 32×32. Our algorithm is able to exploit the regularity present in face regions rather than other methods. Best viewed by zooming in on the screen.

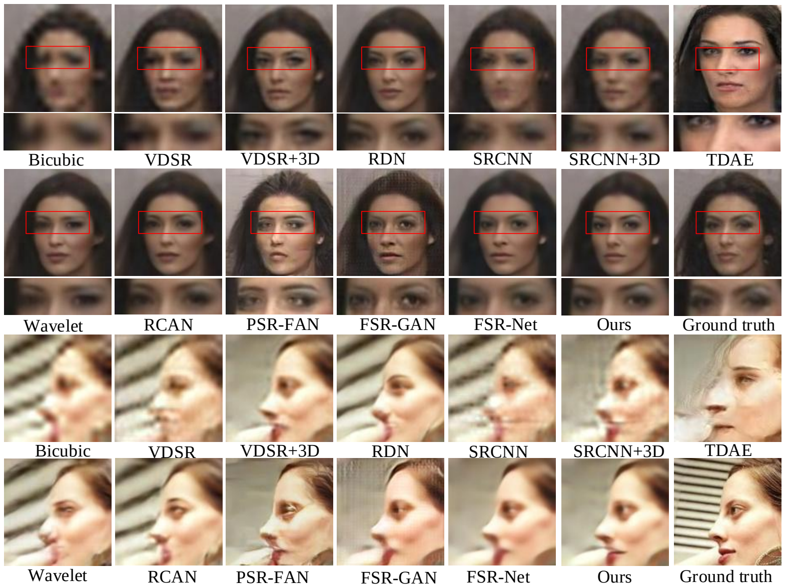


Fig. 6. Comparison with state-of-the-art methods: magnification factors ×8 and the input resolution 16×16. Best viewed by zooming in on the screen.

generate the LR input by the bicubic downsampling method. The HR ground-truth images are obtained by center-cropping the facial images and then resizing them to the 128×128 pixels. The LR face images are generated by downsampling HR ground-truths to 32×32 pixels (×4 scale) and 16×16 pixels (×8 scale). In our network, the ADAM optimizer is used with a batch size of 64 for training, and input images are center-cropped as RGB channels. The initial learning rate

Table 1. Quantitative results on the CelebA test dataset. The best results are highlighted in bold.

-	CelebA			
Scale	×4		×8	
	PSNR	SSIM	PSNR	SSIM
Bicubic	27.16	0.8197	21.90	0.6213
VDSR [17]	28.13	0.8554	22.76	0.6618
RCAN [42]	29.04	0.8643	23.26	0.7362
RDN [43]	29.06	0.8650	23.69	0.7484
SRCNN [7]	27.57	0.8452	22.51	0.6659
TDAE [38]	-	-	20.10	0.5802
Wavelet-SRNet [14]	28.42	0.8698	23.08	0.7147
FSRGAN [4]	-	-	22.27	0.6010
FSRNet [4]	-	-	22.62	0.6410
PSR-FAN [16]	-	-	22.66	0.6850
VDSR+3D	29.29	0.8727	24.66	0.7127
Ours	**29.69**	**0.8817**	**25.39**	**0.7551**

Table 2. Quantitative results of different large facial pose variations (e.g., left, right, and semifrontal) on the Menpo test dataset. The best results are highlighted in bold.

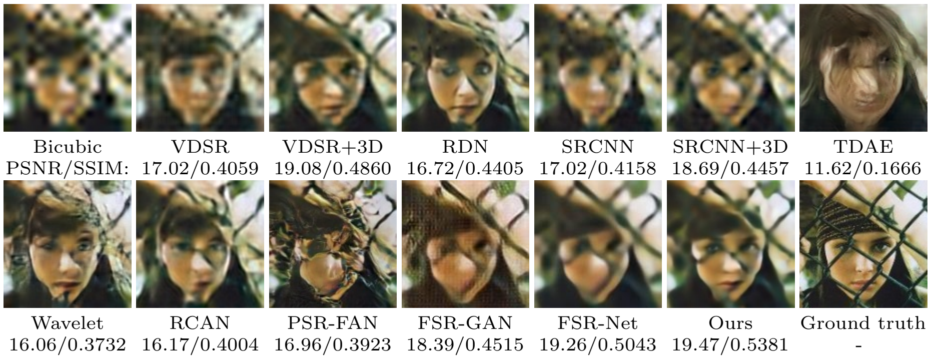
-	Menpo											
Scale	×4						×8					
Pose	Left		Right		Semi-frontal		Left		Right		Semi-frontal	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	26.36	0.7923	26.19	0.7791	24.92	0.7608	22.09	0.6423	21.99	0.6251	20.68	0.5770
VDSR [17]	26.99	0.8024	26.85	0.7908	25.63	0.7794	22.28	0.6315	22.20	0.6163	20.98	0.5752
RCAN [42]	27.47	0.8259	27.27	0.8145	26.11	0.8080	21.94	0.6543	21.87	0.6381	20.60	0.5938
RDN [43]	27.39	0.8263	27.21	0.8150	26.06	0.8088	22.30	0.6706	22.24	0.6552	21.02	0.6160
SRCNN [7]	26.92	0.8038	26.74	0.7913	25.50	0.7782	22.38	0.6408	22.32	0.6272	21.08	0.5857
TDAE [38]	-	-	-	-	-	-	21.22	0.5678	20.22	0.5620	19.88	0.5521
Wavelet-SRNet [14]	26.97	0.8122	26.81	0.8001	25.72	0.7945	21.86	0.6360	21.72	0.6166	20.57	0.5779
FSRGAN [4]	-	-	-	-	-	-	23.00	0.6326	22.84	0.6173	22.00	0.5938
FSRNet [4]	-	-	-	-	-	-	23.56	0.6896	23.43	0.6712	22.03	0.6382
PSR-FAN [16]	-	-	-	-	-	-	22.04	0.6239	21.89	0.6114	20.88	0.5711
VDSR+3D	28.62	0.8439	28.89	0.8326	26.99	0.8236	23.45	0.6845	23.25	0.6653	21.83	0.6239
Ours	**28.98**	**0.8510**	**29.29**	**0.8408**	**27.29**	**0.8332**	**23.80**	**0.7071**	**23.57**	**0.6881**	**22.15**	**0.6501**

is 0.0002 and is divided by 2 every 50 epochs. The whole training process takes 2 days with an NVIDIA Titan X GPU.

4.2 Quantitative Results

Quantitative evaluation of the network using PSNR and the structural similarity (SSIM) scores for the CelebA test set is listed in Table 1. Furthermore, to analyze the performance and stability of the proposed method with respect to large face pose variations, three cases corresponding to different face poses (left, right, and semifrontal) of the Menpo test data are listed in Table 2.

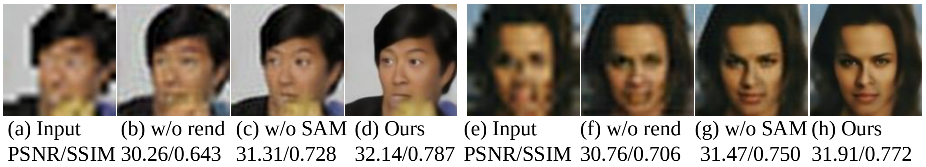
CelebA Test: As shown in Table 1, VDSR+3D (the basic VDSR model [17] guided by the proposed 3D facial priors) achieves significantly better results (1 dB higher than the remaining best method and 2 dB higher than the basic VDSR method in ×8 SR) even for the large-scale parameter methods, such as RDN and



Bicubic PSNR/SSIM: | VDSR 17.02/0.4059 | VDSR+3D 19.08/0.4860 | RDN 16.72/0.4405 | SRCNN 17.02/0.4158 | SRCNN+3D 18.69/0.4457 | TDAE 11.62/0.1666

Wavelet 16.06/0.3732 | RCAN 16.17/0.4004 | PSR-FAN 16.96/0.3923 | FSR-GAN 18.39/0.4515 | FSR-Net 19.26/0.5043 | Ours 19.47/0.5381 | Ground truth -

Fig. 7. Visual comparison with state-of-the-art methods (×8). The results by the proposed method have fewer artifacts on face components (e.g., eyes, mouth, and nose).



(a) Input PSNR/SSIM | (b) w/o rend 30.26/0.643 | (c) w/o SAM 31.31/0.728 | (d) Ours 32.14/0.787 | (e) Input PSNR/SSIM | (f) w/o rend 30.76/0.706 | (g) w/o SAM 31.47/0.750 | (h) Ours 31.91/0.772

Fig. 8. Ablation study results: Comparisons between our proposed model with different configurations, with PSNR and SSIM relative to the ground truth. (a) and (e) are the inputs. (b) and (f) are the SR results without using the rendered priors. (c) and (g) are the SR results without the Spatial Attention Module. (d) and (h) are our SR results.

RCAN. It is worth noting that VDSR+3D still performs slightly worse than the proposed algorithm of SAM3D. These results demonstrate that the proposed 3D priors make a significant contribution to the performance improvement (average 1.6 dB improvement) of face super-resolution. In comparison with 2D priors based methods (*e.g.,* FSRNet and PSR-FAN), our algorithm performs much better (2.73 dB higher than PSR-FAN and 2.78 dB higher than FSRNet).

Menpo Test: To verify the effectiveness and stability of the proposed network towards face pose variations, the quantitative results on the dataset with large pose variations are reported in Table 2. While ours (SAM3D) is the best method superior than the others, VDSR+3D also achieves 1.8 dB improvement compared with the basic VDSR method in the ×4 magnification factor. Our 3D facial priors based method is still the most effective approach to boost the SR performance compared with 2D heatmaps and parsing maps priors.

4.3 Qualitative Evaluation

The qualitative results of our methods at different magnifications (×4 and ×8) are shown respectively in Figures 5 and 6. It can be observed that our proposed method recovers clearer faces with finer component details (e.g., noses, eyes,

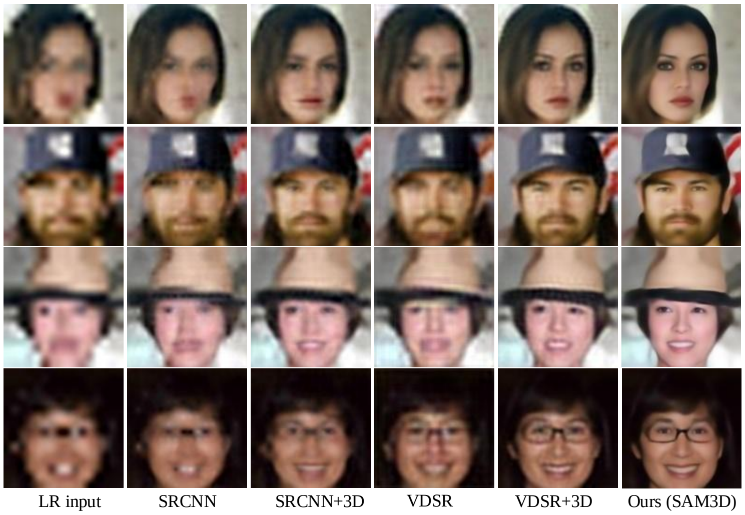


Fig. 9. Qualitative evaluation with different ablation configurations: SRCNN+3D and VDSR+3D denote the basic method (SRCNN and VDSR) incorporating the 3D facial priors; Ours (SAM3D) means the Spatial Attention Module incorporating the 3D facial priors. Our 3D priors enable the basic methods to avoid some artifacts around the key facial components and to generate sharper edges.

and mouths). The outputs of most methods (*e.g.,* PSR-FAN, RCAN, RDN, and Wavelet-SRNet) contain some artifacts around facial components such as eyes and nose, as shown in Figures 1 and 7, especially when facial images are partially occluded. After adding the rendered face priors, our results show clearer and sharper facial structures without any ghosting artifacts, which illustrates that the proposed 3D priors help the network understand the spatial location and the entire face structure and largely avoid the artifacts and significant distortions in facial attributes which are common in facial landmark priors, because the attention is applied merely to the distinct points of facial landmarks.

5 Analyses and Discussions

Ablation Study: In this section, we conduct an ablation study to demonstrate the effectiveness of each module. We compare the proposed network with and without using the rendered 3D face priors and the Spatial Attention Module (SAM) in terms of PSNR and SSIM on the ×8 scale test data. As shown in Figure 8 (b) and (f), the baseline method without using the rendered faces and SAM tends to generate blurry faces that cannot capture sharp structures. Figure 8 (c) and (g) show clearer and sharper facial structures after adding the 3D rendered priors. By using both SAM and 3D priors, the visual quality is further improved in Figure 8 (d) and (h). The quantitative comparisons between (VDSR, our VDSR+3D, and our SAM3D) in Tables 1 and 2 also illustrate the effectiveness of the proposed rendered priors and the spatial attention module.

To verify the advantage of 3D facial structure priors in terms of the convergence and accuracy, three different configurations are designed: basic methods (*i.e.,* SRCNN [7] and VDSR [17]); basic methods incorporating 3D facial priors (*i.e.,* SRCNN+3D and VDSR+3D); the proposed method using the Spatial Attention Module and 3D priors (SAM3D). The validation accuracy curve of each configuration along the epochs is plotted to show the effectiveness of each block. The priors are easy to insert into any network. They only marginally increase the number of parameters, but significantly improve the accuracy and convergence of the algorithms as shown in Supplementary Fig.3. The basic methods of SRCNN and VDSR incorporating the facial rendered priors tend to avoid some artifacts around key facial components and generate sharper edges compared to the baseline methods without the facial priors. By adding the Spatial Attention Module, it helps the network better exploit the priors and easily enables to generate sharper facial structures as shown in Figure 9.

Results on Real-World Images: For real-world LR images, we provide the quantitative and qualitative analysis on 500 LR faces from the WiderFace (x4) dataset in Supplementary Tab.1 and Fig.1.

Model Size and Running Time: We evaluate the proposed method and STOA SR methods on the same server with an Intel Xeon W-2123 CPU and an NVIDIA TITAN X GPU. Our proposed SAM3D, embedded with 3D priors, are more lightweight and less time-consuming, shown in Supplementary Fig.2.

6 Conclusions

In this paper, we proposed a face super-resolution network that incorporates the novel 3D facial priors of rendered faces and multi-dimensional knowledge. In the 3D rendered branch, we presented a face rendering loss to encourage a high-quality guided image providing clear spatial locations of facial components and other hierarchical information (*i.e.,* expression, illumination, and face pose). Compared with the existing 2D facial priors whose attentions are focused on the distinct points of landmarks which may result in face distortions, our 3D priors are explicit, visible and highly realistic, and can largely decrease the occurrence of face artifacts. To well exploit 3D priors and consider the channel correlation between priors and inputs, we employed the Spatial Feature Transform and Attention Block. The comprehensive experimental results have demonstrated that the proposed method achieves superior performance and largely decreases artifacts in contrast with the SOTA methods.

Acknowledgement

This work is supported by the National Key R&D Program of China under Grant 2018AAA0102503, Zhejiang Lab (NO.2019NB0AB01), Beijing Education Committee Cooperation Beijing Natural Science Foundation (No.KZ201910005007), National Natural Science Foundation of China (No.U1736219) and Peng Cheng Laboratory Project of Guangdong Province PCL2018KP004.

References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In ACM SIGGRAPH (1999)
2. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: CVPR (2016)
3. Cao, Q., Lin, L., Shi, Y., Liang, X., Li, G.: Attention-aware face hallucination via deep reinforcement learning. In: CVPR (2017)
4. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: CVPR (2018)
5. Dahl, R., Norouzi, M., Shlens, J.: Pixel recursive super resolution. In: ICCV (2017)
6. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: CVPRW (2019)
7. Dong, C., Loy, C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI **38(2)**, 295–307 (2016)
8. Dong, C., Loy, C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV (2016)
9. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: CVPRW (2019)
10. Grm, K., Scheirer, W., Štruc, V.: Face hallucination using cascaded super-resolution and identity priors. TIP **29**, 2150–2165 (2019)
11. Han, C., Shan, S., Kan, M., Wu, S., Chen, X.: Face recognition with contrastive convolution. In: ECCV pp. 120–135 (2018)
12. Haris, M., Shakhnarovich, G., Ukita, N.: Deep backprojection networks for super-resolution. In: CVPR (2018)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
14. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: ICCV (2017)
15. Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. In: NIPS (2015)
16. Kim, D., Kim, M., Kwon, G., Kim, D.: Progressive face super-resolution via attention to facial landmark. In: BMVC (2019)
17. Kim, J., Lee, J., Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
18. Kim, J., Lee, J., Lee, K.: Deeply recursive convolutional network for image super-resolution. In: CVPR (2016)
19. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR (2017)
20. Li, Z., Tang, J., Zhang, L., Yang, J.: Weakly-supervised semantic guided hashing for social image retrieval. IJCV (2020)
21. Lian, S., Zhou, H., Sun, Y.: A feature-guided super-resolution generative adversarial network for unpaired image super-resolution. In: NIPS (2019)
22. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW pp. 1646–1654 (2017)
23. Liu, C., Shum, H., Freeman, W.: Face hallucination: Theory and practice. IJCV **75(1)**, 115–134 (2007)
24. Liu, W., Lin, D., Tang, X.: Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In: CVPR (2005)

25. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
26. Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: ICML (2016)
27. Ramamoorthi, R., Hanrahan., P.: An efficient representation for irradiance environment maps. in annual conference on computer graphics and interactive techniques. In SIGGRAPH pp. 497–500 (2001)
28. Ren, W., Yang, J., Deng, S., Wipf, D., Cao, X., Tong, X.: Face video deblurring via 3d facial priors. In: ICCV (2019)
29. Shen, Z., Lai, W., Xu, T., Kautz, J., Yang, M.: Deep semantic face deblurring. In: CVPR (2018)
30. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In: CVPR (2016)
31. Tai, Y., Yang, J., Liu, X.: Image superresolution via deep recursive residual network. In: CVPR (2017)
32. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niener, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
33. Wang, X., Tang, X.: Hallucinating face by eigen transformation. TSYST MAN CY C **35(3)**, 425–434 (2005)
34. Wang, X., Yu, K., Dong, C., Loy, C.: Recovering realistic texture in image superresolution by deep spatial feature transform. In: CVPR (2018)
35. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: ECCV (2018)
36. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: CVPR (2018)
37. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: ECCV (2016)
38. Yu, X., Porikli, F.: Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: CVPR (2017)
39. Yu, X., Porikli, F.: Imagining the unimaginable faces by deconvolutional networks. TIP **27(6)**, 2747–2761 (2018)
40. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: A step towards the solution. In: CVPRW (2017)
41. Zhang, K., Zhang, Z., Cheng, C., Hsu, W., Qiao, Y., Liu, W., Zhang, T.: Superidentity convolutional neural network for face hallucination. In: ECCV (2018)
42. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018)
43. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
44. Zhao, J., Xiong, L., Li, J., Xing, J., Yan, S., Feng, J.: 3d-aided dual-agent gans for unconstrained face recognition. TPAMI **41**, 2380–2394 (2019)
45. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM computing surveys (CSUR) **35(4)**, 399–458 (2003)
46. Zhou, E., Fan, H.: Learning face hallucination in the wild. In: AAAI (2015)
47. Zhu, S., Liu, S., Loy, C., Tang, X.: Deep cascaded bi-network for face hallucination. In: ECCV (2016)