**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2022

Face Restoration via Plug-and-Play 3D Facial Priors

Hu, Xiaobin ; Ren, Wenqi ; Yang, Jiaolong ; Cao, Xiaochun ; Wipf, David P ; Menze, Bjoern ; Tong, Xin ; Zha, Hongbin

Abstract: State-of-the-art face restoration methods employ deep convolutional neural networks (CNNs) to learn a mapping between degraded and sharp facial patterns by exploring local appearance knowledge. However, most of these methods do not well exploit facial structures and identity information, and only deal with task-specific face restoration (e.g., face super-resolution or deblurring). In this paper, we propose cross-tasks and cross-models plug-and-play 3D facial priors to explicitly embed the network with the sharp facial structures for general face restoration tasks. Our 3D priors are 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 very 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, for better exploiting this hierarchical information (i.e., intensity similarity, 3D facial structure, and identity content), a spatial attention module is designed for image restoration problems. Extensive face restoration experiments including face super-resolution and deblurring demonstrate that the proposed 3D priors achieve superior face restoration results over the state-of-the-art algorithms

DOI: <https://doi.org/10.1109/tpami.2021.3123085>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-214478>

Journal Article

Accepted Version

Originally published at:

Hu, Xiaobin; Ren, Wenqi; Yang, Jiaolong; Cao, Xiaochun; Wipf, David P; Menze, Bjoern; Tong, Xin; Zha, Hongbin (2022). Face Restoration via Plug-and-Play 3D Facial Priors. IEEE transactions on software engineering, 44(12):8910-8926.

DOI: <https://doi.org/10.1109/tpami.2021.3123085>

Face Restoration via Plug-and-Play 3D Facial Priors

Xiaobin Hu, Wenqi Ren, Jiaolong Yang, Xiaochun Cao, David Wipf, Bjoern Menze, Xin Tong, and Hongbin Zha

Abstract—State-of-the-art face restoration methods employ deep convolutional neural networks (CNNs) to learn a mapping between degraded and sharp facial patterns by exploring local appearance knowledge. However, most of these methods do not well exploit facial structures and identity information, and only deal with task-specific face restoration (e.g., face super-resolution or deblurring). In this paper, we propose cross-tasks and cross-models plug-and-play 3D facial priors to explicitly embed the network with the sharp facial structures for general face restoration tasks. Our 3D priors are 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 very 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, for better exploiting this hierarchical information (i.e., intensity similarity, 3D facial structure, and identity content), a spatial attention module is designed for the image restoration problems. Extensive face restoration experiments including face super-resolution and deblurring demonstrate that the proposed 3D priors achieve superior face restoration results over the state-of-the-art algorithms.

Index Terms—Face restoration, 3D facial priors, 3D morphable knowledge, facial structures, identity knowledge.

1 INTRODUCTION

Face images provide essential knowledge for human observation as well as computer analysis [1], [2]. However, when the resolution of a facial image is relatively low or face images are blurry, the performance of most face image tasks, such as face recognition and facial emotion detection [3], [4], degrades dramatically. Consequently, face restoration, including face super-resolution (SR) and deblurring, was coined from its low-resolution or blurry counterpart to restore a sharp face image. Although a significant influx of deep learning algorithms [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] have been successfully implemented in face restoration problems, rehabilitating sharp facial images is still an open and challenging problem due to the ill-posed nature of the problem and the difficulty in learning and integrating strong priors into a face restoration model.

Some investigations [15], [16], [17], [18], [19], [20], [21] have been proposed on exploiting the face priors to guide neural networks in capturing more facial details. A face hallucination model [15] was proposed by incorporating identity prior which was extracted only from the multi-scale

up-sampling results in the training procedure. Therefore, this kind of identity information cannot provide extra priors to guide the network. Yu et al. [16] used 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. [19] proposed a face alignment network (FAN) for landmark heatmap extraction to improve face SR performance. Chen et al. [20] employed the heatmaps and parsing maps for face SR problems. Shen et al. [22] utilized CNNs to predict the semantic face information that are acted as global priors and local constraints to assist in the position location of facial components. Although these 2D priors provide global component regions, these methods cannot learn the detailed edges, illumination, or other low-level image contexts. In addition, all of these aforementioned face restoration approaches ignore facial structure recovery.

In contrast to the aforementioned approaches, we propose plug-and-play 3D facial priors for face restoration 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 for facilitating the face restoration 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 specific task network is employed to adaptively integrate the 3D facial priors into the network. As shown in Figure 1, for image restoration tasks (e.g., face super-resolution and deblurring), our 3D rendered face priors provide better facial components and skin color similar to the ground truths compared with 2D parsing maps. After embedding the 3D priors, our algorithm generates clearer and sharper facial structures without any ghosting artifacts compared with other 2D prior-based algorithms. The main contributions of

- X. Hu, W. Ren and X. Cao are with School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, P.R. China. email: xbhunanu@gmail.com, {renwenqi, caoxiaochun}@iie.ac.cn. W. Ren is also with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, 100093. X. Hu is also with the Department of Informatics, Technische Universität München, Germany, 80333 Munich, Germany.
- M. Bjoern is with the Department of Informatics, Technische Universität München, Germany, 80333 Munich, Germany. email: bjoern.menze@uzh.ch.
- J. Yang and X. Tong are with Microsoft Research Asia, Beijing, 100080, China. e-mail: {jiaoyan, xtong}@microsoft.com.
- D. Wipf is with the Amazon AI Lab in Shanghai. e-mail: david-wipf@gmail.com.
- H. Zha are with Key Laboratory of Machine Perception (MOE), Department of Machine Intelligence, Peking University, Beijing, 100080, China. e-mail: zha@cis.pku.edu.cn.
- Corresponding author: Wenqi Ren.

Fig. 1. Visual comparison with state-of-the-art face restoration methods. Task I is face super-resolution: (a) 16×16 LR input. (b) 128×128 HR ground-truth. (c) 2D face parsing priors. (d) our 3D rendered priors. (e) Very Deep Residual Channel Attention Network (RCAN) [23]. (f) Progressive Face Super-Resolution using the facial landmark (PSR-FAN) [19]. (g) End-to-End Learning Face Super-Resolution with Facial Parsing Priors (FSRNet) [20]. (h) Our proposed method by incorporating the 3D facial priors. Task II is face deblurring: (i) Blurry input. (j) Ground truth. (k) 2D face parsing priors. (l) our 3D rendered priors. (m) Nah et al. [24]. (n) Pan et al. [25]. (o) Shen et al. [22] using 2D parsing priors. (p) Our proposed method by incorporating the 3D facial priors.

this paper are summarized as follows:

- A novel 3D facial prior 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., semantic maps, landmark heatmaps, and parsing maps).
- To well adapt to the 3D reconstruction of low-resolution or blurry face images, we present a new skin-aware loss function projecting the constructed 3D coefficients onto the rendered images.
- Our proposed 3D facial prior has a high flexibility because its modular structure allows for easy cross-tasks and cross-models plug-in of face restoration tasks (e.g., super-resolution and deblurring) or existing CNN-based face restoration models (e.g., SRCNN [26], VDSR [27], and Nah [24]).

Both qualitative and quantitative experiments demonstrate the proposed network achieves better criteria and superior visual quality compared to state-of-the-art face restoration methods.

2 RELATED WORK

In this section, we discuss recent research on face super-resolution and deblurring 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 [28] demonstrates the effectiveness of CNN to learn the mapping between LR and HR pairs, a lot of CNN architectures have been proposed for SR [29], [30], [31], [32], [33], [34], [35]. Most of the existing high-performance SR networks have residual blocks [27] to go deeper in the network

architecture, and achieve better performance. EDSR [36] improved the performance by removing unnecessary batch normalization layers in residual blocks. A residual dense network (RDN) [37] was proposed to exploit the hierarchical features from all the convolutional layers. Zhang et al. [23] 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. [38] used a spatial feature transform layer to introduce the semantic prior as an additional input of the SR network. Huang et al. [39] presented a wavelet-based CNN approach that can ultra-resolve a very low-resolution face image in a unified framework. Lian et al. [40] 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 and are designed for natural images but fail to leverage facial priors for face super resolution.

Deblurring Neural Networks. Because of the spatially invariant blur kernels, conventional image deblurring methods fail to remove non-uniform motion blurs [41], [42]. To address this issue, deblurring algorithms using deep learning have been proposed to deal with non-uniform motion blurs. Sun et al. [43] proposed a CNN-based model to approximate the blur kernel and then remove the motion blur by estimating the regression relationship between the blurry and sharp images. Su et al. [44] presented an encoder-decoder model for hand-held devices deblurring, which is not restricted by the spatially-aligned pairs. Nah et al. [24] proposed a multi-scale CNN for image deblurring by imitating the coarse-to-fine strategy in conventional deblurring approaches. GAN-based methods (DeblurGAN [45] and DeblurGAN-V2 [46]) are also able to remove blurs by adversarial learning due to their merits in preserving texture details. Zhang et al. [47] proposed a reblurring model that

Fig. 2. The proposed face restoration 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 restoration guided by the facial coefficients and rendered sharp face structures which are concatenated by the Spatial Feature Transform (SFT) layer. For face restoration branch, it comprises four parts: 1). Feature Extraction branch uses a series of convolutional layers to extract the features of the priors. 2). Spatial Attention branch employs SFT layers to well incorporate the facial rendered priors. 3). Residual Channel Attention explores the knowledge and correlations between the channels. 4). HR branch is to reconstruct HR images. The detailed configuration of the proposed method is listed in the appendix.

learns how to blur sharp images and guide the deblurring GAN to efficiently restore sharp images. However, all these networks are intended for natural images or videos and not easily be modified for face deblurring.

Facial Prior Knowledge. Exploiting facial priors in face restoration, such as the spatial configuration of facial components [22], [48], [49], is the key factor that differentiates it from generic restoration tasks. Wang and Tang [50] 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. [51] 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 [52] interweaved multiple spatial transformer networks [53] with the deconvolutional layers to handle unaligned LR faces. Dahl et al. [9] leveraged the framework of PixelCNN [54] to super-resolve very low-resolution faces. Zhu et al. [7] presented a cascade bi-network (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. [16] used a multi-task convolutional neural network to incorporate structural information of faces. Grm et al. [15] built a face recognition model that acts as identity priors for the super-resolution network during training. Yu et al. [20] constructed an end-to-end SR network to incorporate the facial landmark heatmaps and parsing maps. In Super-FAN, Bulat et al. [55] proposed an

end-to-end system to simultaneously tackle two tasks (i.e., improves face resolution and detects the facial landmarks). The joint training helps the network to learn more structural information via integrating a face-alignment sub-network through heatmap regression and loss optimization. Inspired by this idea, Kim et al. [19] proposed a compressed version of the face alignment network (FAN) to obtain landmark heatmaps for the SR network in a progressive method. In [56], considering audio carries information about facial attributes (e.g., gender and age), Givi Meishvili et al. [56] proposed a novel method for face super-resolution task by building the latent representations of a face from the lone audio track and the lone low-resolution image. This work focuses on the benefits of the audio language process to face reconstruction. Nonetheless, combining the images and audios priors for face SR is an interesting strategy to the face restoration community, yet it is beyond the scope of our work that does not have audios information. Zhang et al. [57] proposed a Copy and Paste Generative Adversarial Network for face hallucination using an external guided face that is difficult to obtain in some scenarios. Different from the external priors, our 3D facial priors are generated based on the internal blurry and low-resolution input images, which require no extra guided images. Due to the fact that misalignment LR causes deficient results and artifacts, Yu et al. [58] proposed an end-to-end multi-scale transformative discriminative neural network for hallucinating unaligned face images. In addition, to overcome the limitation that facial priors (e.g., landmark and component maps) are always generated by low-resolution images, Ma et al. [59] proposed

Fig. 3. The visual examples that compare our 3D priors against 2D parsing priors. As shown, it verifies the superiority of our 3D priors and the statement that the 2D parsing priors do not have corresponding position and shape of faces but our 3D priors provide clear spatial positions and skin color of the facial components, especially on large pose variation. The first row shows the facial priors from low-resolution inputs and the last row shows the facial priors from blurry ones. (a) low-resolution inputs. (b) 2D parsing priors obtained from low-resolution inputs. (c) our 3D rendered face structures obtained from low-resolution inputs. (d) ground-truths of the low-resolution inputs. (e) blurry inputs. (f) 2D parsing priors obtained from blurry inputs. (g) our 3D rendered face structures obtained from blurry inputs. (h) the Ground truth of the blurry inputs.

a deep face super-resolution (FSR) model with iterative collaboration between two recurrent networks focusing on facial image recovery and landmark estimation respectively.

There are also some face deblurring methods that exploit facial priors to obtain sharp face images. HaCohen et al. [60] guided face deblurring by using the other reference images with some similar content. Anwar et al. [61] reconstructed the band-pass frequency components for face images via using a frequency-domain class-specific prior. Pan et al. [62] proposed a face image deblurring algorithm by matching the blurred image with a sharp face from an external exemplar dataset. Shen et al. [22] generated semantic labels of faces as a global prior to constrain the location of facial components for restoration, but it was affected by the misleading face segmentation. However, existing face restoration 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 restoration. The identity descriptor from the 3D face reconstruction network is also incorporated to retain identity-aware facial details.

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) [63], [64], [65] 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.

3 THE PROPOSED METHOD

The proposed face restoration framework presented in Figure 2 mainly consists of the 3D priors extraction and the downstream face restoration, which focuses on the face super-resolution and deblurring tasks. Given a low-resolution or blurry face image, we first use the 3D rendering branch to extract the 3D face coefficients. A high-resolution and sharp rendered image is then generated using the 3D coefficients and regarded as the high-resolution facial prior, which facilitates the face restoration process in the spatial attention module.

3.1 Motivations and Advantages of 3D Facial Priors

Existing face restoration 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 different from recently related 2D prior works (e.g., the parsing maps and facial landmark heatmaps by FSRNet [20], the landmark heatmap extraction by FAN [19], and semantic-based priors by Shen et al. [22]). The 3D coefficients contain 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 realistic and robust HR results, significantly reducing artifacts even for large pose variations and partial occlusions. Given low-resolution or blurry face images, the generated 3D rendered reconstructions are shown in Figure 3. The rendered face predictions contain the explicit spatial knowledge and sharp visual quality of facial components close to the ground-truth, even in images containing large pose variations as shown in the second and third rows of Figure 3. Therefore, we concatenate the reconstructed face image as an additional feature in the face restoration network. The facial expression, identity, texture, element-concatenation of illumination, and face pose are transformed into four feature maps and fed into the spatial feature transform block of the face restoration network.

There are typical problems to overcome for real-world applications of the 3D face morphable model, including large pose variations and partial occlusions. As shown in Figure 3, the morphable model can generate realistic reconstructions of large pose variations, containing 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 face restoration algorithms which are blind to unknown degradation types, our 3D model can robustly generate the 3D morphable priors to guide the face restoration branch to grasp the explicit spatial knowledge and facial components, even for complicated real-world applications. Furthermore, our 3D priors can be plugged into any network and significantly improve the performance of existing natural image restoration networks (e.g., SRCNN [26], VDSR [27], and Nah [24] demonstrated in Section 7).

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 or blurry 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, inspiration came from the idea that the 3DMM coefficients can analytically model the pose variations with a simple mathematical derivation [64], [65] 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 $x = (\alpha, \beta, \delta, \gamma, \rho) \in \mathbb{R}^{239}$, where $\alpha \in \mathbb{R}^{80}$, $\beta \in \mathbb{R}^{64}$, $\delta \in \mathbb{R}^{80}$, $\gamma \in \mathbb{R}^9$, and $\rho \in \mathbb{R}^6$ represent the identity, facial expression, texture, illumination, and face pose [64], respectively.

$\mathbf{T}$ and $\mathbf{S}$ are the 3D texture vector and shape vector, respectively. $\mathbf{T}$ and $\mathbf{S}$ are the basic ingredients which constitute 3D facial priors via a linear superposition approach. In other words, face images can be linearly decomposed as different texture vector and shape vector [63]. For more details regarding the 3D face model and the rendering process, we refer the readers to [64]. According to the Morphable model [63], we transform the face coefficients to a 3D shape $\mathbf{S}$ and texture $\mathbf{T}$ of the face image as

$$\mathbf{S} = \mathbf{S}(\alpha, \beta) = \bar{\mathbf{S}} + \mathbf{B}_{id}\alpha + \mathbf{B}_{exp}\beta, \quad (1)$$

and

$$\mathbf{T} = \mathbf{T}(\delta) = \bar{\mathbf{T}} + \mathbf{B}_t\delta, \quad (2)$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ 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) [66] to derive the illumination coefficient $\gamma \in \mathbb{R}^9$. The 3D face pose $\rho \in \mathbb{R}^6$ is represented by rotation $\mathbf{R} \in \text{SO}(3)$ 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

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

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 [64]. In addition, x represent the LR or blurry inputs, $B(x)$ denotes the regressed coefficients obtained by the ResNet-50 with input x , and R denotes the image rendered by 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 restoration.

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 $\alpha, \beta, \delta, \gamma, \rho$ 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 ($\gamma + \rho$) are reshaped to four matrices by setting extra elements to zeros. Afterward, these four matrices are expanded to the same size as the input image 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 [38] 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 modulation parameter pair adaptively controls the outputs of the SFT layer 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:

$$(\mu, \nu) = \Theta(\psi), \quad (4)$$

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

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

where 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 guidance to bias the allocation of available processing resources towards the most informative

Fig. 5. Residual channel attention block (RCAB): an attention mechanism to explore the inter-dependency between inputs and facial priors.

components of the input [67]. 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 [23], the attention mechanism is composed of a series of residual channel attention blocks (RCAB) shown in Figure 5. For the b -th block, the output F_b of RCAB is obtained by:

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

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

4 FACE SUPER-RESOLUTION

Image super-resolution is to rehabilitate the corresponding high-resolution images from the low-resolution inputs. Generally, the low-resolution image I_x is modeled as the degradation process of high-resolution image:

$$I_x = \mathcal{D}(I_y; \delta), \quad (7)$$

where $\mathcal{D}$ denotes a degradation function, I_y is a high-resolution image, and δ is the parameters of degradation process (e.g., noise or scaling factor). Assuming the low-resolution image is provided and the degradation process is unknown, an high-resolution image $\hat{I}_y$ is approximated from the LR image I_x :

$$\hat{I}_y = \mathcal{F}(I_x; \theta), \quad (8)$$

where $\mathcal{F}$ is the super-resolution model with the parameters θ . Although the degradation process of HR images in real-world is unknown, most works ([68], [14], [19], [20]) directly model the process as a down-sampling operation as:

$$\mathcal{D}(I_y; \delta) = (I_y) \downarrow_s, \{s\} \in \delta, \quad (9)$$

where $\downarrow_s$ is a down-sampling operation with the scaling factor s . Furthermore, considering the combination of several degraded operations [69], this formation can be reorganized as:

$$\mathcal{D}(I_y; \delta) = (I_y \otimes k) \downarrow_s + n_\epsilon, \{k, s, \epsilon\} \in \delta, \quad (10)$$

where $I_y \otimes k$ means the convolution between the blur kernel k and unknown HR image I_y , n_ϵ is the independent Gaussian noise term with standard deviation ϵ . This combination of degraded pattern is closer to the real world scenario. Then, the objective of the single image super-resolution task can be summarized as:

$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}(\hat{I}_y, I_y) \quad (11)$$

where $\mathcal{L}(\hat{I}_y, I_y)$ denotes the loss function between the generated HR images $\hat{I}_y$ and ground truth I_y .

4.1 Comparisons with state-of-the-arts

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) [27], the Very Deep Residual Channel Attention Network (RCAN) [23], the Residual Dense Network (RDN) [37], the Super-Resolution Convolutional Neural Network (SRCNN) [28], the Transformative Discriminative Autoencoder (TDAE) [52], the Wavelet-based CNN for Multi-scale Face Super Resolution (Wavelet-SRNet) [39], the deep end-to-end trainable face SR network (FSRNet) [20], face SR generative adversarial network (FSRGAN) [20] incorporating the 2D facial landmark heatmaps and parsing maps, and the progressive face Super Resolution network via face alignment network (PSR-FAN) [19] 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 [28] and VDSR [27]. The implementation code will be made available to the public.

4.2 Datasets and Implementation Details

CelebA [70] and Menpo [71] 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. 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., [19], [20], [16], [10]) to 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 128×128 pixels. The LR face images are generated by downsampling HR ground-truths to 32×32 pixels ($\times 4$ scale) and 16×16 pixels ($\times 8$ scale). We employ the random mirror as our data augmentation strategy. 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 is 0.0002 and is divided by 2 every 50 epochs. The whole training process takes 2 days with an NVIDIA Titan X GPU.

4.3 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

Fig. 6. Face super-resolution comparison with state-of-the-art methods: magnification factors $\times 8$ and the input resolution 16×16 . Best viewed by zooming in on the screen. Here we compare our algorithm with single image super-resolution approaches (VDSR [27], RCAN [23], RDN [37], SRCNN [28]), face super-resolution methods (TDAE [52], Wavelet-SRNet [39]), and 2D facial prior-based methods (FSRNet [20], FSRGAN [20], PSR-FAN [19]).

face poses (left, right, and semifrontal) of the Menpo test data are listed in Table 1.

CelebA Test: As shown in Table 1, VDSR+3D (the basic VDSR model [27] 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 $\times 8$ SR) even for the large-scale parameter methods, such as RDN and RCAN. Our 3D priors algorithm performs much better (1 dB higher) than the recent landmark-based deep iterative collaboration model (DIC) [59]. 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 significantly contribute to the performance improvement (average 1.6 dB improvement) of face super-resolution. Our algorithm performs much better in comparison with 2D priors-based methods (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 1. While ours (SAM3D) is the best method superior to the others, VDSR+3D also achieves 1.8

dB improvement compared with the basic VDSR method in the $\times 4$ 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.

4.4 Qualitative Evaluation

The qualitative results of our method at the magnifications ($\times 8$) are shown in Figure 6. It can be observed that our proposed method recovers clearer faces with finer component details (e.g., noses, eyes, 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 ghosting artifacts. These results illustrate that the proposed 3D priors help the network understand the spatial location and the entire facial 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.

TABLE 1
Face super-resolution quantitative results on the CelebA and Menpo test dataset with different large facial pose variations (e.g., left, right, and semifrontal).

-	Menpo												CelebA			
Scale	×4						×8						×4		×8	
Pose	Left		Right		Semi-frontal		Left		Right		Semi-frontal		-			
	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	PSNRSSIM	
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	27.16	0.8197	21.90	0.6213
VDSR [27]	26.99	0.8024	26.85	0.7908	25.63	0.7794	22.28	0.6315	22.20	0.6163	20.98	0.5752	28.13	0.8554	22.76	0.6618
RCAN [23]	27.47	0.8259	27.27	0.8145	26.11	0.8080	21.94	0.6543	21.87	0.6381	20.60	0.5938	29.04	0.8643	23.26	0.7362
RDN [37]	27.39	0.8263	27.21	0.8150	26.06	0.8088	22.30	0.6706	22.24	0.6552	21.02	0.6160	29.06	0.8650	23.69	0.7484
SRCNN [28]	26.92	0.8038	26.74	0.7913	25.50	0.7782	22.38	0.6408	22.32	0.6272	21.08	0.5857	27.57	0.8452	22.51	0.6659
TDAE [52]	-	-	-	-	-	-	21.22	0.5678	20.22	0.5620	19.88	0.5521	-	-	20.10	0.5802
Wavelet-SRNet [39]	26.97	0.8122	26.81	0.8001	25.72	0.7945	21.86	0.6360	21.72	0.6166	20.57	0.5779	28.42	0.8698	23.08	0.7147
FSRGAN [20]	-	-	-	-	-	-	23.00	0.6326	22.84	0.6173	22.00	0.5938	-	-	22.27	0.6010
FSRNet [20]	-	-	-	-	-	-	23.56	0.6896	23.43	0.6712	22.03	0.6382	-	-	22.62	0.6410
PSR-FAN [19]	-	-	-	-	-	-	22.04	0.6239	21.89	0.6114	20.88	0.5711	-	-	22.66	0.6850
DIC [59]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.43	0.7166
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	29.29	0.8727	24.66	0.7127
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	29.69	0.8817	25.39	0.7551

Fig. 7. Face super-resolution 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). Here we compare our algorithm with single image super-resolution approaches (VDSR [27], RCAN [23], RDN [37], SRCNN [28]), face super-resolution methods (TDAE [52], Wavelet-SRNet [39]), and 2D facial prior-based methods (FSRNet [20], FSRGAN [20], PSR-FAN [19]).

Comparison with latent-space-based SR method: PULSE [72] and others [73], [74], [75] can generate high-fidelity images by using the priors of a 2D generative model in latent space. Instead of directly building the mapping between Low-resolution and High-resolution images, PULSE explores the high-resolution natural image manifold and then searches for images that downscale to the original LR images. But there exist some concerns: *i*). PULSE fails to deal with an unknown downscaling function but only focused on the bicubically downsampled images. It unavoidably limits its performance on real-world super-resolution or deblurring scenarios. *ii*). In the PULSE, the search space of outputs is restricted to be realistic by using the high-dimensional Gaussians. However, the identity information of low-resolution images is challenging to be retained and

further recovered especially in some difficult scenarios (e.g., real-world data or image data beyond latent space). As admitted by authors, PULSE makes imaginary faces of people who do not exist and does not help identify or reconstruct the original image. We qualitatively compare our algorithms against PULSE on the super-resolution task in Figure 8 and on the deblurring task in Figure 9. We found that our proposed method recovers clear and identical identity information compared with the Ground-truth (GT) while the PULSE can recovery very high-realistic and high-fidelity images but fails to retain the same identity information. PULSE [72] is a self-supervised face super-resolution model without requiring training on the LR-HR paired dataset. Following the comparison way [74], [75], PULSE is also trained on the high-resolution FFHQ dataset which consists

TABLE 2

Real-world face super-resolution quantitative results (PIQE and NIQE) with different configurations. 3D denotes the 3D rendered structure priors; ↓ means that the lower of criteria, the better of the image quality.

Criteria	Bicubic	VDSR	VDSR+3D	Ours
NIQE ↓	14.69	14.38	14.17	14.15
PIQE ↓	41.65	45.92	34.86	33.64

of 70,000 high-quality images, and then is evaluated on our CelebA-Test set.

Comparison with Deep Iterative Collaboration: For deep iterative collaboration (DIC) [59], we fine-tune the pre-trained DIC model from authors on our training set with recommended parameters and then evaluate it on the same test set. From qualitative results in Fig. 10, it can be observed that our proposed method recovers clearer faces with finer component details (e.g., noses, eyes, and mouths) while the outputs of DIC [59] contain some artifacts around facial components such as eyes and nose. It could verify that 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.

Fig. 8. Face super-resolution visual comparison with PULSE ($\times 8$).

Fig. 9. Face deblurring visual comparison with PULSE ($\times 8$).

4.5 Results on Real-World Images

For real-world LR images, we provide the quantitative analysis on 500 LR faces from the WiderFace (x4) dataset [76] using two no-reference image criteria (Perception Image Quality (PIQE) and Naturalness Image Quality (NIQE)). The results are sorted in Table 2 (the higher of criteria, the worse the image quality). These results demonstrate that the proposed 3D priors boost the performance even in real-world conditions by capturing the good visual quality of

Fig. 10. Visual comparison with state-of-the-art face super-resolution method [59]. (a) 16×16 LR input. (b) Deep Iterative Collaboration [59]. (c) Our proposed method by incorporating the 3D facial priors. (d) 128×128 HR ground-truth.

facial components and restoring large pose variations and partial occlusions. We add some visual comparisons shown in Figure 11 to illustrate the effectiveness of our 3D priors on real-world faces.

Fig. 11. Visual comparisons of real world face super-resolution images with state-of-the-art methods ($\times 4$).

5 FACE DEBLURRING

Blur artifacts are caused by different reasons (e.g., camera shakes and object movement). Typically, the mathematical formulation of the blurring process ([77], [24]) can be described as:

$$I_x = (I_y \otimes k) + n_\epsilon, \{k, \epsilon\} \in \delta, \quad (12)$$

where I_x , I_y and n_e denote blurry, sharp images, and additive noise, respectively. k is the blur kernel that is unknown in practice. The blind deblurring methods try to estimate the sharp images I_y and the unknown blur kernels k . When the blurry inputs are provided, the corresponding estimated sharp images $\hat{I}_y$ is:

$$\hat{I}_y = \mathcal{N}(I_x; \theta), \quad (13)$$

where $\mathcal{N}$ is the deblurring model with the parameters θ .

5.1 Face image deblurring

5.1.1 Comparisons with state-of-the-arts

In order to evaluate the performance of the proposed algorithm, we provide quantitative and qualitative comparisons with other eight state-of-the-art single image deblurring algorithms, including MAP-based methods ([78], [79], [80], [81], [82]), multi-scale CNN-based method [24], and facial priors methods ([62], [22]).

5.1.2 Datasets and Implementation Details

Following the work [22], we collect a total of 6464 images as our training dataset (2000 images from Helen dataset [83], 2164 images from CMU PIE dataset [84], and 2300 images from the CelebA dataset [70]). We also synthesize 200 motion blur kernels from random 3D camera trajectories [85] as same as the work [22], whose size ranges from 13×13 to 17×17 . We get total 1.3 million blurred images as inputs of our deblurring network after convolving the sharp images with blur kernels and adding 1 percent of Gaussian noise. Besides, for our test dataset, we synthesize another 40 random blur kernels which are different from previous 200 training blur kernels. After the same process of convolving kernels and adding noise, we finally obtain 8000 blurred images for testing from 200 clear images in Helen and CelebA validation datasets. We implement our network using the Pytorch framework. We use a batch size of 16 and set the learning rate to 0.0001 and is divided by 2 every 1 million iterations. We train the deblurring network for 5 million iterations taking about 3 days on NVIDIA Titan X GPU.

5.1.3 Quantitative Results

As shown in Table 3, our algorithm guided by the proposed 3D facial priors achieves much better results (0.92 dB higher than the semantic-priors method [22] on the CelebA test dataset and 0.45 dB higher on the Helen test dataset). It indicates that our algorithm using 3D priors still performs much better than the 2D prior-based method in the face image deblurring task.

5.1.4 Qualitative Evaluation

The qualitative results of our methods are shown in Figure 12. Compared with semantic-based priors, it can be found that our proposed method recovers clearer faces with finer component details (e.g., noses and mouths). When the input images are severe blurry, the outputs of semantic-priors method [22] are easy to generate some artifacts around facial components. After incorporating 3D rendered priors, our algorithm can avoid some artifacts in facial attributes.

Fig. 12. Qualitative results for face image deblurring. Here we compare our algorithm with 2D semantic-based priors deblurring approach [22].

TABLE 3
Face deblurring quantitative comparison with state-of-the-art methods. We compute the average PSNR and SSIM on two test sets.

Methods	Helen		CelebA	
	PSNR	SSIM	PSNR	SSIM
Nah et al. [24]	21.99	0.684	21.41	0.644
Nah et al. [24] +3D	23.01	0.722	22.53	0.689
Shen et al. [22]	25.98	0.822	24.98	0.790
Ours	26.43	0.828	25.90	0.801

5.2 Face Deblurring extended to multi-frame video

For the performance evaluation, we qualitatively and quantitatively compare our algorithm with six start-of-the-art deblurring methods including one video deblurring method [44], two natural image deblurring methods [24], [86], and one face deblurring method [25], [22]. In order to provide fair comparison results, we fine-tuned the model [86] with another 50K iterations and retrained the open-source implementations of the video deblurring model [44] on the same dataset in our work.

5.2.1 Datasets and Implementation Details

Similar to the work [44], we synthesize our training videos from the 300VW face dataset [88]. Since some low-quality videos are inappropriate for our synthesis purpose, we remove blurry or low-resolution videos and use the remaining videos for data synthesis. Overall, 83 videos are selected as our training data and 9 videos are the testing data from the 114 videos in the 300VW dataset. Furthermore, we analyze the generalization ability of our algorithm in the VidTIMIT dataset [89]. The testing data are synthesized by 11 videos randomly selected from the VidTIMIT dataset.

Fig. 13. Qualitative results for real-world blurry face videos. Here we compare our algorithm with single image deblurring approaches [24], [86], video deblurring algorithms [44], [87], and face deblurring methods [25], [22].

TABLE 4

Quantitative PSNR and SSIM results of face video deblurring on the synthetic datasets using different deblurring methods.

	Pan et al. [25]	Shen et al. [22]	Su et al. [44]	Nah et al. [24]	Tao et al. [86]	Ours
9 synthetic testing videos from the 300VW dataset [88]						
PSNR/SSIM	25.46/0.9002	21.94/0.8803	34.40/0.9218	33.72/0.9559	36.36/0.9784	37.70/0.9849
11 synthetic testing videos from the VidTIMIT dataset [89]						
PSNR/SSIM	24.01/0.9113	20.97/0.8684	37.00/0.9850	34.95/0.9805	37.25/0.9757	38.16/0.9871

Fig. 14. Ablation study results of face super-resolution: 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.

5.2.2 Comparisons with state-of-the-arts

300VW dataset. Table 4 reports average PSNR and SSIM results on the synthetic testing data from 300VW. Our deblurring algorithm performs significantly better in terms of quantitative criteria than prior-based methods ([25], [22]).

VidTIMIT dataset. We directly run the different methods on the VidTIMIT dataset to evaluate the ability of model generalization. As shown in Table 4, our algorithm generalizes well to these testing videos and gets higher PSNR and SSIM values than other competitors.

5.2.3 Results on Real-World Images

To evaluate the performance of the proposed method on real-world images, we collect some videos from YouTube which contains blurred face frames caused by camera shake and physical motion. Our algorithm is compared with other six deblurring models including video deblurring [44], [87], natural image deblurring [24], [86], and face image deblurring [25], [22]. Since there are no ground truths, the clearly visible comparisons of the rehabilitated facial components are provided in Figure 13. In contrast with other visible results, our algorithm restores higher higher-fidelity faces with finer facial components details in Figure 13(i), while other 2D priors algorithms [25], [22] show less effectiveness for facial components deblurring in Figure 13 (g) and (h).

5.2.4 Identity Similarity of Facial Recognition

Here we evaluate the identity similarity between the deblurred result and the ground truth computed as the cosine distance of deep face features extracted by a face recognition network from [90]. As shown in Table 5, with the rendered facial structure, the proposed algorithm obtains the highest identity similarity information on the test data. In addition, adding both facial structure and identity knowledge obtains optimal performance in terms of identity similarity scores.

TABLE 5
Identity similarity with different configurations on 9 synthetic testing videos generated from the 300VW dataset.

		Identity similarity	
ident.	rend.	Su [44]	Ours
×	×	0.8352	0.8352
✓	×		0.8335
×	✓		0.8364
✓	✓		0.8373

Fig. 15. Qualitative plug-and-play evaluation of 3D priors 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.

Fig. 16. 3D face rendered results with/without rendering loss. (a) Blurred inputs. (b) Rendered results without using the proposed face rendering loss ℓ_r . (c) Rendered results with the proposed face rendering loss ℓ_r . (d) Ground-truths.

6 ANALYSES AND DISCUSSIONS

6.1 Running Time

In terms of running time, the proposed algorithm performs favorably against the state-of-the-art image and video deblurring methods [24], [86], [44]. The average running times for two image resolutions are shown in Table 6. All the methods are evaluated on the same machine with an Intel(R) Xeon(R) CPU and an Nvidia Titan X GPU. The methods of [24], [86] use a multi-scale strategy which inevitably in-

creases the computational cost. They are clearly less efficient than our method. The video deblurring approach of [44] runs faster than our algorithm on low-resolution frames. However, our method tends to have better scalability and is slightly faster than [44] on higher resolutions. For the super-resolution task, our method still maintains comparable running times than other face SR methods [23], [37] and 2D priors [20] in the appendix.

TABLE 6
Average running time on deblurring testing video frames.

Image size	Nah [24]	Tao [86]	Su [44]	Ours
360×450	2.03s	1.81s	0.31s	0.67s
580×610	2.56s	1.98	0.77s	0.70s

6.2 Ablation Study

Face super-resolution: 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 $\times 8$ scale test data. As shown in Figure 14 (b) and (f), the baseline method without using the rendered faces and SAM tends to generate blurry faces that cannot capture sharp structures. Figure 14 (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 14 (d) and (h). The quantitative comparisons between (VDSR, our VDSR+3D, and our SAM3D) in Table 1 also illustrate the effectiveness of the proposed rendered priors and the spatial attention module.

Face deblurring: As shown in Table 7, the quantitative results demonstrate the effectiveness of our 3D priors. After only incorporating the rendered facial structure, the proposed algorithm achieves the best SSIM values and the second best PSNR values. When adding both facial structure and identity knowledge, only PSNR value is further improved (0.17 dB).

The number determination of Spatial Attention Blocks: We determine the number of spatial attention blocks by mainly considering the balance between the accuracy and the computation time, since the increment of the block number leads to the increase of the computation time and computational complexity. We designed three spatial attention block schemes (6, 8, 14) to show the PSNR results: 6 blocks [29.6920 dB], 8 blocks [29.6942 dB], and 14 blocks [29.7042 dB]. The results indicate that the ensemble number of spatial attention blocks marginally affects the accuracy to some extent.

SAM vs. 3D priors: We conduct two experiments of only employing 3D priors or Spatial attention module (SAM) to demonstrate which part plays a major role. The results are listed here: only SAM (28.92dB) and only 3D priors (29.67dB), which indicate that the 3D priors contribute a major part while the SAM, a tailored mechanism, helps get a further improvement.

6.3 Effectiveness of the Face Rendering Loss

We verify the effectiveness of the proposed face rendering loss ℓ_r in Eq. 3 on the 300VM dataset. From Figure 16(b), although the results without using the face rendering loss are sharp, some face components (e.g., eyes) are not close to the ground truths in (d). Conversely, the results with the face rendering loss ℓ_r in Figure 16(c) recover high-realistic facial structures (e.g., noses, mouths, and eyes) closer to the ground-truths in Figure 16(d) even for large pose variations and occlusions.

Fig. 17. Plug-and-play evaluation of 3D priors: test accuracy curves with different configurations along the training epochs. 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.

6.4 Plug-and-Play Analysis of 3D priors

To verify the advantage of 3D facial structure priors in terms of the flexibility in the plug-and-play character of cross-models, three different configurations are designed: basic methods (*i.e.*, SRCNN [28] and VDSR [27]); 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

TABLE 7
Ablation study results of face deblurring with different configurations on 9 synthetic testing videos generated from the 300VW dataset.

ident.	rend.	Deblurring	
		PSNR	SSIM
×	×	36.74	0.9817
✓	×	36.82	0.9812
×	✓	37.53	0.9869
✓	✓	37.70	0.9849

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 Figure 17. 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. Adding the Spatial Attention Module helps the network better exploit the priors and easily enables generating sharper facial structures as shown in Figure 15.

6.5 Position Deviation Analysis of Rendered 3D Faces

Face super-resolution: As shown in Table 8, we conduct the pixel-offset experiments on 32x32 input images (x4 scale) for the CelebA dataset to analyze the accuracy and the robustness to the position deviation of inaccurate 3D priors. We vertically shift the pixels of rendered 3D face priors from one pixel to twenty pixels. The quantitative results (PSNR) are listed in Table 8. The performance of the inaccurate face priors also behaves better even for 10 pixels-offset than without using the 3D facial priors (28.92dB).

Face image deblurring: We conduct the same pixel-offset experiments on 128x128 blurry images for the CelebA dataset to study the position deviation effect on the deblurring results. As shown in Table 8, the deblurring performance is very robust and resistant to the pixel-shift of 3D priors (only 0.05 dB decrease even for 20 pixels shift).

TABLE 8
Position deviation analysis of rendered 3D priors in the tasks of face super-resolution and deblurring

Pixels Error	Super-resolution	Deblurring
	PSNR(dB)	PSNR(dB)
0	29.69	26.43
1	29.53	26.43
2	29.32	26.43
5	29.08	26.43
10	29.05	26.42
20	29.03	26.38

6.6 Limitations

Failure on the hair rehabilitation: Since the 3D render priors can recover facial components but fail to rehabilitate detailed hairs, our proposed method on the super-resolution

Fig. 18. Failure cases. The first row is the failure case on the hair rehabilitation. The second row is the 3D prior position deviation caused by the blurry boundary and the red line is the low bound of the jaw.

task ignores hair texture and generates blurry hair details in some cases as shown in Figure 18.

3D priors position deviation on the severe blur face: For some severe blurry face images, their face boundaries are indistinguishable. Thus, our 3D rendered branch fails to locate the exact position and leads to the 3D priors position deviation. As shown in Figure 18, the jaw positions of rendered priors move upward compared with the ground truth. It is well-noting that our algorithm still recovers the correct jaw position even for the position deviation of 3D priors. However, when our 3D reconstruction branch fails to exactly locate structures in challenging situations (e.g., severe blur), our deblurring result will also degrade, such as some small artifacts around facial components.

7 CONCLUSIONS

In this paper, we proposed a face restoration network that incorporates the novel 3D facial priors of rendered faces and multi-dimensional knowledge. The 3D facial priors are tailored as a highly cross-tasks and cross-models plug-and-play modular. 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 that may result in face distortions, our 3D priors are explicit, visible, highly realistic, and can largely decrease the occurrence of face artifacts. In addition, we employed the Spatial Feature Transform and Attention Block to well exploit 3D priors and consider the channel correlation between priors and inputs. The comprehensive experimental results have demonstrated that the proposed method largely decreases artifacts and performs favorable against to the state-of-the-art methods.

ACKNOWLEDGEMENT

This work is supported by the National Key R&D Program of China under Grant 2020AAA0109304, National Natural Science Foundation of China (No. 62172409, 62072454, 62025604, 61971016), Beijing Natural Science Foundation (No.4202084), Beihang University (No.VRLAB2021C06), Beijing Nova Program (No. Z201100006820074), and Youth Innovation Promotion Association CAS.

REFERENCES

- [1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35(4), pp. 399–458, 2003.
- [2] Z. Li, J. Tang, L. Zhang, and J. Yang, "Weakly-supervised semantic guided hashing for social image retrieval," *International Journal of Computer Vision*, 2020.
- [3] C. Han, S. Shan, M. Kan, S. Wu, and X. Chen, "Face recognition with contrastive convolution," in *European Conference on Computer Vision*, 2018, pp. 120–135.
- [4] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [5] E. Zhou and H. Fan, "Learning face hallucination in the wild," in *American Association for Artificial Intelligence*, 2015.
- [6] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *European Conference on Computer Vision*, 2016.
- [7] S. Zhu, S. Liu, C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," *European Conference on Computer Vision*, 2016.
- [8] Q. Cao, L. Lin, Y. Shi, X. Liang, and G. Li, "Attention-aware face hallucination via deep reinforcement learning," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [9] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," *IEEE International Conference on Computer Vision*, 2017.
- [10] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [11] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3d-aided dual-agent gans for unconstrained face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2380–2394, 2019.
- [12] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2019.
- [13] W. Liu, D. Lin, and X. Tang, "Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation," *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
- [14] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [15] K. Grm, W. Scheirer, and V. Štruc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 2150–2165, 2019.
- [16] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," *European Conference on Computer Vision*, 2018.
- [17] W. Ren, J. Yang, S. Deng, D. Wipf, X. Cao, and X. Tong, "Face video deblurring via 3d facial priors," *IEEE International Conference on Computer Vision*, 2019.
- [18] K. Zhang, Z. Zhang, C. Cheng, W. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," *European Conference on Computer Vision*, 2018.
- [19] D. Kim, M. Kim, G. Kwon, and D. Kim, "Progressive face super-resolution via attention to facial landmark," *British Machine Vision Conference*, 2019.
- [20] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsnet: End-to-end learning face super-resolution with facial priors," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [21] X. Yu and F. Porikli, "Imagining the unimaginable faces by deconvolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2747–2761, 2018.
- [22] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," *European Conference on Computer Vision*, 2018.
- [24] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.
- [25] J. Pan, W. Ren, Z. Hu, and M.-H. Yang, "Learning to deblur images with exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1412–1425, 2018.

- [26] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014.
- [27] J. Kim, J. Lee, and K. Lee, "Accurate image super-resolution using very deep convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [28] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38(2), pp. 295–307, 2016.
- [29] C. Dong, C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," *European Conference on Computer Vision*, 2016.
- [30] W. Shi, J. Caballero, F. Huszar, J. Totz, A. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [31] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [32] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep backprojection networks for super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [33] J. Kim, J. Lee, and K. Lee, "Deeply recursive convolutional network for image super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [34] Y. Tai, J. Yang, and X. Liu, "Image super resolution via deep recursive residual network," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [35] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 185–200.
- [36] B. Lim, S. Son, H. Kim, S. Nah, and K. Lee, "Enhanced deep residual networks for single image super-resolution," in: *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp. 1646–1654, 2017.
- [37] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [38] X. Wang, K. Yu, C. Dong, and C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [39] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-smnet: A wavelet-based cnn for multi-scale face super resolution," *IEEE International Conference on Computer Vision*, 2017.
- [40] S. Lian, H. Zhou, and Y. Sun, "A feature-guided super-resolution generative adversarial network for unpaired image super-resolution," *Advances in Neural Information Processing Systems*, 2019.
- [41] T. Hyun Kim and K. Mu Lee, "Segmentation-free dynamic scene deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2766–2773.
- [42] T. Hyun Kim, B. Ahn, and K. Mu Lee, "Dynamic scene deblurring," in *IEEE International Conference on Computer Vision*, 2013, pp. 3160–3167.
- [43] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [44] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1279–1288.
- [45] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [46] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8878–8887.
- [47] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, "Deblurring by realistic blurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2737–2746.
- [48] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.
- [49] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting crfs with boltzmann machine shape priors for image labeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2019–2026.
- [50] 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.
- [51] C. Liu, H. Shum, and W. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision*, vol. 75(1), pp. 115–134, 2007.
- [52] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3760–3768.
- [53] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, 2015.
- [54] A. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *International Conference on Machine Learning*, 2016.
- [55] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 109–117.
- [56] G. Meishvili, S. Jenni, and P. Favaro, "Learning to have an ear for face super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1364–1374.
- [57] Y. Zhang, I. W. Tsang, Y. Luo, C.-H. Hu, X. Lu, and X. Yu, "Copy and paste gan: Face hallucination from shaded thumbnails," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7355–7364.
- [58] X. Yu, F. Porikli, B. Fernando, and R. Hartley, "Hallucinating unaligned face images by multiscale transformative discriminative networks," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 500–526, 2020.
- [59] 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/CVF conference on computer vision and pattern recognition*, 2020, pp. 5569–5578.
- [60] Y. Hachohen, E. Shechtman, and D. Lischinski, "Deblurring by example using dense correspondence," in *IEEE International Conference on Computer Vision*, 2013.
- [61] S. Anwar, C. P. Huynh, and F. Porikli, "Image deblurring with a class-specific prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [62] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring face images with exemplars," in *European Conference on Computer Vision*, 2014.
- [63] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *ACM SIGGRAPH*, 1999.
- [64] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in: *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2019.
- [65] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [66] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps. in annual conference on computer graphics and interactive techniques," in *SIGGRAPH*, pp. 497–500, 2001.
- [67] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [68] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, "Face super-resolution guided by 3d facial priors," in *European Conference on Computer Vision*, 2020, pp. 763–780.
- [69] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.
- [70] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *IEEE International Conference on Computer Vision*, 2015.
- [71] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in: *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017.
- [72] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration

- of generative models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2437–2445.
- [73] D. Berthelot, P. Milanfar, and I. Goodfellow, "Creating high resolution images with a latent adversarial generator," *arXiv preprint arXiv:2003.02365*, 2020.
- [74] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9168–9178.
- [75] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 672–681.
- [76] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [77] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *arXiv preprint arXiv:2008.13751*, 2020.
- [78] S. Cho and S. Lee, "Fast motion deblurring," in *ACM SIGGRAPH Asia 2009 papers*, 2009, pp. 1–8.
- [79] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 233–240.
- [80] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *Acm transactions on graphics (tog)*, vol. 27, no. 3, pp. 1–10, 2008.
- [81] L. Xu, S. Zheng, and J. Jia, "Unnatural 10 sparse representation for natural image deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1107–1114.
- [82] L. Zhong, S. Cho, D. Metaxas, S. Paris, and J. Wang, "Handling noise in single image deblurring using directional filters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 612–619.
- [83] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*, 2012, pp. 679–692.
- [84] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2002, pp. 53–58.
- [85] G. Boracchi and A. Foi, "Modeling the performance of image restoration from motion blur," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3502–3517, 2012.
- [86] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [87] T. Hyun Kim and K. Mu Lee, "Generalized video deblurring for dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5426–5434.
- [88] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossai, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *IEEE international conference on computer vision workshops*, 2015, pp. 50–58.
- [89] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, no. 5, pp. 449–480, 2004.
- [90] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4362–4371.

Xiaobin Hu is working toward the Ph.D. degree in Informatics, Technische Universität München (TUM) supervised by Prof. Bjoern Menze since the November of 2017. He received master degree of Hunan University at the 2017. His research interests include image processing (e.g., deblurring and super resolution), medical image analysis, and uncertainty quantification.

Wenqi Ren is an associate professor in Institute of Information Engineering, Chinese Academy of Sciences, China. During 2015 to 2016, he was supported by China Scholarship Council as a joint-training Ph.D. student at the University of California at Merced. He received MSRA Star Track Program in 2018. His research interests include image processing and related high-level vision problems.

Jiaolong Yang is currently a senior researcher at Microsoft Research Asia, Beijing, China. He received the dual Ph.D. degrees in Computer Science and Engineering from the Australian National University and Beijing Institute of Technology in 2016. His research interests include 3D vision for human face and body. He received the outstanding PhD thesis award from China Society of Image and Graphics in 2017.

Xiaochun Cao is a professor with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, P.R.China. Prof. Cao is a Fellow of the IET. His dissertation was nominated for the University of Central Florida's university-level Outstanding Dissertation Award. In 2004 and 2010, he was the recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

David Wipf completed his Ph.D. at the University of California, San Diego as an NSF Fellow in Vision and Learning in Humans and Machines. Subsequently he was an NIH Postdoctoral Fellow at the University of California, San Francisco developing robust Bayesian statistical models for imaging functional brain activity and for finding sparse representations. After several years at Microsoft Research in Beijing, he has now moved to a new position with the Amazon AI Lab.

Bjoern Menze was Professor for Machine Learning in Biomedical Imaging in the Department of Informatics at TU München. He organized workshops at MICCAI, ISBI, NeurIPS, and CVPR in the fields of medical computer vision, served as guest editor for Medical Image Analysis and as a member of the program committee of MICCAI. He received the Medical Image Analysis Award for the best paper of MICCAI 2014, and the Young Scientist Publication Impact Award at MICCAI 2015.

Xin Tong is now a principal researcher in Internet Graphics Group of Microsoft Research Asia. He obtained his Ph.D. degree in Computer Graphics from Tsinghua University in 1999. His Ph.D. thesis is about hardware assisted volume rendering. He got his B.S. Degree and Master Degree in Computer Science from Zhejiang University in 1993 and 1996 respectively. His research interests include appearance modeling and rendering, texture synthesis, and image based modeling and rendering.

Hongbin Zha is a professor in Department of Machine Intelligence, School of EECS, Peking University (PKU). He has served as the Director of Key Lab of Machine Perception (MOE), PKU, since 2007. He obtained his B.Sc. from Hefei University of Technology in 1983, and Ph.D. from Kyushu University, Japan, in 1990, respectively. His research interests include 3D reconstruction and modeling, geometric data processing, sensor localization and mapping, all in computer vision.