Face Video Deblurring using 3D Facial Priors Supplementary Material

Wenqi Ren[1†*], Jiaolong Yang[2†], Senyou Deng[1], David Wipf[2], Xiaochun Cao[1,3‡], and Xin Tong[2]
[1]SKLOIS, IIE, CAS [2]Microsoft Research Asia [3]University of Chinese Academy of Sciences

Overview

In this supplemental material, we given the split of training and testing videos of the paper in Section 1, and validate the performance of the proposed 3D face rendering branch and analyze the effectiveness of the proposed face rendering loss $\mathcal{L}_r$ in Section 2. In addition, we show more visual comparisons in Section 5.

1. Split of the training and testing videos

In this section, we give the list of the videos used in our training and testing data selected from the 300VW dataset [4]. We excluded videos that are already blurred or of low resolutions (thus inappropriate for synthesis) from the 114 videos in the 300VW dataset. All the names of the used videos are listed below:

Training data: vd001-vd003, vd005, vd009-vd024, vd027-vd033, vd035-vd041, vd043, vd045-vd050, vd052, vd053, vd056, vd058-vd062, vd064-vd069, vd072, vd076-vd080, vd082-vd084, vd087, vd089-vd091, vd093-vd099, vd103-vd107, vd110, vd113, vd114.

Testing data: vd006, vd007, vd025, vd026, vd034, vd051, vd057, vd070, vd092.

2. 3D Face Rendering Branch

In this section, we evaluate the performance of the proposed 3D face rendering branch and the proposed face rendering loss function $\mathcal{L}_r$ in (3) on the 300VM dataset. As shown in Figure 2(b), without using the proposed face rendering loss, the 3D rendering branch tends to generate some facial components that differ from the ground-truths in Figure 2(d) due to the motion blur. In contrast, with the proposed rendering loss $\mathcal{L}_r$, our proposed 3D face rendering branch is more robust to motion blur and recovers detailed facial structures (e.g., shape, eyebrows, eyes, noses, and mouths) as shown in Figure 2(c). In addition, the proposed rendering branch is robust to various facial poses, expressions, and even occlusions as shown in the first row of Figure 2(c).

3. Effects of Multiple Frames

Inspired by the work of Deep Video Deblurring (DVD) [6] which demonstrates that simply stacking neighboring frames without any alignment performs better than the single image based method, we also perform an early fusion of neighboring frames by concatenating five images in the input layer. Multi-image input will provide not only motion cues but also complementary information across frames[1], thus benefiting the deblur process.

To demonstrate the effectiveness of multiple frames in the input layer, we train a network with the same architecture in the proposed algorithm, but only feed the central frame into the network instead of inputing a stack of neighboring frames. In the testing stage, we deblur face videos frame-by-frame. Quantitative results are shown in Table 1. It can be seen that using neighboring frames significantly improves the quality of results.

Table 1. Quantitative PSNR and SSIM results on the 9 synthetic testing data from the 300VW dataset.

Su et al. [6]	Tao et al. [7]	Our single	Our multiple
34.40/0.9218	36.36/0.9784	37.20/0.9841	**37.70/0.9849**

4. Failure Cases

When our 3D reconstruction branch fails to recover authentic structures in challenging situations (e.g., severe blur or occlusion), our deblurring result will also degrade. Figure 1 shows one example where our deblurred image is not satisfactory especially at the eye regions due to the inaccurate structural information on the rendered image.

5. Quantitative Results on Benchmark

In this supplemental material, we provide more examples to evaluate the proposed method in Figure 3-6. We compare

*This work was done while W. Ren was a visiting scholar at MSRA.
†Equal contributions
‡Corresponding author

[1]Due to temporally-varying motion velocity, the same image content may appear blurry on some frames but clearer on others.

(a) Blurred frame (b) Su et al. [6] (c) Our rendered (d) Our deblurred

Figure 1. A failure case. Since the rendered face (c) contains misguided structural information at the eye regions, the proposed method fails to reconstruct a sharp image at the eye regions.

the proposed algorithm with following six methods: video deblurring [6, 1], natural image deblurring [2, 7], and face image deblurring [3, 5] methods. Note that for fair comparisons, we fine-tuned the image deblurring network [7] with further 50,000 iterations and retrained the video deblurring network [6] using the same training data in this work.

References

[1] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, 2015. 2, 4, 5, 6, 7

[2] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2, 4, 5, 6, 7

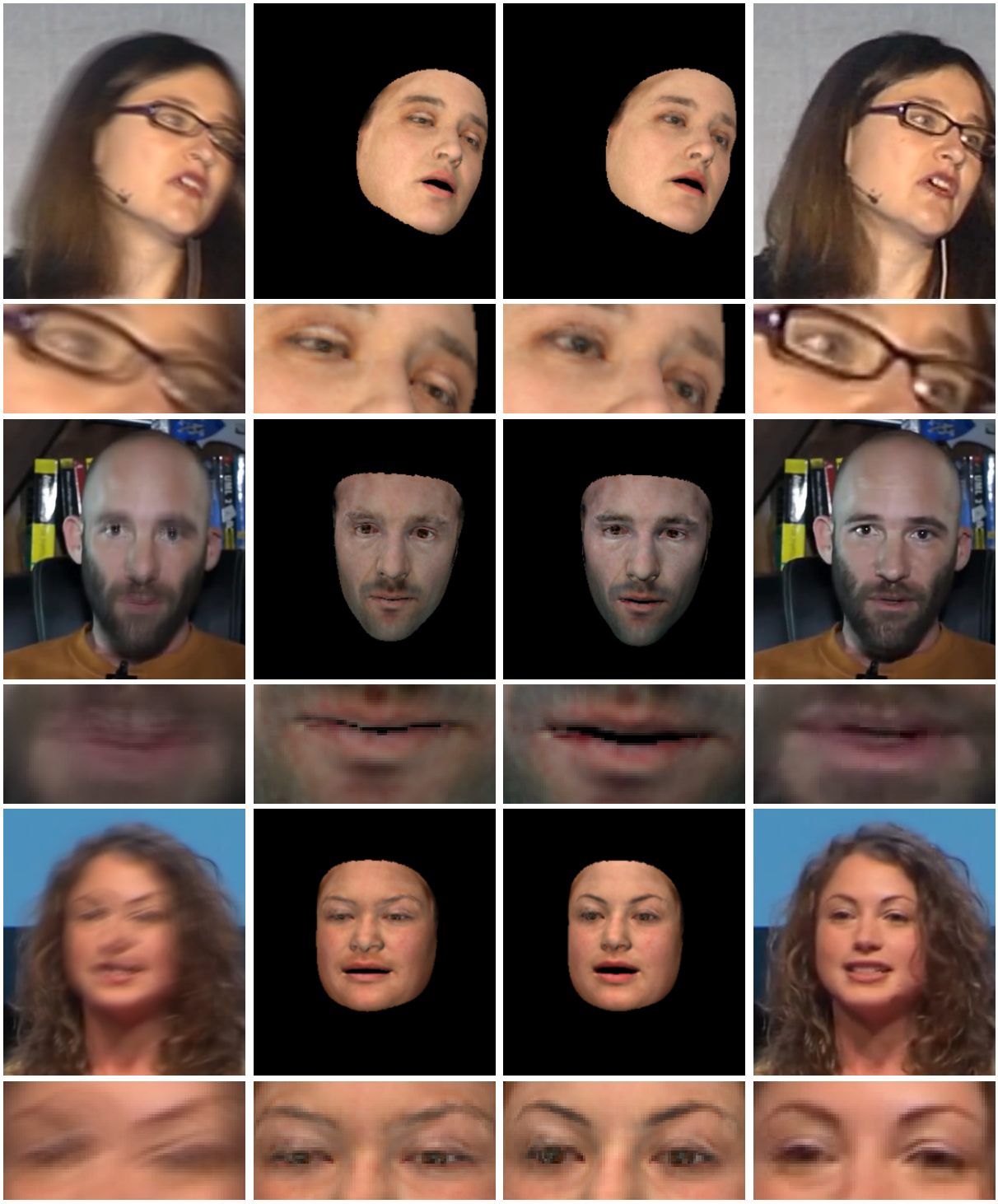
[3] Jinshan Pan, Wenqi Ren, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Learning to deblur images with exemplars. *TPAMI*, 2019. 2, 4, 5, 6, 7

[4] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCVW*, 2015. 1

[5] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 2, 4, 5, 6, 7

[6] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 1, 2, 4, 5, 6, 7

[7] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 1, 2, 4, 5, 6, 7

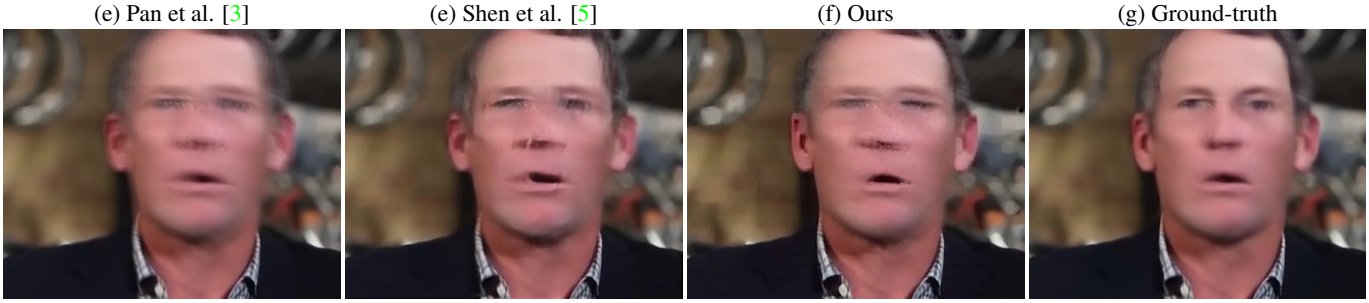


(a) Blurred inputs (b) w/o the rendering loss (c) w/ the proposed rendering loss (d) Ground-truths

Figure 2. 3D face rendered results of the proposed rendering branch. (a) Blurred face frames. (b) Rendered results without using the proposed face rendering loss $\mathcal{L}_r$. (c) Rendered results with the proposed face rendering loss $\mathcal{L}_r$. (d) Ground-truths.

PSNR | 29.40 dB | 28.96 dB | 30.32 dB
(a) Blurred inputs | (b) Su et al. [6] | (c) Nah et al. [2] | (d) Tao et al. [7]

23.85 dB | 21.38 dB | **31.10 dB** | $+\infty$
(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth



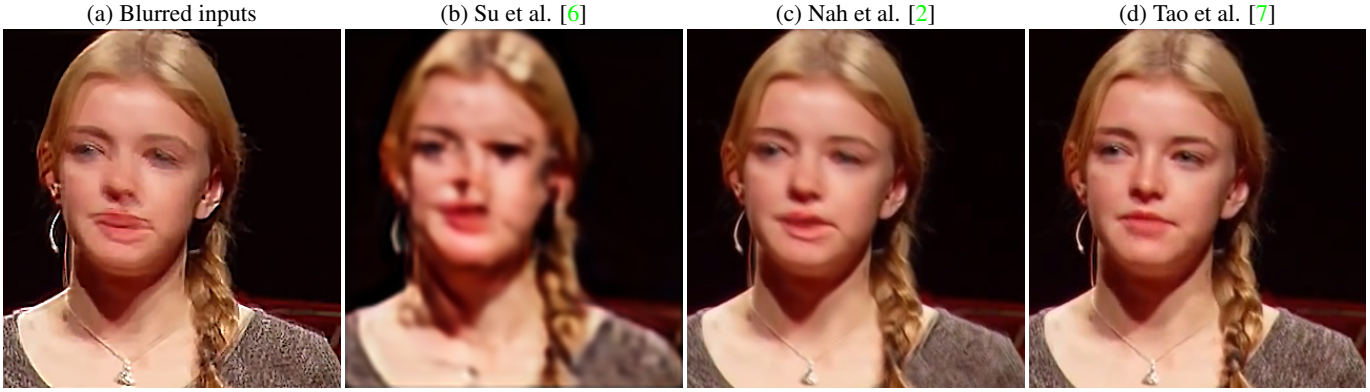
PSNR | 29.51 dB | 29.06 dB | 30.99 dB
(a) Blurred inputs | (b) Su et al. [6] | (c) Nah et al. [2] | (d) Tao et al. [7]

23.57 dB | 25.68 dB | **31.52 dB** | $+\infty$
(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth

Figure 3. Quantitative results from our testing set, with PSNR and SSIM relative to the ground truth. Here we compare algorithm with single image deblurring approaches [7, 2], video deblurring [6, 1], and face deblurring methods [3, 5].

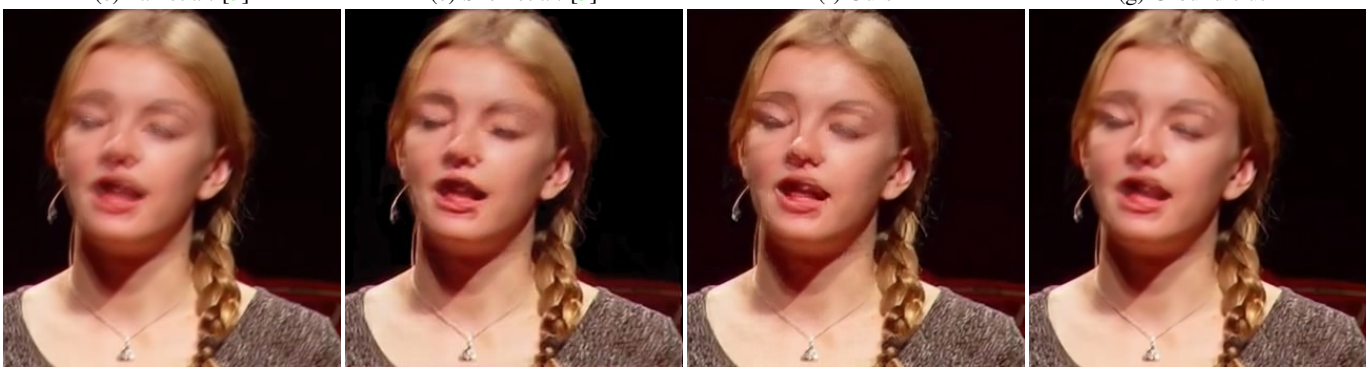
PSNR | 27.56 | 28.59 dB | 28.35 dB

(a) Blurred inputs | (b) Su et al. [6] | (c) Nah et al. [2] | (d) Tao et al. [7]



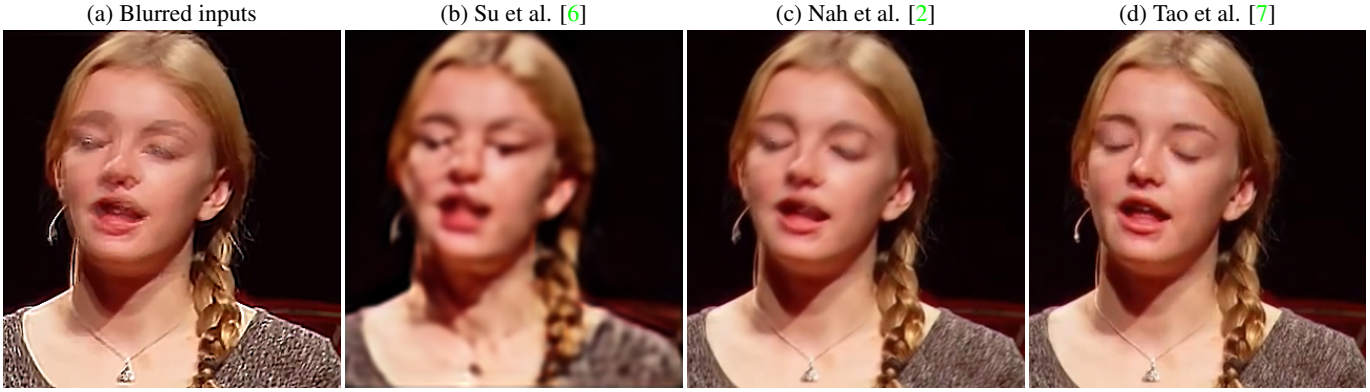
22.82 dB | 19.29 dB | **34.63 dB** | $+\infty$

(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth



PSNR | 29.91 dB | 30.08 dB | 30.17 dB

(a) Blurred inputs | (b) Su et al. [6] | (c) Nah et al. [2] | (d) Tao et al. [7]

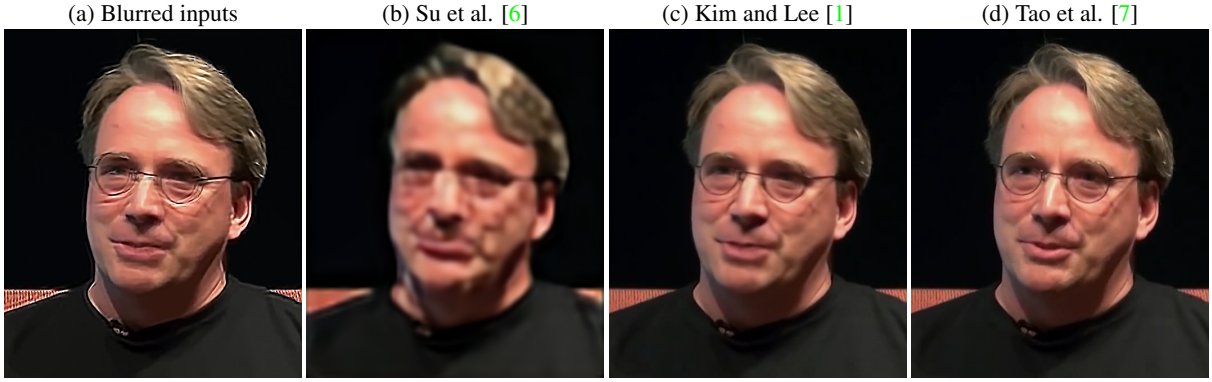


22.62 dB | 21.41 dB | **31.96 dB** | $+\infty$

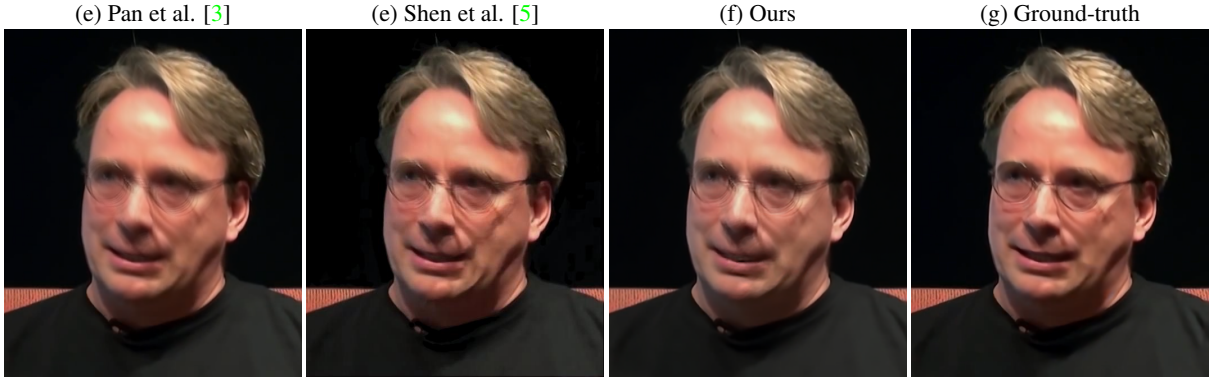
(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth

Figure 4. Quantitative results from our testing set, with PSNR and SSIM relative to the ground truth. Here we compare algorithm with single image deblurring approaches [7, 2], video deblurring [6, 1], and face deblurring methods [3, 5].

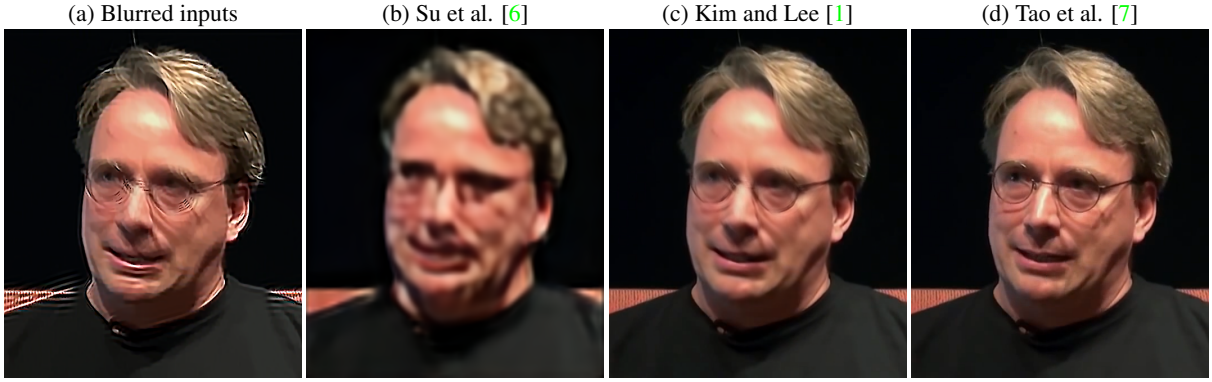
PSNR | 32.81 dB | 32.58 dB | 32.56 dB
(a) Blurred inputs | (b) Su et al. [6] | (c) Kim and Lee [1] | (d) Tao et al. [7]



23.58 dB | 23.42 dB | **35.32 dB** | $+\infty$
(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth

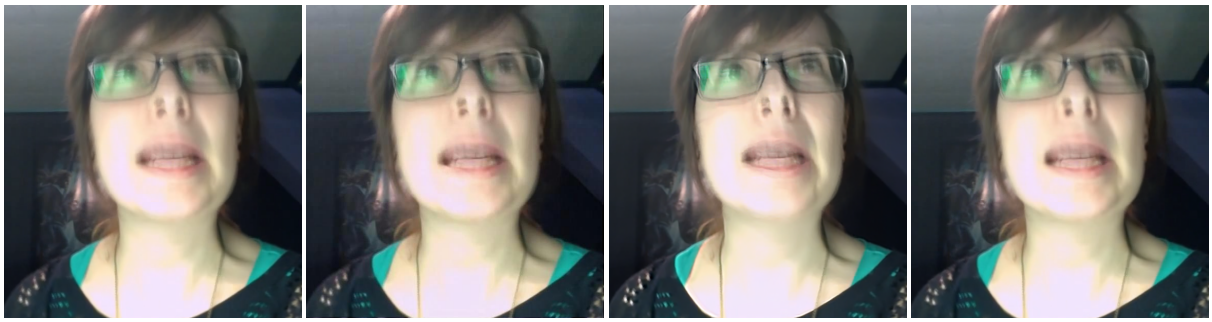


PSNR | 32.46 dB | 32.22 dB | 31.87 dB
(a) Blurred inputs | (b) Su et al. [6] | (c) Kim and Lee [1] | (d) Tao et al. [7]



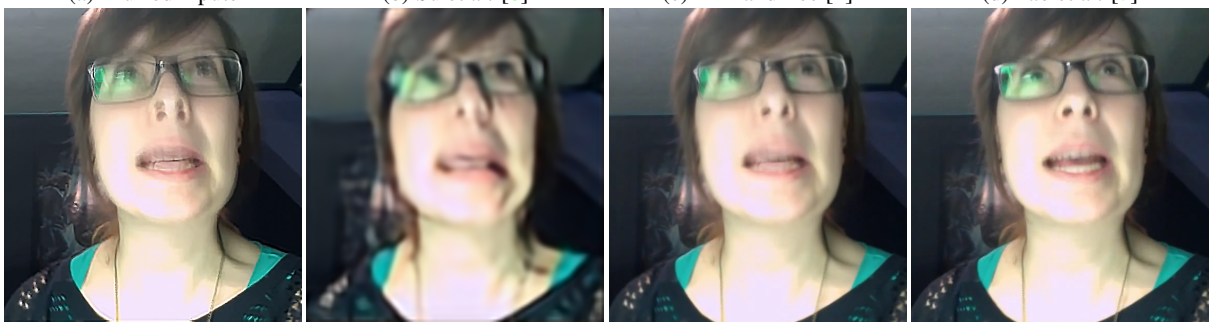
22.11 dB | 21.90 dB | **35.33 dB** | $+\infty$
(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth

Figure 5. Quantitative results from our testing set, with PSNR and SSIM relative to the ground truth. Here we compare algorithm with single image deblurring approaches [7, 2], video deblurring [6, 1], and face deblurring methods [3, 5].



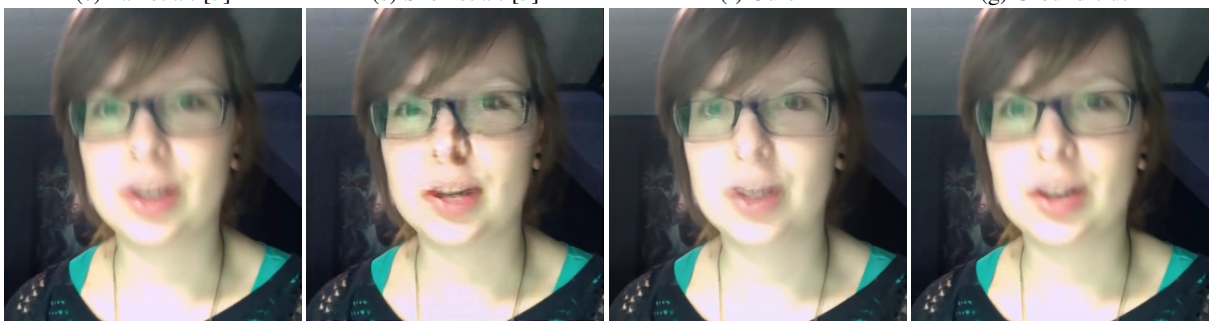
PSNR | 30.77 dB | 31.69 dB | 31.53 dB

(a) Blurred inputs | (b) Su et al. [6] | (c) Kim and Lee [1] | (d) Tao et al. [7]



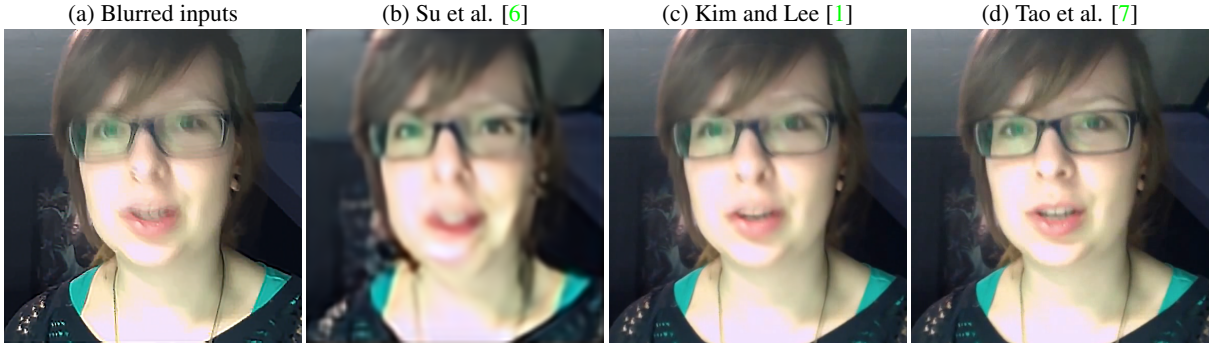
27.00 dB | 23.51 dB | **32.83 dB** | $+\infty$

(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth



PSNR | 32.08 dB | 33.52 dB | 34.09 dB

(a) Blurred inputs | (b) Su et al. [6] | (c) Kim and Lee [1] | (d) Tao et al. [7]



28.57 dB | 24.20 dB | **34.78 dB** | $+\infty$

(e) Pan et al. [3] | (e) Shen et al. [5] | (f) Ours | (g) Ground-truth

Figure 6. Quantitative results from our testing set, with PSNR and SSIM relative to the ground truth. Here we compare algorithm with single image deblurring approaches [7, 2], video deblurring [6, 1], and face deblurring methods [3, 5].