FACE SUPER-RESOLUTION GUIDED BY 3D FACIAL PRIORS

Anonymous authors
Paper under double-blind review

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 and misalignments. In this paper, we propose a novel face super-resolution method that explicitly incorporates 3D facial priors which grasp the sharp facial structures. Firstly, a 3D face rendering branch is set up to obtain 3D priors of salient facial structures and identity knowledge. Secondly, the Spatial Attention Mechanism 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 algorithm achieves superior face super-resolution results and outperforms the state-of-the-art.

1 INTRODUCTION

Face images provide crucial clues for human observation as well as computer analysis (Fasel & Luettinb, 2003; Zhao et al., 2003). However, the performance of most face image tasks, such as face recognition and facial emotion detection (Han et al., 2018; Thies et al., 2016), 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 low-resolution face image to its high-resolution counterpart.

A multitude of deep learning methods (Zhou & Fan, 2015; Yu & Porikli, 2016; 2017a; 2018; Zhu et al., 2016; Cao et al., 2017; Dahl et al., 2017a; Yu et al., 2018b) have been successfully applied in face Super-Resolution (SR) problems and achieve state-of-the-art results. But 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 (Grm et al., 2018; Yu et al., 2018a; Ren et al., 2019) on exploiting the face priors to assist neural networks to capture more facial details have been proposed recently. A face hallucination model incorporating identity priors is presented in Grm et al. (2018). But the identity prior is extracted only from the multi-scale up-sampling results in the training procedure and therefore cannot provide enough extra priors to guide the network to achieve a better result. Yu et al. (2018a) employ facial component heatmaps to encourage the upsampling stream to generate super-resolved faces with higher-quality details, especially for large pose variations. Although heatmaps can provide global component regions, it cannot learn the reconstruction of detailed edges, illumination or expression priors. Besides, all of these aforementioned face SR approaches ignore facial structure and identity recovery.

In contrast to previous methods, we propose a novel face super-resolution method that embeds 3D face structures and identity priors. 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 face render prior is generated by the ResNet-50 network (He et al., 2016). It contains rich hierarchical information, such as low-level (e.g., sharp edge, illumination) and perception level (e.g., identity). The Spatial Attention Mechanism is proposed here to adaptively integrate the 3D facial prior into the network. Specifically, we employ the Spatial Feature Transform (SFT) (Wang et al., 2018) to generate affine transformation parameters for spatial feature modulation. Afterwards, it encourages the network to learn the spatial interdepenencies of features between 3D facial priors and input images after adding the attention module into the network.

The main contributions of this paper are: 1. 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 visibility information, which are ignored by the pixel-level content. 2. We propose a feature-fusion-based network to better extract and integrate the face rendered priors by employing the Spatial Attention Mechanism (SAM). 3. We qualitatively and quantitatively explore 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. Thus, 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 can learn to map the relationship between LR and HR (Dong et al., 2016a), a lot of CNN architectures have been proposed for SR (Dong et al., 2016b; Shi et al., 2016; Lai et al., 2017; Haris et al., 2018; Kim et al., 2016; Tai et al., 2017). Most of the existing high-performance SR networks have residual blocks (Jiwon Kim & Lee, 2016) to go deeper in the network architecture, and achieve better performance. EDSR (Lim et al., 2017) improves the performance by removing unnecessary batch normalization layers in residual blocks. A residual dense network (RDN) (Zhang et al., 2018a) was proposed to exploit the hierarchical features from all the convolutional layers. Zhang et al. (2018b) 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. (2018) used a spatial feature transform layer to introduce the semantic prior as an additional input of SR network. Huang et al. (2017) presented a wavelet-based CNN approach that can ultra-resolve a very low resolution face image in a unified framework. However, these networks require a lot of time to train the large-scale parameters to obtain good results. In our work, we largely decrease the training parameters, but still achieve the superior performance in SR criteria (SSIM and PSNR) and visible quality.

Facial Prior Knowledge: Exploiting facial priors in face hallucination, such as spatial configuration of facial components, is the key factor that differentiates it from generic super-resolution tasks. There are some face SR methods that use facial prior knowledge to better super-resolve LR faces. Wang & Tang (2005) learned subspaces from LR and HR face images respectively, and then reconstructed an HR output from the PCA coefficients of the LR input. Liu et al. (2007) set up a Markov Random Field (MRF) to reduce ghosting artifacts because of the misalignments in LR images. These methods are prone to generate severe artifacts, especially in large pose variations and misalignments in LR images. Yu & Porikli (2017b) interweaved multiple spatial transformer networks (Jaderberg et al., 2015) with the deconvolutional layers to handle unaligned LR faces. Dahl et al. (2017b) leveraged the framework of PixelCNN (Van Den Oord et al., 2016) to super-resolve very low-resolution faces. Zhu et al. (2016) 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. (2018a) used a multi-task convolutional neural network (CNN) to incorporate structural information of faces. Grm et al. (2018) built a face recognition model that acts as identity priors for the super-resolution network during training. In our paper, we used the 3D face reconstruction branch to extract the facial structure, detailed edges, illumination, and identity priors. Furthermore, we recover these priors in an explicit way.

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) (Blanz & Vetter, 1999; Deng et al., 2019; Booth et al., 2016) based on the fusion of parametric descriptions of face attributes (e.g., gender, identity, and distinctiveness) to reconstruct the 3D facial priors. The reconstructed face will inherit the facial features and present the clear and sharp facial components.

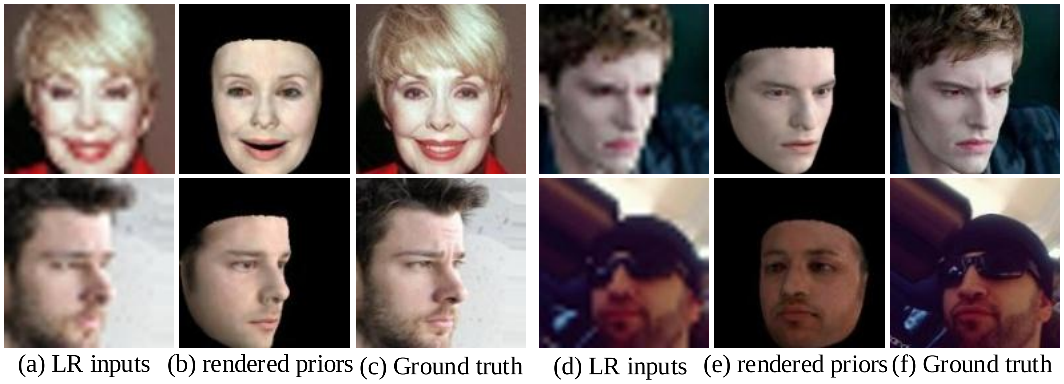


Figure 1: The rendered prior by our method. (a) and (d) low-resolution input. (b) and (e) our rendered face structures. (c) and (f) ground-truths. As shown, the reconstructed facial structures provide clear spatial locations and sharp visualization of facial components even for large pose variations (e.g., left and right facial pose positions).

3 3D FACE RENDERING PRIOR

Given a low-resolution facial image, the 3D rendering branch aims to extract the 3D face coefficients based on the 3D Morphable Model (3DMM). The high-resolution face rendered image is generated after obtaining the 3D coefficients and regarded as the high-resolution facial priors which facilitate the face super-resolution. The 3D coefficients contain abundant hierarchical knowledge, such as identity, facial expression, texture, illumination, and face pose. The proposed face super-resolution framework is presented in Figure 2, and it consists of two branches: the 3D rendering network to extract the facial prior and the Spatial Attention Mechanism aiming to exploit the prior for the face super-resolution problem. 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. Besides, 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 variation with a simple math derivation (Booth et al., 2016; Deng et al., 2019) and does not require a large training set, 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 $\boldsymbol{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) \in \mathbb{R}^{239}$, 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 $\boldsymbol{\zeta} \in \mathbb{R}^{6}$ represent the identity, facial expression, texture, illumination, and face pose (Deng et al., 2019), respectively.

According to the Morphable model (Blanz & Vetter, 1999), 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 the $\mathbf{S}$ and $\mathbf{T}$. Besides, $\mathbf{B}_t$, $\mathbf{B}_{id}$ and $\mathbf{B}_{exp}$ denote the base vector of texture, identity, and expression calculated by the PCA method. A modified L_2 based loss function for the 3D face reconstruction is presented based on a paired training set

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

where j is the paired image index, and L is the total number of training pairs. i and M denote the pixel index and face region, respectively. A, I and B represent the skin color based attention mask, the sharp image, and the up-sampling of low-resolution image, respectively. $R(B_j^i(\boldsymbol{x}))$ denotes the reconstructed face image based on the learned face vector by the ResNet-50 network.

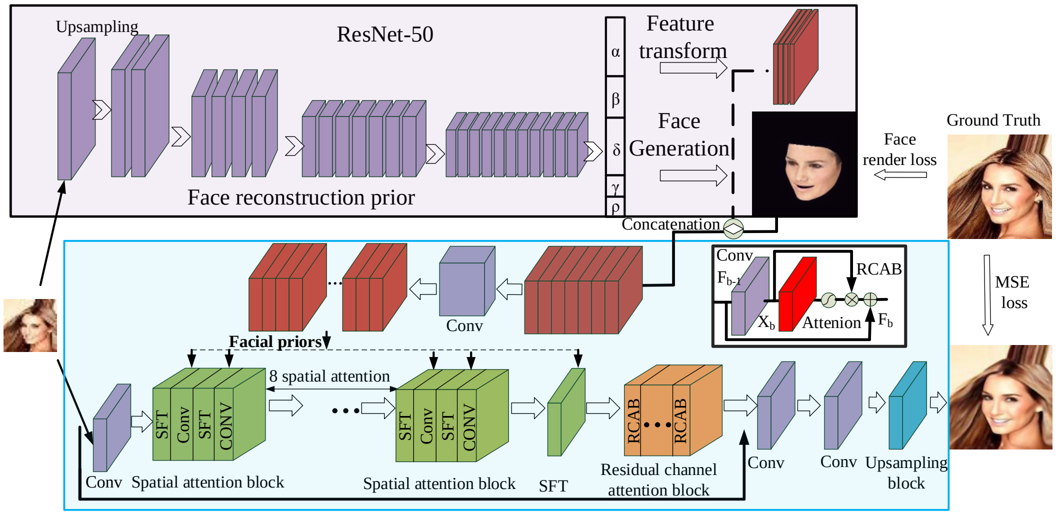


Figure 2: The proposed face super-resolution architecture. Our model consists of two branches: the top block is a ResNet-50 Network to extract the 3D facial coefficients and restore a sharp face rendered structure. The bottom block is dedicated to face super-resolution guided by the facial coefficients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

Given the LR images, the generated 3D face rendered reconstructions are shown in Figure 1. The rendered face predictions contain the clear spatial knowledge and good visual quality of facial components which are very close to the information of the ground-truths. The 3D priors grasp very well the pose variations and skin colour, and further embed pose variations into the SR networks which improve the accuracy and stability in face images with large pose variations. Therefore, we concatenate the reconstructed face image as an additional feature in the SR network. The face expression, identity, texture, illumination, and face pose are transformed into four feature maps and fed into the spatial feature transform block of the SR network.

4 Spatial attention Mechanism

As shown in Figure 2, our Spatial Attention Mechanism aims to exploit the 3D face rendered priors which grasp the precise locations of face components and the facial identity. In order to explore the interdependence and correlation of priors and input images between channels, the attention block is added into the Spatial Attention Mechanism. The proposed network, also named the Spatial Attention Mechanism (SAM), consists of three simple parts: a spatial transform block, an attention block, and an upscale module.

4.1 Spatial Feature Transform Block

We import the 3D face priors into the Spatial Attention Transform Block after a convolutional layer. The 3D face priors consist of two parts: one directly from the rendered face images (as the RGB input), and the other from the feature transformation of the coefficient parameters. The feature transformation procedure is described as follows: firstly, the coefficients of (identity, expression, texture, and the fusion of illumination and face pose) are reshaped to a matrix by setting extra elements to zeros. Afterwards, it is expanded to the same size as the LR images by zero-padding, and then scaled to the interval [0,1]. Finally, the coefficient features are concatenated with the priors from the rendered face images. The Spatial Feature Transform (SFT) learns a mapping function Θ that provides a modulation parameter pair (μ, ν) according to the priors ψ, such as segmentation probability. Instead, the 3D face priors are taken as the input. The outputs of the SFT layer are adaptively controlled by the modulation parameter pair by way of applying an affine transformation spatially to each intermediate feature map. Specifically, the intermediate transformation parameters (μ, ν) are derived from the priors ψ by a mapping function as:

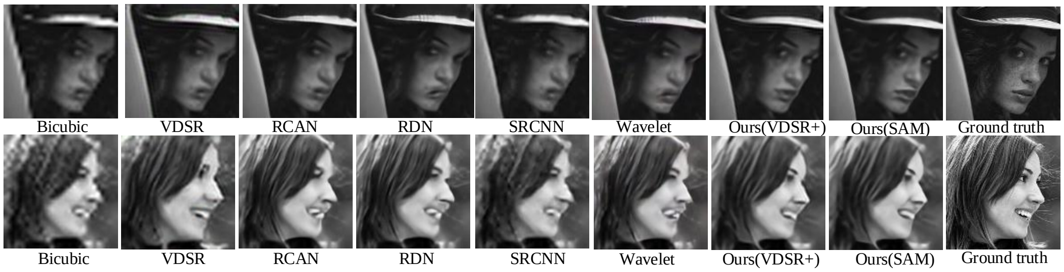


Figure 3: Comparison of state-of-the-art methods: magnification factors ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

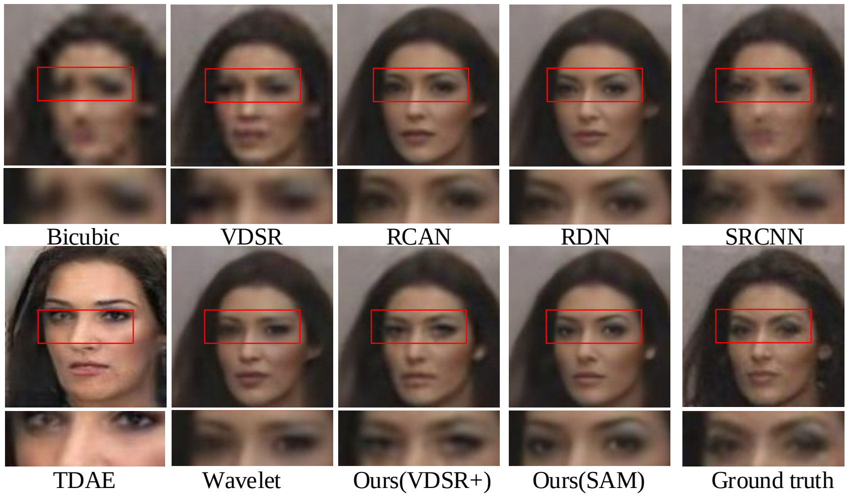


Figure 4: Comparison of state-of-the-art methods: magnification factors ×8 and the input resolution 16×16. Best viewed by zooming in the screen.

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

and then

$$\widetilde{\boldsymbol{y}} = N_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\nu}), \tag{5}$$

where N denotes the SR network, and $\boldsymbol{\theta}$ represents trainable parameters of the network. The intermediate feature maps are modified by scaling and shifting feature maps according to the transformation parameters:

$$\boldsymbol{SFT}(\boldsymbol{F}|\boldsymbol{\mu}, \boldsymbol{\nu}) = \boldsymbol{\mu} \otimes \boldsymbol{F} + \boldsymbol{\nu}, \tag{6}$$

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

4.2 RESIDUAL CHANNEL ATTENTION BLOCK

Attention mechanism can be viewed as a guide to bias the allocation of available processing resources towards the most informative components as input (Hu et al., 2017). Consequently, the channel module is presented to explore the most informative components and the interdependency between the channels. The attention module is composed of a series of residual channel attention blocks (RCAB) shown in Figure 2. Inspired by the integration of channel attention and residual blocks, we ensemble a series of residual channel attention blocks. For the b-th block, the output F_b of RCAB is obtained by:

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

-	CelebA			
Scale	x4		x8	
	PSNR	SSIM	PSNR	SSIM
Bicubic	27.1650	0.8197	21.9001	0.6213
VDSR[1]	28.1393	0.8554	22.7634	0.6618
RCAN[2]	29.0410	0.8643	23.2671	0.7362
RDN[3]	29.0631	0.8650	23.6925	0.7484
SRCNN[4]	27.5775	0.8452	22.5161	0.6659
TDAE[5]	-	-	20.1001	0.5802
Wavelet-SRNet[6]	28.4292	0.8698	23.0874	0.7147
Ours(VDSR+)	29.2991	0.8727	24.6681	0.7127
Ours (SAM)	**29.6942**	**0.8817**	**25.3921**	**0.7551**

$$\boldsymbol{F_b} = \boldsymbol{F_{b-1}} + C_b(\boldsymbol{X_b}) \cdot \boldsymbol{X_b}, \tag{7}$$

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.

5 Experimental results

In order to evaluate the performance of our priors and algorithms, we compare them with the start-of-art methods qualitatively and quantitatively. The six most recent state-of-the-art super-resolution and face hallucination models are listed as follows: the Very Deep Super Resolution Network (VDSR[1]) from Jiwon Kim & Lee (2016), the Very Deep Residual Channel Attention Network (RCAN[2]) (Zhang et al., 2018b), the Residual Dense Network (RDN[3]) (Zhang et al., 2018a), the Super-Resolution Convolutional Neural Network (SRCNN[4]) from Dong et al. (2016a), the Transformative Discriminative Autoencoder (TDAE[5]) (Yu & Porikli, 2017b), and the Wavelet-based CNN for Multi-scale Face Super Resolution (Wavelet-SRNet[6]) (Huang et al., 2017). We use open-resource implementations from the authors and train all the networks on the same dataset for a fair comparison. We propose two models: first is the VDSR+ which is the basic VDSR model embedded with the 3D facial prior as extra RGB channel information and the other is our SR network incorporating facial priors by the Spatial Attention Mechanism (SAM). The implementation code will be made available to the public. More results are shown in the supplementary material.

5.1 Datasets and Implementation Details

CelebA (Liu et al., 2015) and Menpo (Zafeiriou et al., 2017) dataset 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. 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 is 0.0002 and is divided by 2 every 50 epochs. The whole training process takes 2 days with an NVIDIA Titan X GPU.

5.2 Quantitative results by PSNR/SSIM

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

CelebA Test: Ours (VDSR+) achieves significantly better results (1 dB higher than the remaining best method and 2 db higher than the basic VDSR method in x8 SR) even for the large-scale parameter methods, such as RDN and RCAN. But it does perform slightly worse than ours (SAM). It should be noted that ours (VDSR+) is the same as VDSR except for the extra 3D face priors as the

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	x4						x8					
Pose	Left		Right		Semi-frontal		Left		Right		Semi-frontal	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	26.3641	0.7923	26.1996	0.7791	24.9278	0.7608	22.0998	0.6423	21.9968	0.6251	20.6878	0.5770
VDSR[1]	26.9988	0.8024	26.8539	0.7908	25.6345	0.7794	22.2876	0.6315	22.2058	0.6163	20.9821	0.5752
RCAN[2]	27.4712	0.8259	27.2772	0.8145	26.1184	0.8080	21.9499	0.6543	21.8711	0.6381	20.6088	0.5938
RDN[3]	27.3953	0.8263	27.2186	0.8150	26.0646	0.8088	22.3068	0.6706	22.2422	0.6552	21.0214	0.6160
SRCNN[4]	26.9231	0.8038	26.7426	0.7913	25.5044	0.7782	22.3863	0.6408	22.3252	0.6272	21.0805	0.5857
TDAE[5]	-	-	-	-	-	-	21.2201	0.5678	20.2234	0.5620	19.8820	0.5521
Wavelet-SRNet[6]	26.9794	0.8122	26.8146	0.8001	25.7239	0.7945	21.8649	0.6360	21.7296	0.6166	20.5787	0.5779
Ours(VDSR+)	28.6206	0.8439	28.8961	0.8326	26.9951	0.8236	23.4519	0.6845	23.2501	0.6653	21.8344	0.6239
Ours(SAM)	**28.9810**	**0.8510**	**29.2988**	**0.8408**	**27.2978**	**0.8332**	**23.8003**	**0.7071**	**23.5724**	**0.6881**	**22.1507**	**0.6501**

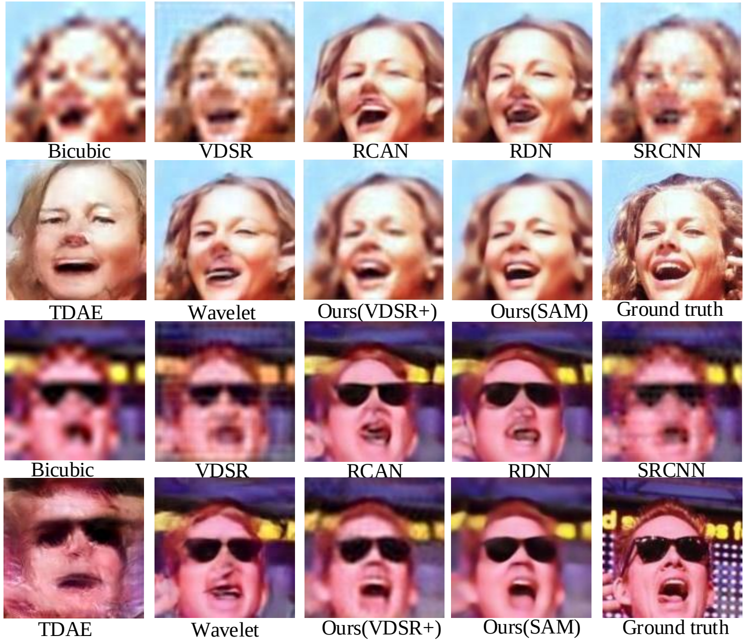


Figure 5: Visual comparison with state-of-the-art methods(×8). The results from the proposed method have less visual artifacts and more details on key face components (e.g., eyes, mouth, and nose)

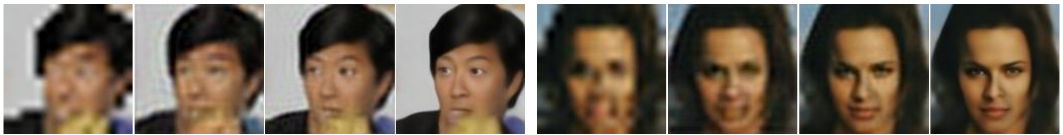
RGB channel inputs. It indicates that the 3D priors make a great contribution to the performance improvement (average 1.6 db improvement) of face super-resolution.

Menpo Test: To verify the effectiveness and stability of face priors and our proposed network towards large pose variations, the PSNR and SSIM results of face poses are listed in Table 2. While ours (SAM) is the best method superior than others, VDSR+ achieves 1.8db improvement compared with the basic VDSR method in the magnification factors (×4).

5.3 Qualitative Evaluation

Super-resolution: The qualitative results of our methods at different magnifications (×4 and ×8) are shown respectively in Figures 3 and 4. It can be observed that our proposed method recovers clearer faces with finer component details (e.g., nose and eyes).

Artifacts: The outputs of most methods (e.g., RCAN, RDN, and Wavelet-SRNet) contain some artifacts around facial components, such as the eyes, nose, and mouth shown in Figure 5. After adding the rendered face priors, ours results show clear and sharp facial structures without any



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

Figure 6: 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 Mechanism. (d) and (h) are our SR results.

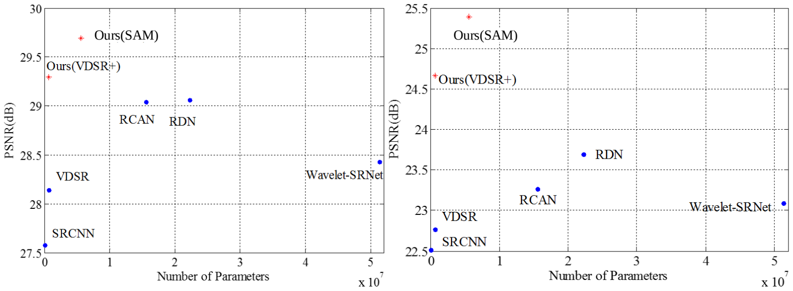


Figure 7: Performance vs. number of parameters. Results are evaluated on CelebA dataset. Left: ×4 scale; Right:×8 scale.

ghosting artifacts. It illustrates that the proposed 3D priors can help the network understand the spatial location and the entire face structure.

5.4 Ablation study and Model Size Analysis

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 Mechanism (SAM) in terms of PSNR and SSIM on the test data. As shown in Figure 6 (b, f), the baseline method without rendered faces and SAM tends to generate blurry faces that cannot capture sharp edges. Figure 6 (c and g) shows clearer and sharper facial structures after adding the rendered priors. By using SAM, the visual quality is further improved in Figure 6 (d and h). The quantitative comparisons between (VDSR, our VDSR+, and our SAM) in Tables 1 and 2 also illustrate the effectiveness of the rendered priors and the Spatial Attention Mechanism.

Model Size Analysis: Figure 7 shows comparisons of model size and performance. Our networks, VDSR+ and SAM, embedded with 3D priors are more lightweight while still achieving the best performance even compared with other state-of-the-art methods (e.g., RCAN and RDN) with a larger scale of parameters.

6 Conclusion

In this paper, we proposed a novel network that incorporates 3D facial priors of rendered faces and identity knowledge. The 3D rendered branch utilizes the 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). To well exploit 3D priors and consider the channel correlation between priors and inputs, the Spatial Attention Mechanism is presented by employing the Spatial Feature Transform and Attention block. The comprehensive experimental results have demonstrated that the proposed method can deliver the better performance and largely decrease artifacts in comparison with the state-of-the-art methods by using significantly fewer parameters.

References

V Blanz and T Vetter. A morphable model for the synthesis of 3d faces. *In SIGGRAPH*, 1999.

J Booth, A Roussos, S Zafeiriou, A Ponniah, and D Dunaway. A 3d morphable model learnt from 10,000 faces. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Q Cao, L Lin, Y Shi, X Liang, and G Li. Attention-aware face hallucination via deep reinforcement learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 690–698, 2017.

R Dahl, M Norouzi, and J Shlens. Pixel recursive super resolution. *International Conference on Computer Vision (ICCV)*, pp. 5439–5448, 2017a.

R Dahl, M Norouzi, and J Shlens. Pixel recursive super resolution. in: International. *Conference on Computer Vision (ICCV)*, pp. 5439–5448, 2017b.

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. *arXiv preprint arXiv:1903.08527*, 2019.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2016a.

Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. *European Conference on Computer Vision*, pp. 391–407, 2016b.

B Fasel and Juergen Luettinb. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

K Grm, W. J S. Dobrišek, and Scheirer, and V Štruc. Face hallucination using cascaded super-resolution and identity priors. *arXiv preprint arXiv:1805.10938*, 2018.

C Han, S Shan, M Kan, S Wu, and X Chen. Face recognition with contrastive convolution. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 120–135, 2018.

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep backprojection networks for super-resolution. *In Conference on Computer Vision and Pattern Recognition*, 2018.

K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1063–6919, 2016.

J Hu, L Shen, and G Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

H Huang, R He, Z Sun, and T Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. *International Conference on Computer Vision (ICCV)*, 6, 2017.

M Jaderberg, K Simonyan, and A Zisserman. Spatial transformer networks. *Advances in Neural Information Processing Systems (NIPS)*, pp. 2017–2025, 2015.

Jung Kwon Lee Jiwon Kim and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply recursive convolutional network for image super-resolution. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645, 2016.

Wei Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and MingHsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *In IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2017.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *In The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, 1:1646–1654, 2017.

C Liu, H.Y Shum, and W.T Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75(1):115–134, 2007.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *In Proceedings of International Conference on Computer Vision (ICCV)*, 6:3730–3738, 2015.

Wenqi Ren, Jiaolong Yang, Senyou Deng, David Wipf, Xiaochun Cao, and Xin Tong. Face video deblurring via 3d facial priors. *IEEE International Conference on Computer Vision*, 2019.

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Ying Tai, Jian Yang, and Xiaoming Liu. Image superresolution via deep recursive residual network. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1:1637–1645, 2017.

J Thies, M Zollhofer, M Stamminger, C Theobalt, and M Niener. Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

A Van Den Oord, N Kalchbrenner, and K Kavukcuoglu. Pixel recurrent neural networks. *Proceedings of International Conference on International Conference on Machine Learning (ICML)*, pp. 1747–1756, 2016.

X Wang and X Tang. Hallucinating face by eigen transformation. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 35(3):425–434, 2005.

X Wang, K Yu, C Dong, and C.C Loy. Recovering realistic texture in image superresolution by deep spatial feature transform. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

X Yu and F Porikli. Ultra-resolving face images by discriminative generative networks. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 318–333, 2016.

X Yu and F Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3760–3768, 2017a.

X Yu and F Porikli. Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3760–3768, 2017b.

X Yu and F Porikli. Imagining the unimaginable faces by deconvolutional networks. *IEEE Transactions on Image Processing*, 27(6):2747–2761, 2018.

X Yu, B Fernando, B Ghanem, F Porikli, and R Hartley. Face super-resolution guided by facial component heatmaps. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 219–235, 2018a.

X Yu, B Fernando, R Hartley, and F Porikli. Super-resolving very low-resolution face images with supplementary attributes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 908–917, 2018b.

S Zafeiriou, G Trigeorgis, G Chrysos, J Deng, and J Shen. The menpo facial landmark localisation challenge: A step towards the solution. *In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 2116–2125, 2017.

Y Zhang, Y Tian, Y Kong, B Zhong, and Y Fu. Residual dense network for image super-resolution. *In Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 2018a.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. *In Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018b.

W Zhao, R Chellappa, P.J Phillips, and A Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

E Zhou and H Fan. Learning face hallucination in the wild. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3871–3877, 2015.

S Zhu, S Liu, C.C Loy, and X Tang. Deep cascaded bi-network for face hallucination. *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 614–630, 2016.

A Appendix

A.1 Visualization quality of Super-resolution

Semi-Frontal Facial Pose Visualization: For the semi-frontal pose, the SR results of RCAN, RDN and Wavelet-SRNet have a lot of artifacts around facial components (e.g., eyes, teeth, nose, mouth). Fortunately, after incorporating the rendered face priors, it largely avoids the appearance of ghosting artifacts, seen in Figure.8 (ours).

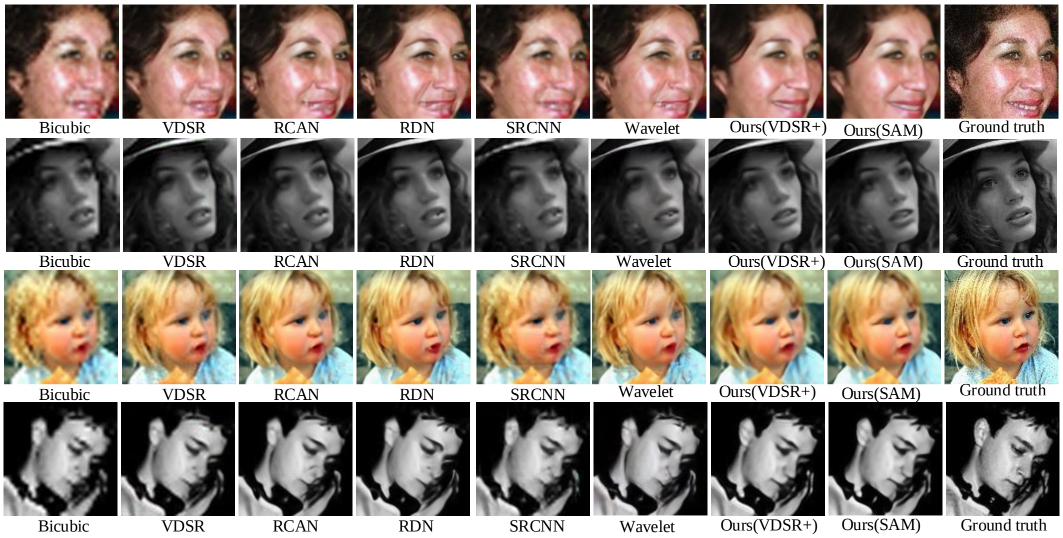


Figure 8: Comparison of state-of-the-art methods on Semi-frontal facial pose:magnification factor ×4 and the input resolution 32×32.

Left Facial Pose Visualization: For the left pose, the high-resolution results of the proposed method perform much better. Ours (VDSR+) results which exploiting the 3D facial priors can grasp the facial structure knowledge and restore the high-resolution facial components (e.g. mouth) much closer to the ground-truth compared with the basic VDSR method without priors.

Right Facial Pose Visualization: For the right pose, the high-resolution results of the proposed method are still the best. Adding the facial structure priors can help network to learn the location of facial components even for the large pose variation.

High Magnification Factor × 8 Visualization: It is still a challenge to generate the sharp super-resolution images for a large magnification factor (×8). The 3D rendered facial priors provide extra facial structure knowledge that are crucial for SR problems. As shown in Figure 12 and 13, the proposed method generates a high visible quality of SR images even for the large magnification factor.

A.2 Ablation study

Learning Curves with Different Ablation Configurations:To verify the effectiveness of 3D facial structure priors, we design the three different configurations (w/o 3D priors, w/o Spatial Attention Mechanism): baseline methods (i.e., VDSR, SRCNN); baseline incorporating 3D facial priors (i.e., VDSR+,SRCNN+); the method using the Spatial Attention Mechanism and 3D priors (our proposed method: +priors and +SAM). The learning curves of each configuration are plotted to show the effectiveness of the each block. The priors are easy to insert into any network without increasing any parameters, but largely improve the accuracy and the convergence of the algorithms shown in Figure 14.

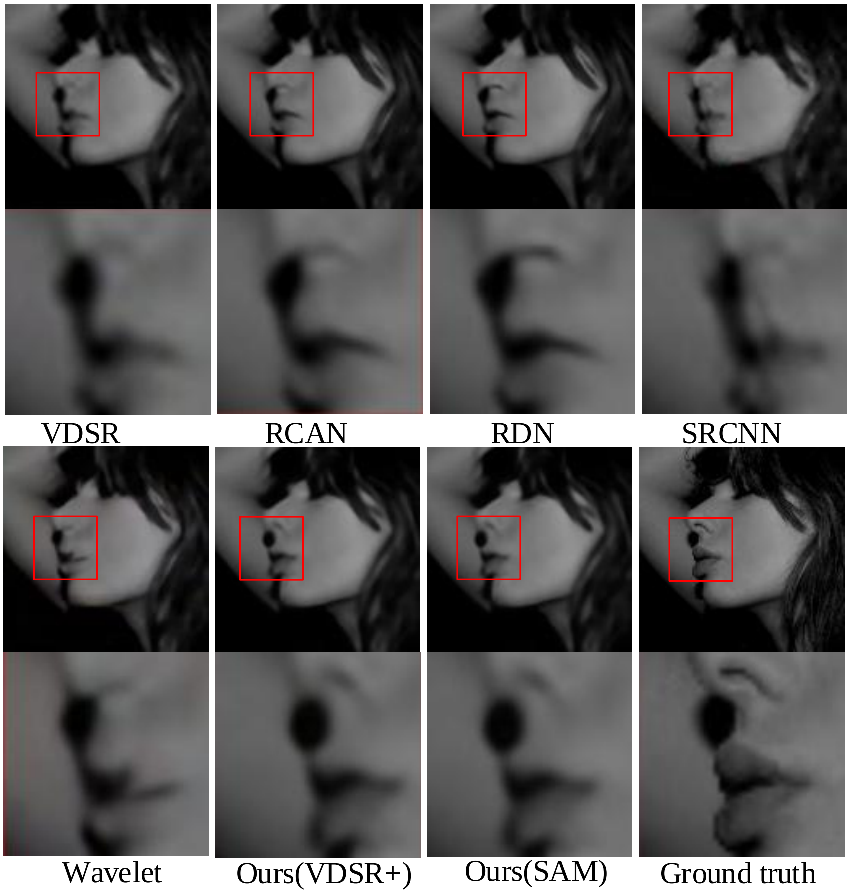


Figure 9: Comparison of state-of-the-art methods on left facial pose:magnification factor ×4 and the input resolution 32×32.

Quantitative Results with Different Ablation Configurations: As shown in Table 3, each block boosts the accuracy of baseline algorithms: the average performance improvement stemming from 3D facial priors and from Spatial Attention Mechanism are 1.6db and 0.57db, respectively.

Table 3: Quantitative results (PSNR/SSIM) with different ablation configurations. Priors denotes the 3D rendered structure priors; SAM denotes the Spatial Attention Mechanism.

Factor	SRCNN	VDSR	SRCNN+prior	VDSR+prior	ours(+prior+SAM)
4scale	27.57/0.8452	28.13/0.8554	28.66/0.8501	29.29/0.8727	**29.69/0.8817**
8scale	22.51/0.6659	22.76/0.6618	24.18/0.6959	24.66/0.7127	**25.39/0.7551**

Qualitative Evaluation with different ablation configurations: The baseline incorporated with the facial rendered priors tends to avoid some artifacts around the key facial components and generate more sharp edges compared with the basic baseline method without the facial priors. By adding the Spatial Attention Mechanism, it could help the network better exploit the priors and is easier to generate more sharp facial structures, shown in Figure 15.

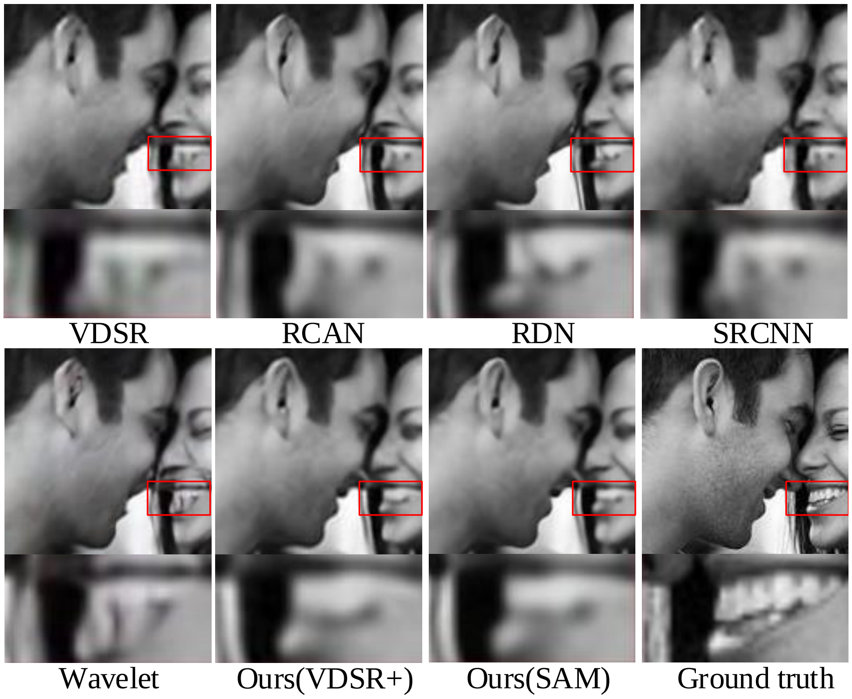


Figure 10: Comparison of state-of-the-art methods on left facial pose:magnification factor ×4 and the input resolution 32×32.

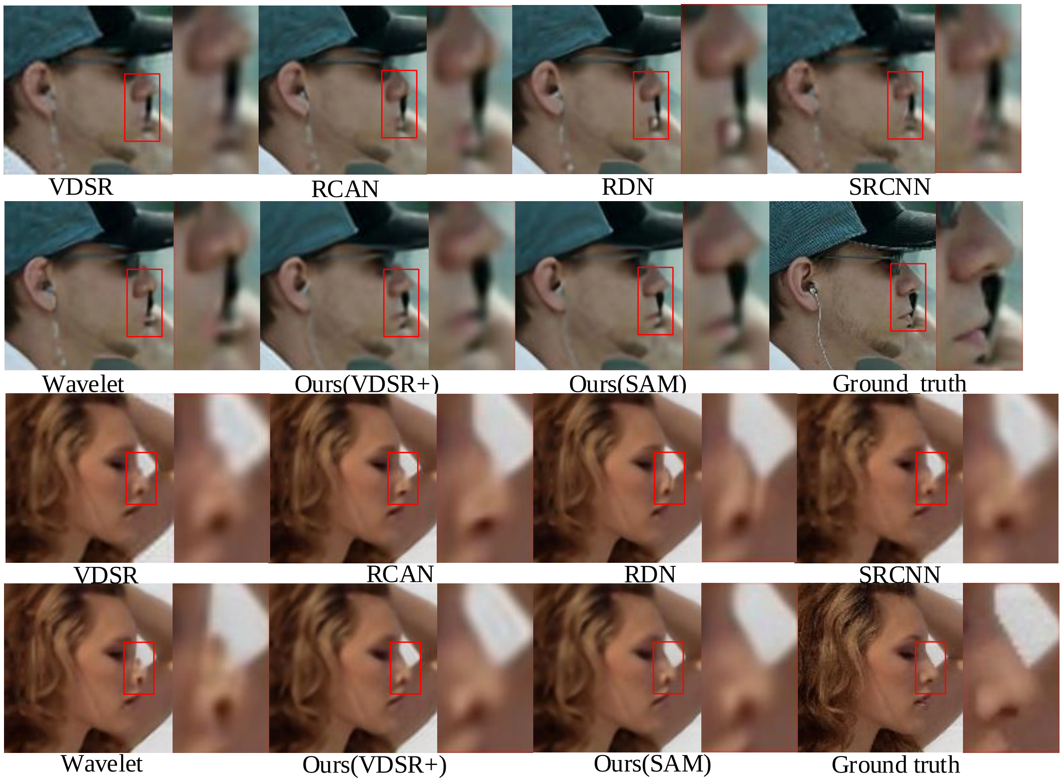


Figure 11: Comparison of state-of-the-art methods on right facial pose:magnification factor ×4 and the input resolution 32×32.

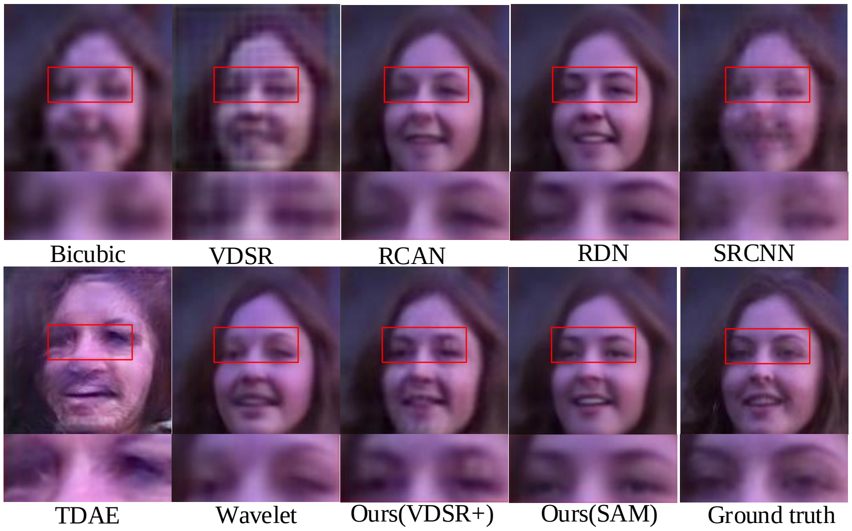


Figure 12: Comparison of state-of-the-art methods: magnification factor ×8 and the input resolution 16×16.

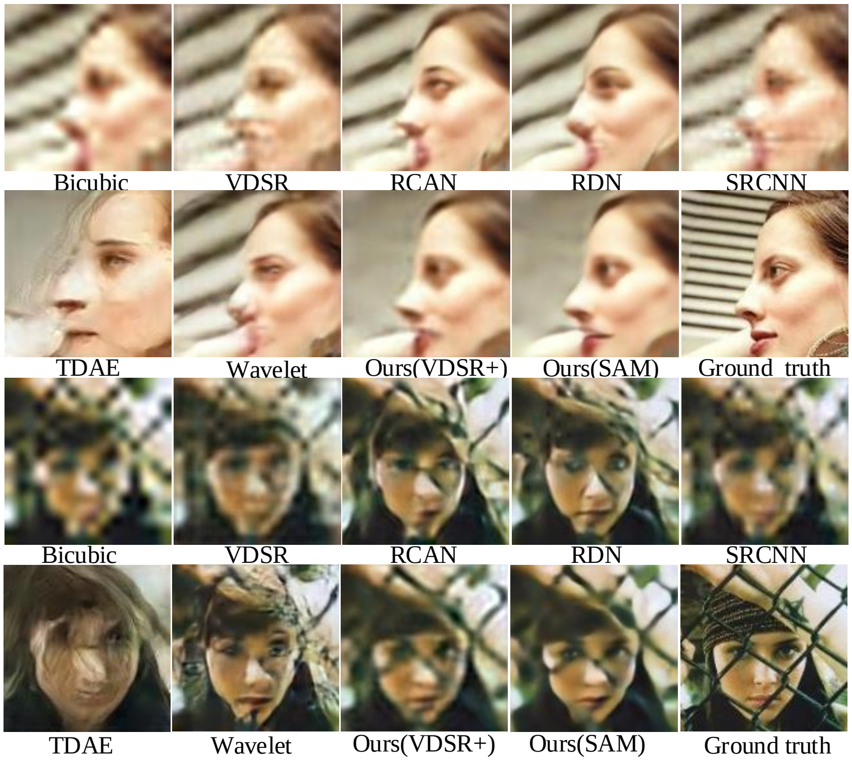
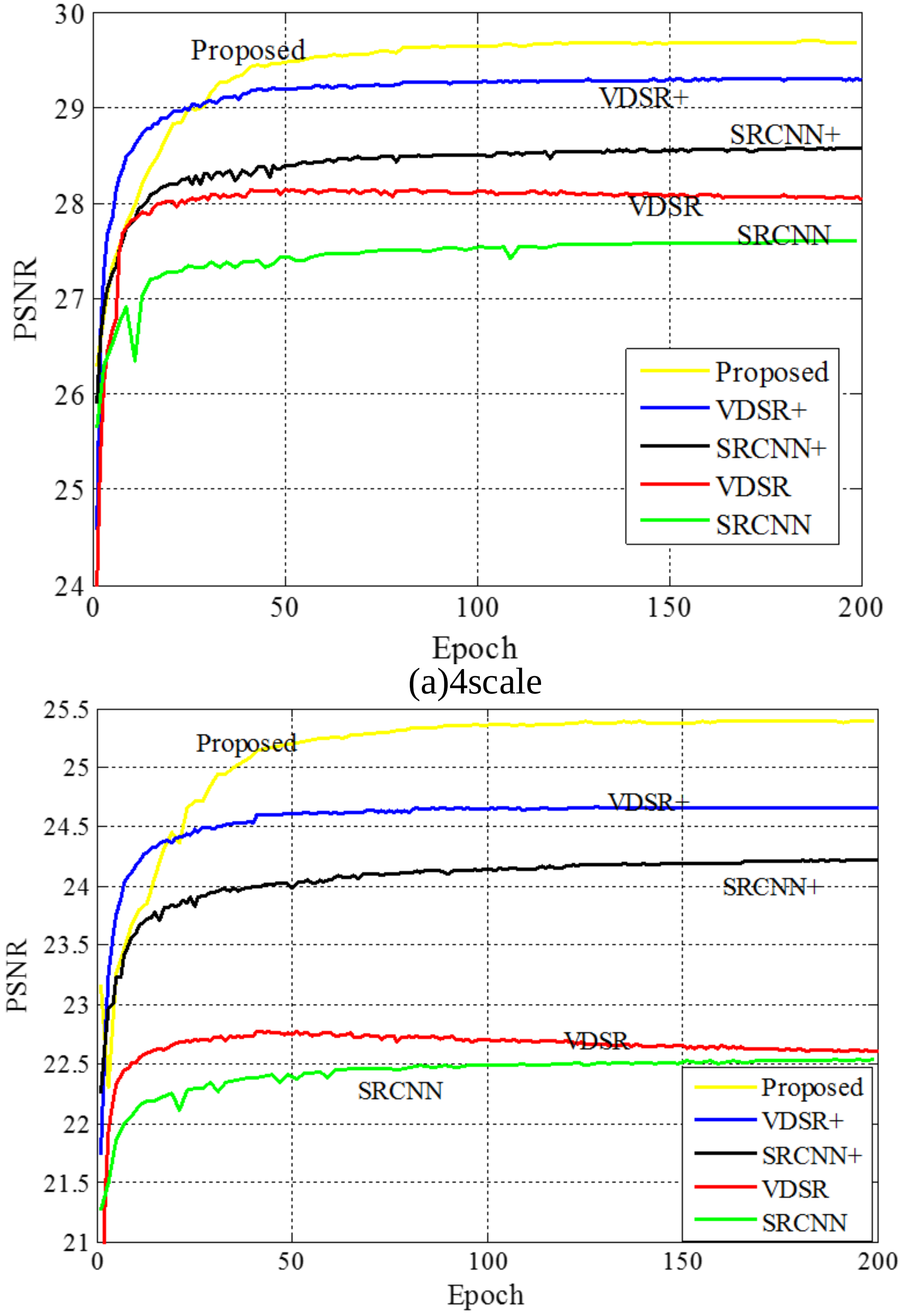


Figure 13: Comparison of state-of-the-art methods: magnification factor ×8 and the input resolution 16×16.



(a)4scale

(b)8scale

Figure 14: Learning curves with different ablation configurations: baseline + denotes the baseline (SRCNN and VDSR) incorporating the 3D facial priors.

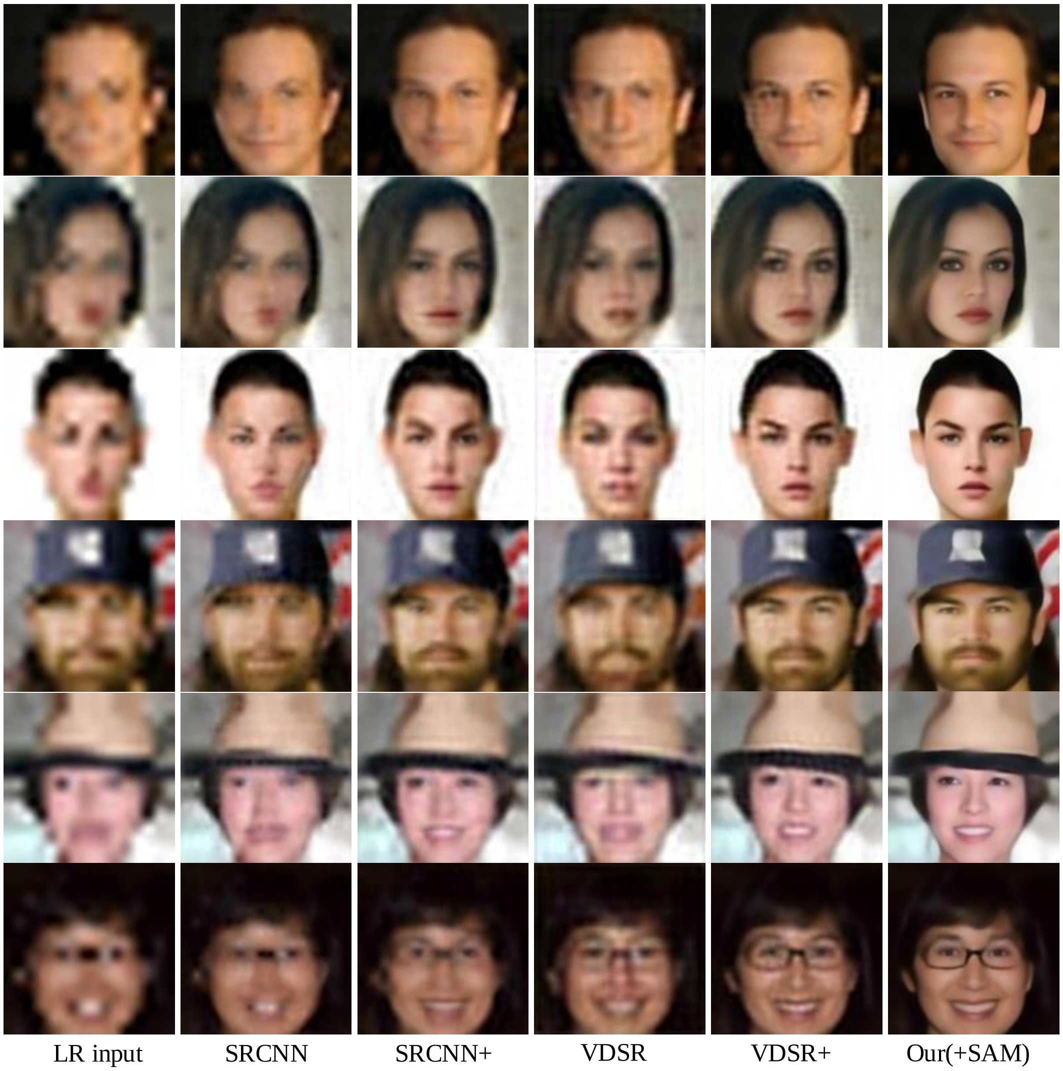


Figure 15: Qualitative evaluation with different ablation configurations: baseline + denotes the baseline (SRCNN and VDSR) incorporating the 3D facial priors; +SAM means adding the Spatial Attention Mechanism.