Face Super-resolution Guided by 3D Facial Priors Supplementary Material

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

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

1 Overview

In this supplementary document, we present additional results to complement the main paper. Firstly, we provide the detailed configurations and parameters of the proposed methods. Secondly, the quantitative ablation results, the result on real world images and the comparisons of model size and performance are added to verify the superiority of the facial rendered priors and proposed Spatial Attention Mechanism. Thirdly, the reconstructed rendered faces of large pose variations and occlusion are also presented here to show the stability and superior performance of the face reconstruction. Lastly, more qualitative comparisons with the state-of-the-art algorithms are added in the supplementary.

2 Results on Real-World Images

Quantitative and qualitative results: For real-world LR images, we provide the quantitative analysis on 500 LR faces from the WiderFace (x4) dataset using two no-reference image criteria (Perception Image Quality (PIQE) and Naturalness Image Quality (NIQE)). The results are sorted in Table 1 (the higher of criteria, the worse of the image quality). These results demonstrate that the proposed 3D priors boost the performance even in the real-world conditions by capturing the good visual quality of facial components and restoring large pose variations and partial occlusions. We add some visual comparisons shown in Figure 1 to illustrate the effectiveness of our 3D priors on real-world faces.

Position deviation of rendered 3D faces: 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 as: 0 pixel [29.69dB]; 1 pixel [29.53dB]; 2 pixels [29.32dB]; 5 pixels [29.08dB]; 10 pixels [29.05dB]; 20 pixels [29.03dB]. The performance of the inaccurate face priors also behaves better even for 10 pixels-offset than without using the 3D facial priors (28.92dB).

[*] indicates the corresponding author.

[†] these authors contributed equally to this work.

Table 1. 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**

3 Network Architectures

The proposed network consists of four branches: 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. Table 2 lists the detailed configuration of the proposed method for the 4× scale factor.

4 Ablation Studies

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.

Advantage of 3D facial structure priors: To verify the advantage of 3D facial structure priors in terms of the convergence and accuracy, three different configurations are designed: basic methods (*i.e.,* SRCNN and VDSR); basic methods incorporating 3D facial priors (*i.e.,* SRCNN+3D and VDSR+3D); the proposed method using the Spatial Attention Module and 3D priors (SAM3D). The validation accuracy curve of each configuration along the epochs is plotted to show the effectiveness of each block. The priors are easy to insert into any network. They only marginally increase the number of parameters, but significantly improve the accuracy and convergence of the algorithms as shown in Figure 2.

5 Model Size and Running Time

We evaluate the proposed method and state-of-the-art super-resolution methods on the same server with an Intel Xeon W-2123 CPU and an NVIDIA TITAN X GPU. Figure 3 (a) shows comparisons of model size and PSNR performance on the CelebA dataset with ×8 magnification factor. Our proposed SAM3D as well as VDSR+3D, embedded with 3D priors, is more lightweight while still achieving the best performance even compared with the recent state-of-the-art SR methods (*e.g.,* RCAN and RDN) and face priors based SR methods (*e.g.,* FSRNet and PSRFAN). In addition, as shown in Figure 3 (b), our proposed method Spatial

Table 2. Detailed configurarion of the proposed network for the 4 × scale factor.

Block name	Input	Output	(Size,Channel,Stride)	Repeated Block Number
	LR images	Conv0	(3,64,1)	1
	Priors	CondNet	(1,128,1)/ReLU;	
	CondNet	CondNet1	(1,128,1)/ReLU;	
	CondNet1	CondNet2	(1,128,1)/ReLU;	
	CondNet2	CondNet3	(1,32,1)	
	Conv0 and CondNet3	SFT_Layer_1	(1,32,1)/Relu (1,64,1) (1,32,1)/Relu (1,64,1)	
	SFT_Layer_1	Conv1	(3,64,1)	
	Conv1	SFT_Layer_2	(1,32,1)/Relu (1,64,1) (1,32,1)/Relu (1,64,1)	
	SFT_Layer_2	Conv2_8	(3,64,1)	
	Conv2_8	SFT_Layer_3	(1,32,1)/Relu (1,64,1) (1,32,1)/Relu (1,64,1)	
	SFT_Layer_3	Conv3	(3,64,1)/Relu (3,64,1)	
	Conv3	CA_layer32	(1,64/16,1)/Relu (1,64,1)/Relu Sigmoid	
	CA_layer32	Conv4	(3,64,1)	
	Conv4	Up1	upsampling(x2)	
	Up1	Conv5	(3,64,1)/Relu	
	Conv5	Up2	upsampling(x2)	
	Up2	Conv6	(3,64,1)/Relu	
	Conv6	Conv7	(3,64,1)/Relu	
	Conv7	Conv8	(3,3,1)	

Attention Module incorporating 3D priors (SAM3D) and VDSR+3D improves PSNR for scale factor ×8 on the dataset CelebA in comparison to the state-of-the-art methods. Our methods outperform the other approaches by a large margin while maintaining comparable running times to face SR methods with 2D priors. Our test running time includes the time required for ResNet-50.

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**

6 Coefficients Feature Transform

The output of ResNet-50 is the representative feature vector of $\boldsymbol{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\rho}) \in \mathbb{R}^{239}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{80}, \boldsymbol{\beta} \in \mathbb{R}^{64}, \boldsymbol{\delta} \in \mathbb{R}^{80}, \boldsymbol{\gamma} \in \mathbb{R}^{9}$, and $\boldsymbol{\rho} \in \mathbb{R}^{6}$ represent the identity, facial expression, texture, illumination, and face pose, respectively. The feature transformation procedure is described as follows shown in Figure 4: firstly, the coefficients of identity, expression, texture, and the element-concatenation of illumination and face pose ($\boldsymbol{\gamma} + \boldsymbol{\rho}$) are reshaped to four matrices by setting extra elements to zeros. Afterwards, it is expanded to the same size as the LR images (16×16 or 32×32) by zero-padding, and then scaled to the interval [0,1].

7 Rendered Face Generation

In order to evaluate the quality of rendered face generation, we present plenty of rendered face images to judge whether the rendered faces grasp the gender and pose variations priors. Given the low-resolution images, the generated 3D face rendered reconstructions of the gender (male and female) are shown in Figure 5. The 3D face rendered reconstructions of the large pose variations are shown in Figure 6. 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 color, and further embed pose variations into the super-resolution networks which improve the accuracy and stability in face images with large pose variations. The algorithm is stable and robust to well-reconstruct the rendered faces even for the face images which are partly occluded by glasses, hairs, etc., as shown in Figure 7.

8 Visualization Quality of Super-Resolution

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 is crucial for SR problems. As shown in Figures 8-12, the proposed method generates a high visible quality of SR images even for the large magnification factor.

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 and mouth). Fortunately, after incorporating the rendered face priors, it largely avoids the appearance of ghosting artifacts, seen in Figures 13-17.

Left Facial Pose Visualization: For the left pose, the high-resolution results of the proposed method perform much better. Ours (VDSR+) which exploits 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 shown in Figures 18-21.

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 the network learn the location of facial components even for the large pose variations shown in Figures 22-26.

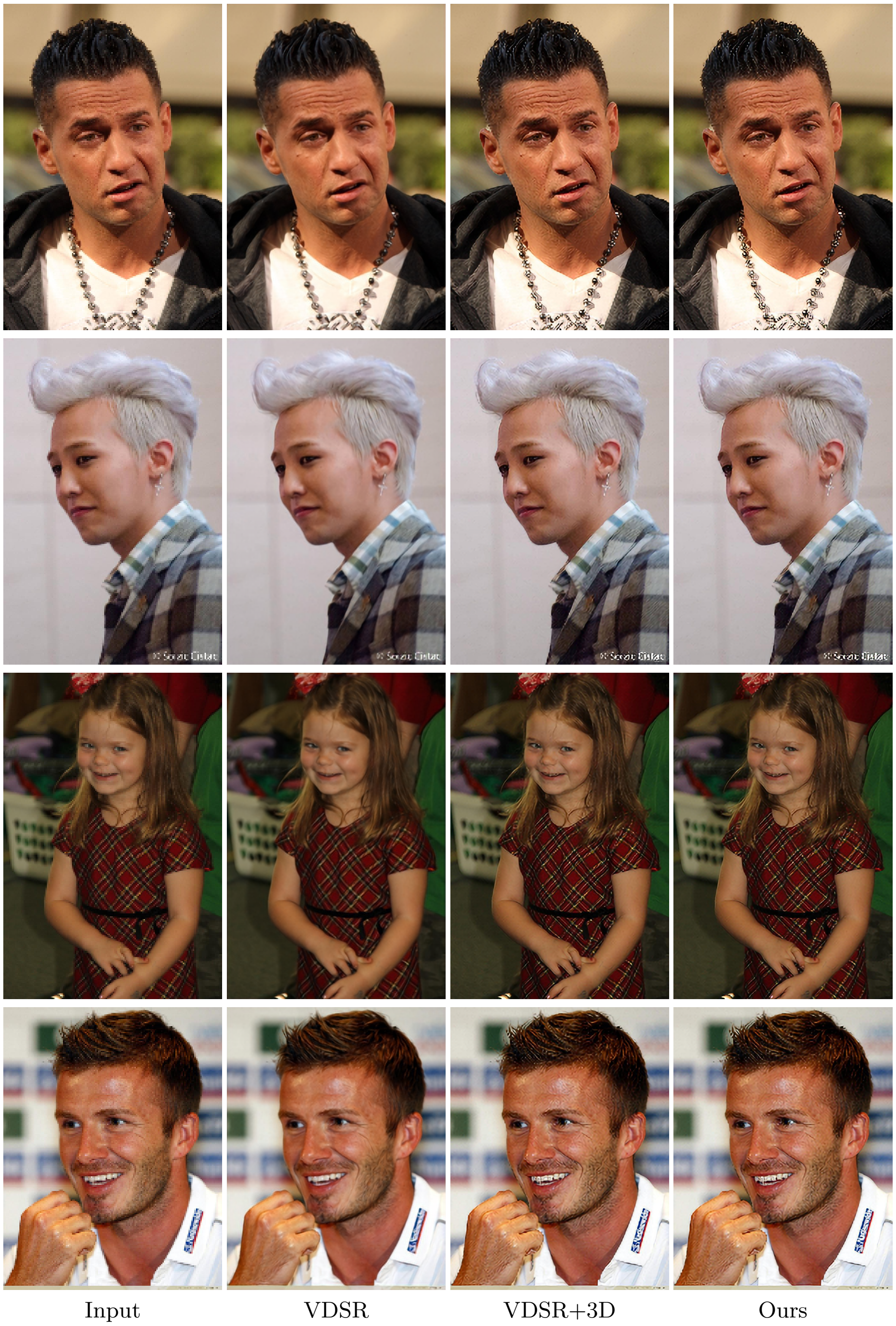
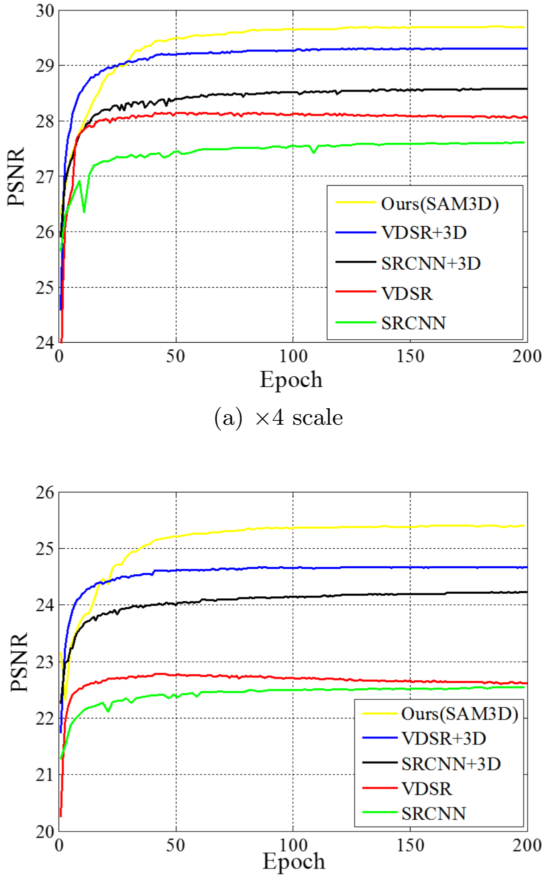


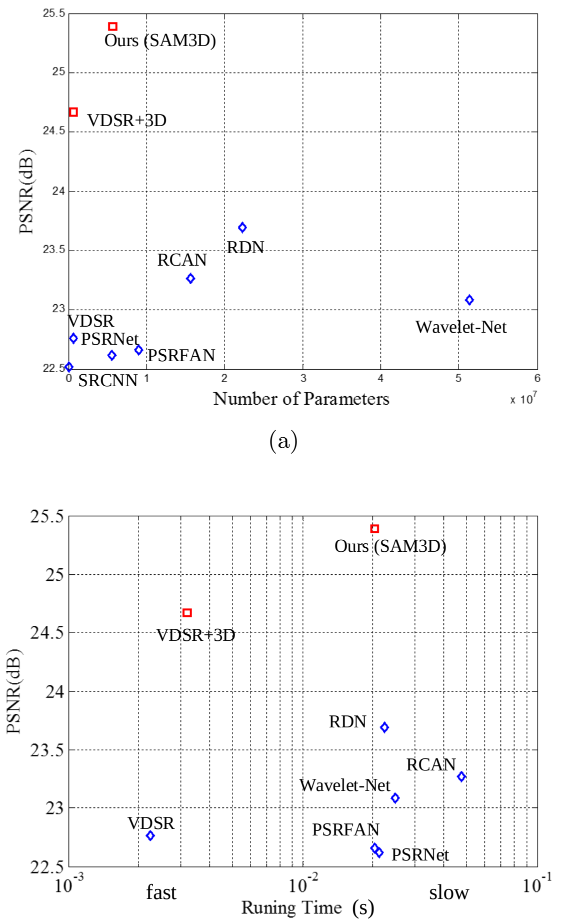
Fig. 1. Visual comparison of real world images with state-of-the-art methods (×4). Best viewed by zooming in the screen.



(a) ×4 scale

(b) ×8 scale

Fig. 2. Test accuracy curves with different configurations along the training epochs.



(a)

(b)

Fig. 3. (a) Performance vs. number of parameters. Results are evaluated on the CelebA dataset: ×8 scale. (b) Average running time on CelebA.

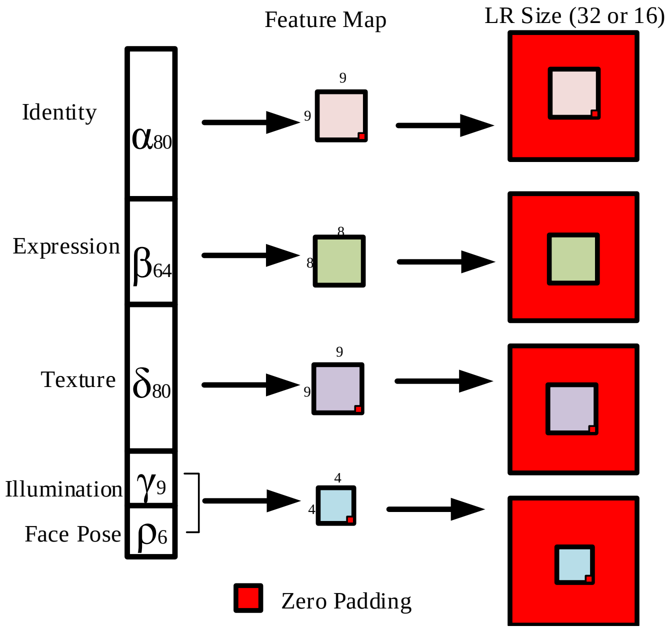
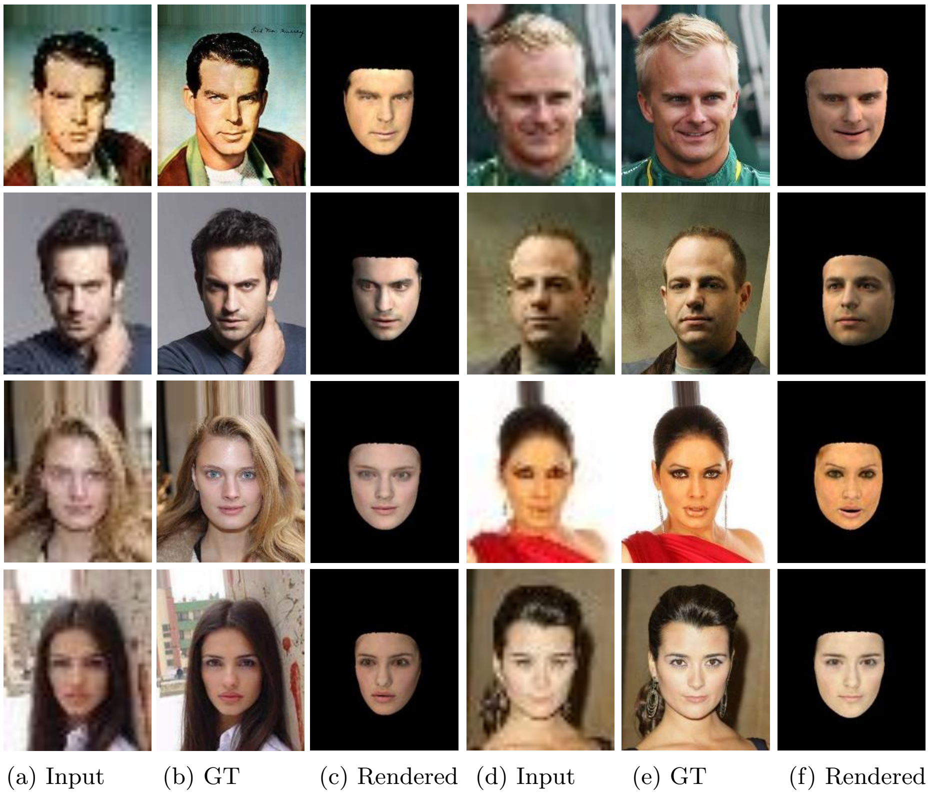
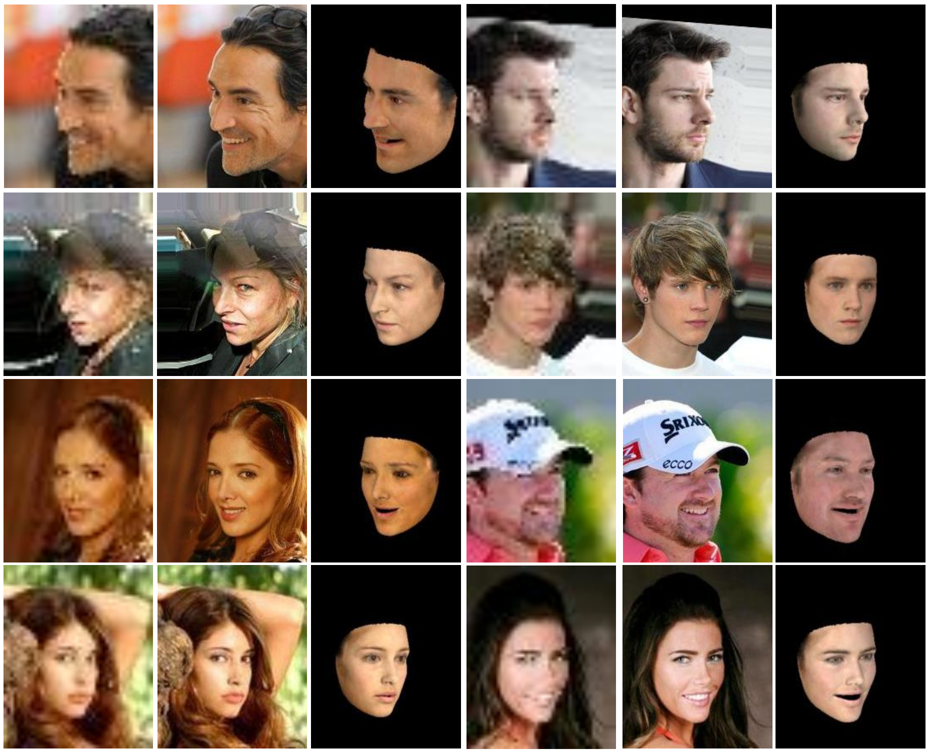


Fig. 4. Coefficients (the identity, facial expression, texture, illumination, and face pose) Feature Transform: Coefficients are reshaped, zero-padded, expanded and scaled into [0,1].



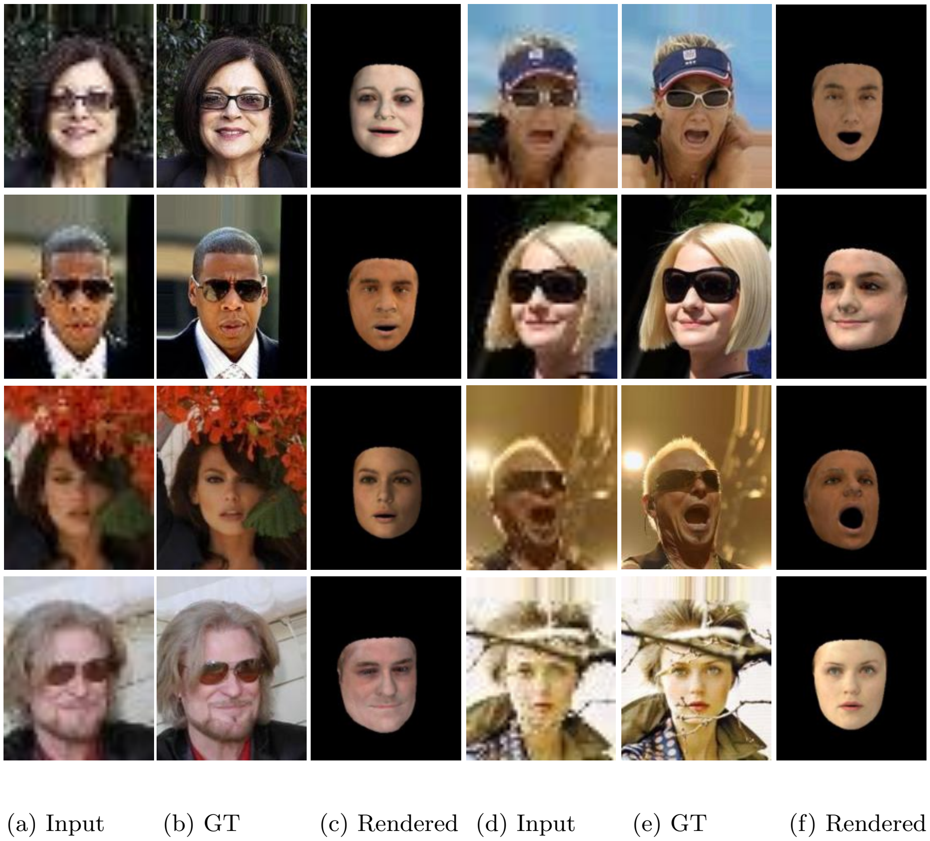
(a) Input (b) GT (c) Rendered (d) Input (e) GT (f) Rendered

Fig. 5. The rendered prior by our method. (a) and (d) inputs. (b) and (e) ground-truths. (c) and (f) our rendered face structures. The reconstructed facial structures provide clear gender knowledge (male and female).



(a) Input (b) GT (c) Rendered (d) Input (e) GT (f) Rendered

Fig. 6. The rendered prior by our method. (a) and (d) inputs. (b) and (e) groundtruths. (c) and (f) our rendered face structures. 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)



(a) Input (b) GT (c) Rendered (d) Input (e) GT (f) Rendered

Fig. 7. The rendered prior by our method. (a) and (d) inputs. (b) and (e) ground-truths. (c) and (f) our rendered face structures. Our method can recover the occluded parts of face images and grasp the clear spatial location.

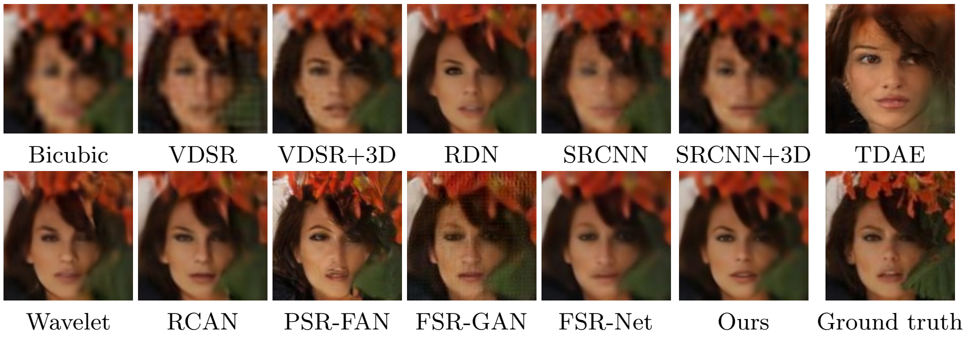


Fig. 8. Visual comparison with state-of-the-art methods (×8). The results by the proposed method have fewer visual artifacts and sharper facial structures. Best viewed by zooming in the screen.

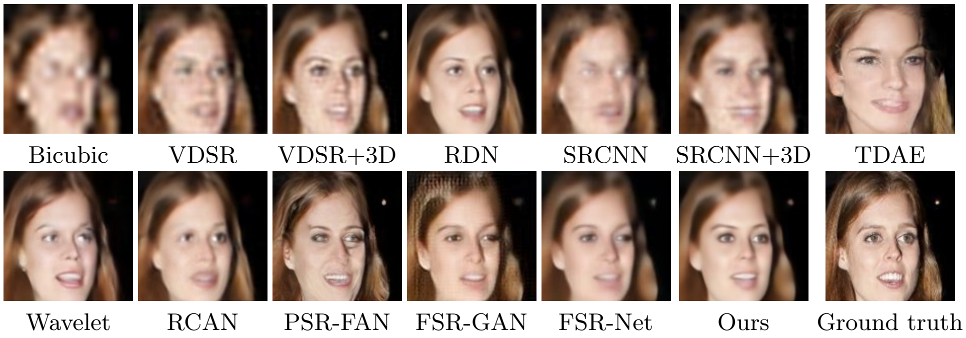


Fig. 9. Visual comparison with state-of-the-art methods (×8). The results by the proposed method have fewer visual artifacts and sharper facial structures. Best viewed by zooming in the screen.

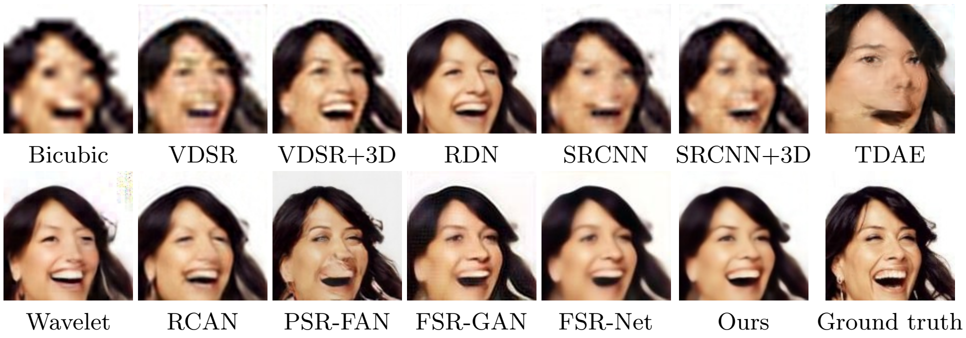


Fig. 10. Visual comparison with state-of-the-art methods (×8). The results by the proposed method have fewer visual artifacts and sharper facial structures. Best viewed by zooming in the screen.

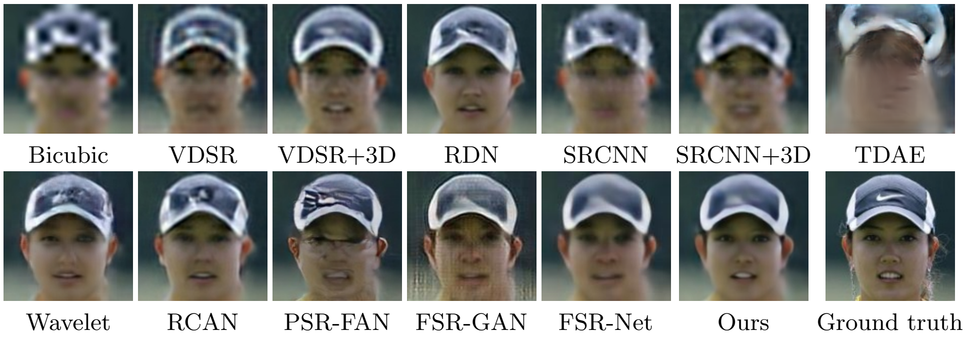


Fig. 11. Visual comparison with state-of-the-art methods (×8). The results by the proposed method have fewer visual artifacts and sharper facial structures. Best viewed by zooming in the screen.

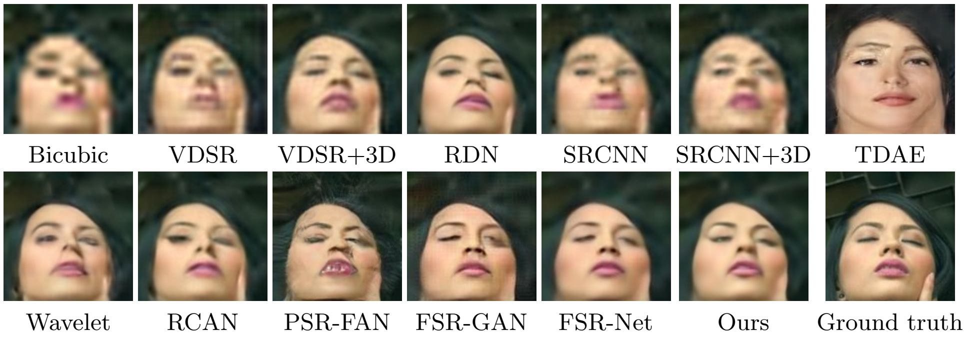


Fig. 12. Visual comparison with state-of-the-art methods (×8). The results by the proposed method have fewer visual artifacts and sharper facial structures. Best viewed by zooming in the screen.

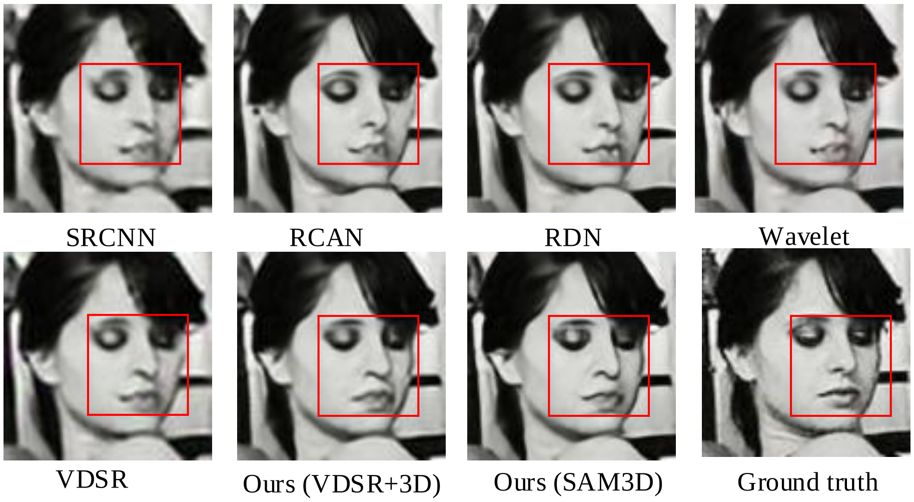


Fig. 13. Comparison of state-of-the-art methods on **semi-frontal** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

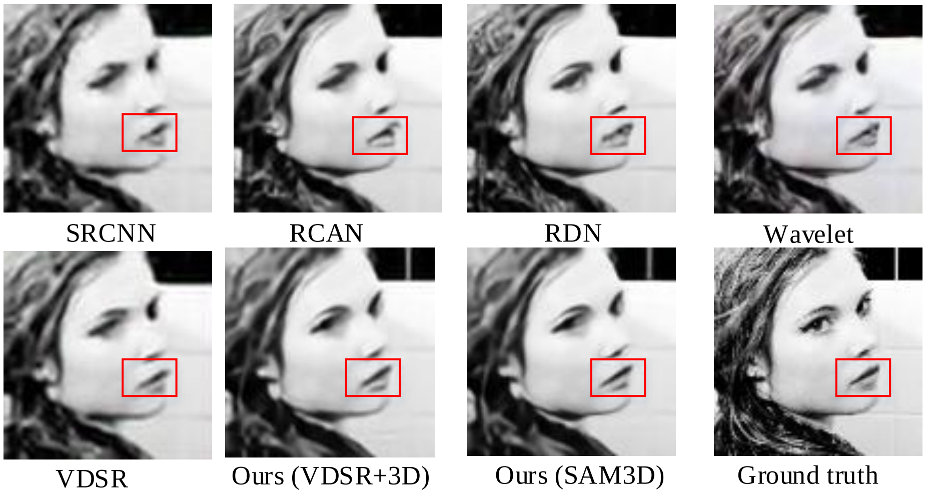


Fig. 14. Comparison of state-of-the-art methods on **semi-frontal** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

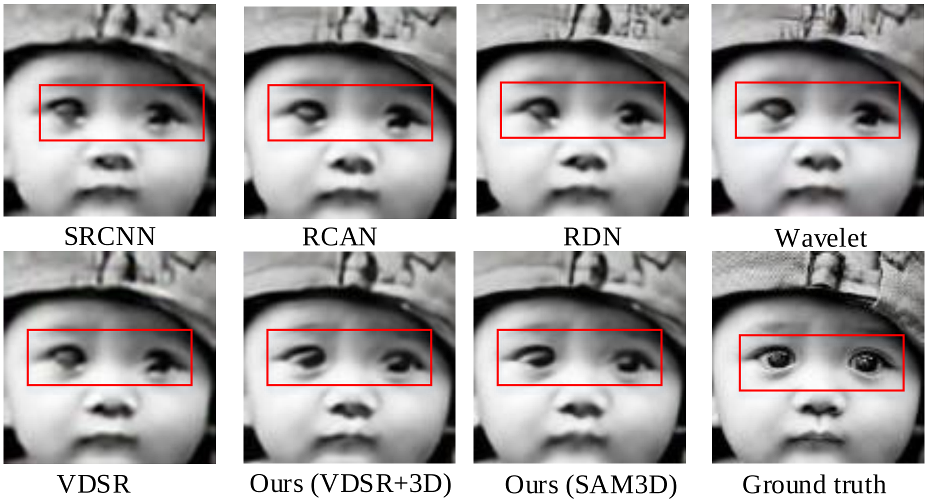
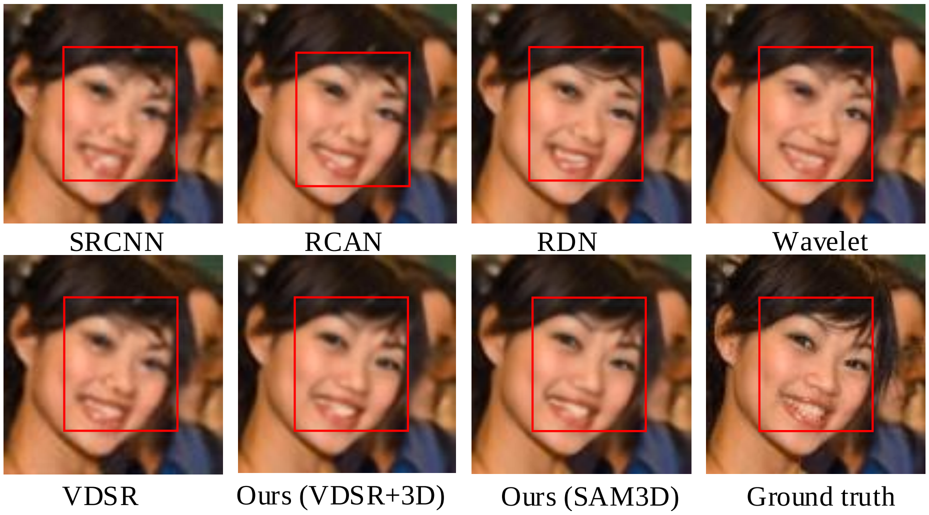


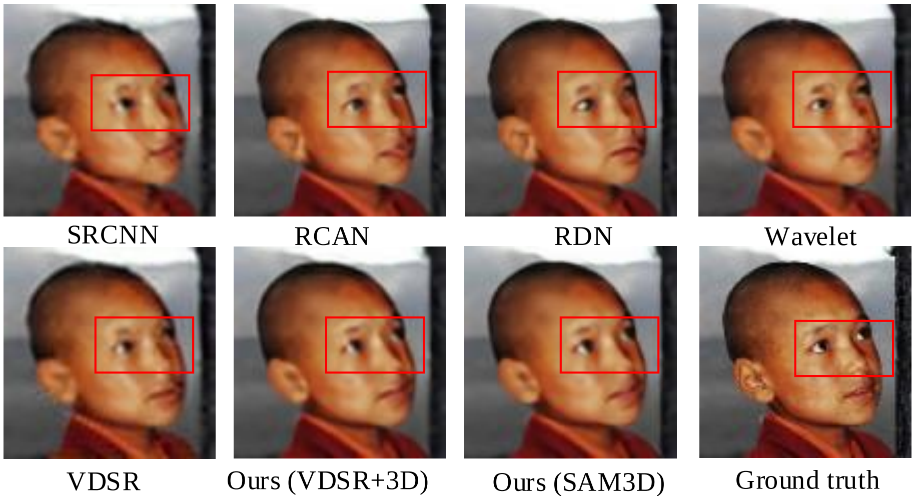
Fig. 15. Comparison of state-of-the-art methods on **semi-frontal** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.



SRCNN RCAN RDN Wavelet

VDSR Ours (VDSR+3D) Ours (SAM3D) Ground truth

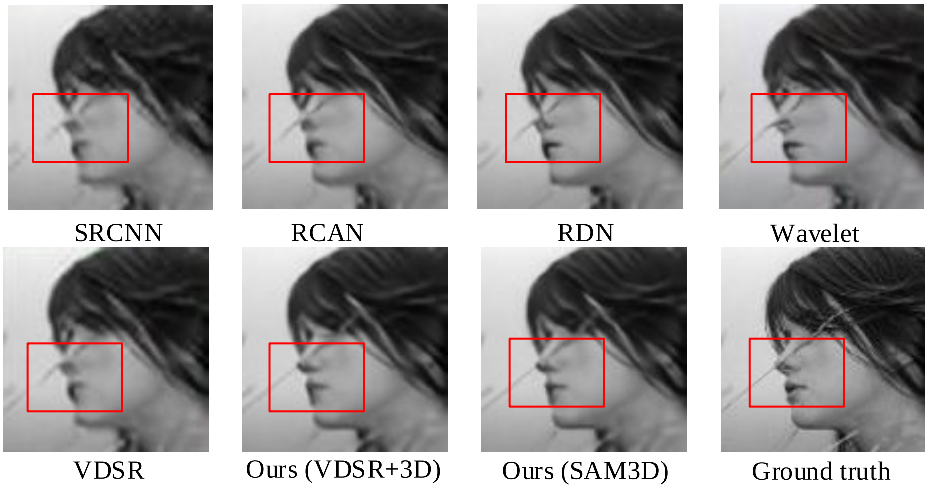
Fig. 16. Comparison of state-of-the-art methods on **semi-frontal** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.



SRCNN RCAN RDN Wavelet

VDSR Ours (VDSR+3D) Ours (SAM3D) Ground truth

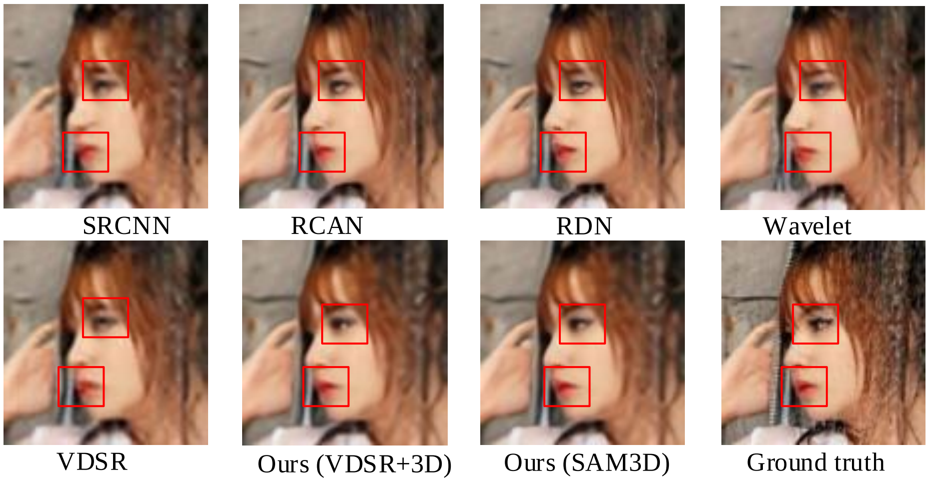
Fig. 17. Comparison of state-of-the-art methods on **semi-frontal** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.



SRCNN RCAN RDN Wavelet

VDSR Ours (VDSR+3D) Ours (SAM3D) Ground truth

Fig. 18. Comparison of state-of-the-art methods on **left** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.



SRCNN RCAN RDN Wavelet

VDSR Ours (VDSR+3D) Ours (SAM3D) Ground truth

Fig. 19. Comparison of state-of-the-art methods on **left** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

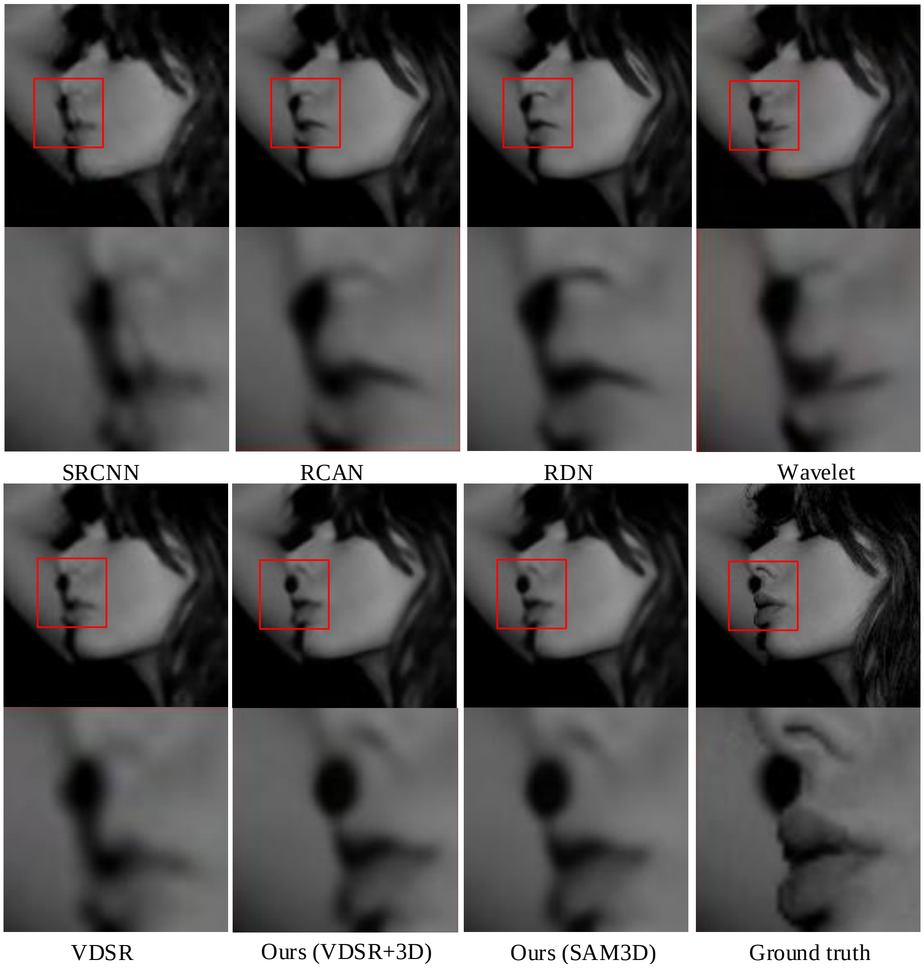


Fig. 20. Comparison of state-of-the-art methods on **left** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

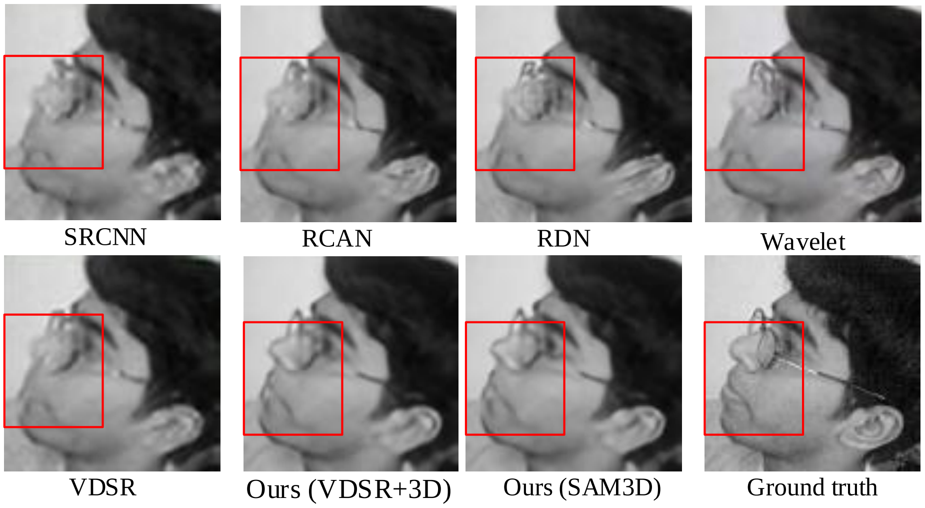


Fig. 21. Comparison of state-of-the-art methods on **left** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

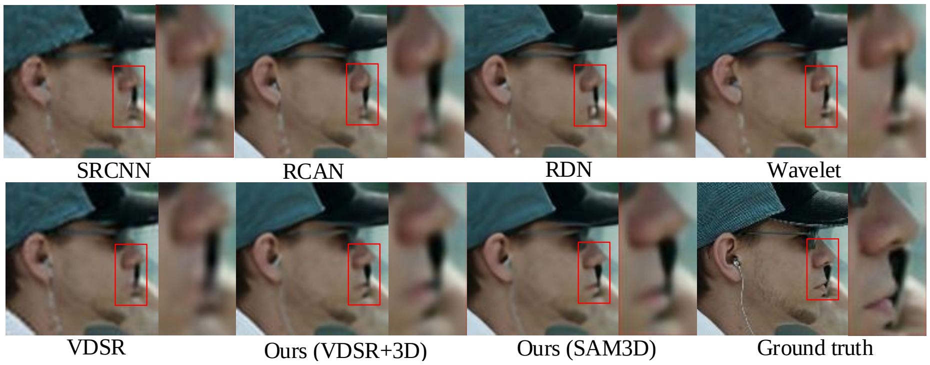


Fig. 22. Comparison of state-of-the-art methods on **right** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

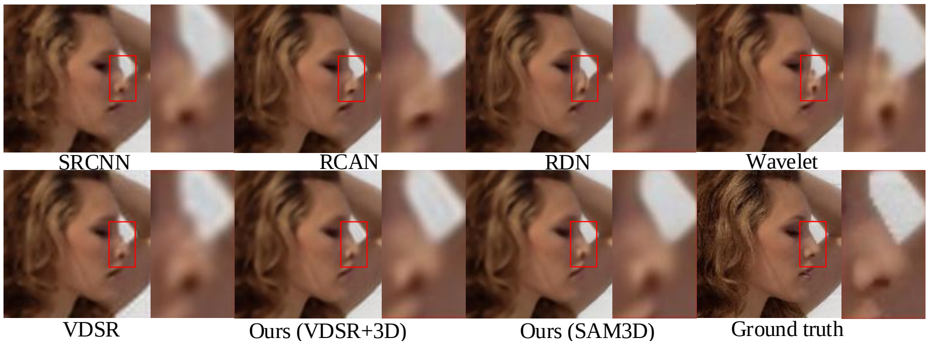


Fig. 23. Comparison of state-of-the-art methods on **right** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

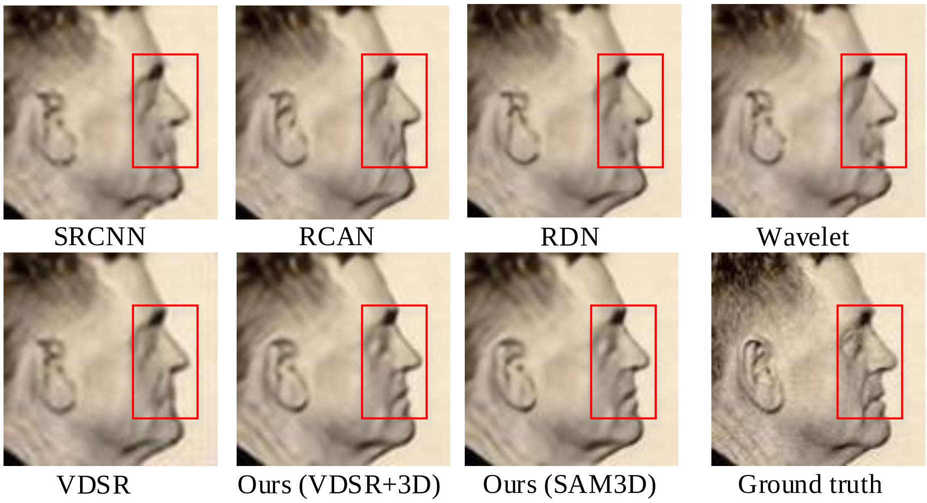


Fig. 24. Comparison of state-of-the-art methods on **right** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

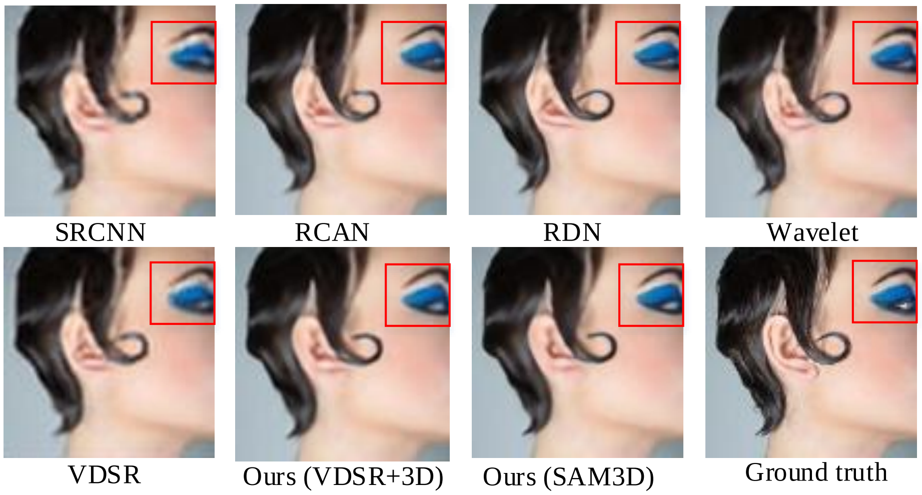


Fig. 25. Comparison of state-of-the-art methods on **right** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.

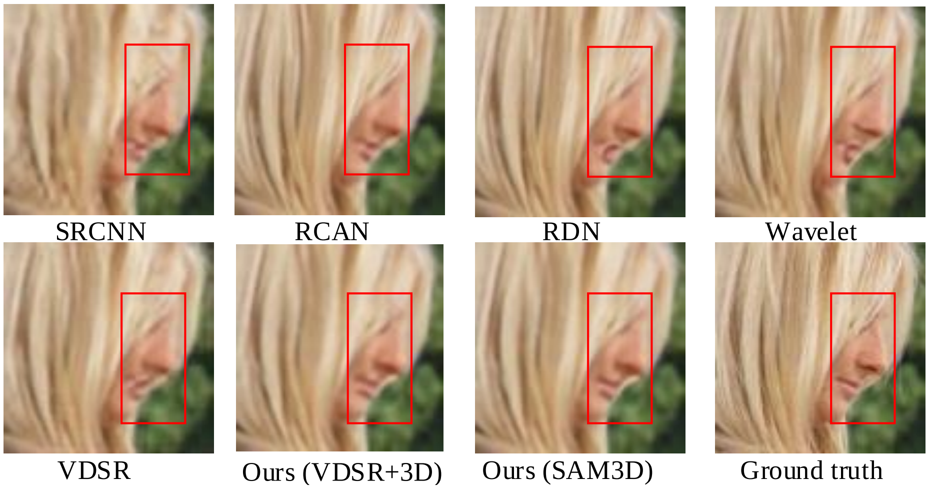


Fig. 26. Comparison of state-of-the-art methods on **right** facial pose:magnification factor ×4 and the input resolution 32×32. Best viewed by zooming in the screen.