Face Video Deblurring using 3D Facial Priors

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

Abstract

Existing face deblurring methods only consider single frames and do not account for facial structure and identity information. These methods struggle to deblur face videos that exhibit significant pose variations and misalignment. In this paper we propose a novel face video deblurring network capitalizing on 3D facial priors. The model consists of two main branches: i) a face video deblurring subnetwork based on an encoder-decoder architecture, and ii) a 3D face reconstruction and rendering branch for predicting 3D priors of salient facial structures and identity knowledge. These structures encourage the deblurring branch to generate sharp faces with detailed structures. Our method leverages both image intensity and high-level identity information derived from the reconstructed 3D faces to deblur the input face video. Extensive experimental results demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.

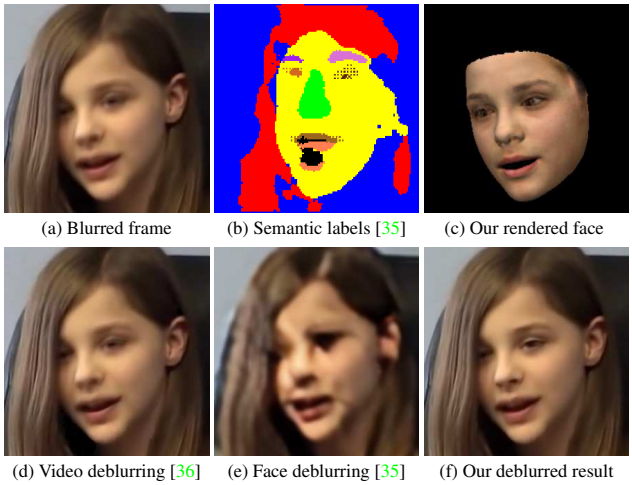
1. Introduction

Face videos captured by hand-held cameras in amateur filming often contain significant camera shake, which results in unpleasant blurry frames in the captured videos. Even for a fixed camera, active people may lead to motion blur that can significantly degrade the video quality. Removing blur and recovering sharp faces from blurry videos are highly desirable under such situations.

Traditional image deblurring algorithms have designed various sharp image priors (e.g., sparsity gradient constraints [43], patch prior [22, 29, 38], and dark channel prior [26, 46]) to constrain the solution space. However, these priors are less effective for face images as the gradient distributions and patch recurrences do not closely follow the generic statistics of natural images.

Given that many studies deblur images based on intermediate salient edges or sharp gradients estimation [8], a sensible way to solve face deblurring is also to implicitly or explicitly extract salient edges or structures from blurred faces [11, 24]. To this end, Pan et al. [24] collect an exemplar dataset of face images and select an exemplar from the dataset with the closest structural similarity to the blurred input, and then use the matched structure to reconstruct salient edges and guide the kernel estimation process. However, this method involves manual image annotations for the exemplar images. Furthermore, it is computationally expensive due to the searching process and the iterative optimization of latent images and blur kernels.

Recently, Convolutional Neural Networks (CNNs) have been successfully applied to natural image deblurring [9, 37]. In this context, some face deblurring networks have been proposed to build a mapping between blurry and sharp faces using large-scale datasets [15, 4, 48]. Shen et al. [35] use CNNs to generate semantic face labels for guiding the deblurring process. The semantic face segmentation serves as global priors and local constraints to determine which component should be in the corresponding region. However, semantic labels can only provide global component regions, but not the detailed edges or other low-level im-



(a) Blurred frame (b) Semantic labels [35] (c) Our rendered face
(d) Video deblurring [36] (e) Face deblurring [35] (f) Our deblurred result

Figure 1. (a) Blurred frame of a face sequence. (b) Predicted semantic labels by [35]. (c) Rendered face by our pipeline, which has clear and sharp facial structure that provides guided spatial location and intensity information of face components. (d) Deblurred result by the video deblurring approach [36]. (e) Deblurred face by [35]. (f) Our deblurred result.

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

age context. Moreover, as shown in Figure 1(b), the method of Shen et al. [35] fails to localize facial components accurately and consequently produces a result with severe distortions in Figure 1(e). Therefore, using semantic labels in face deblurring is suboptimal and may lead to ghosting artifacts in the deblurred result. In addition, it is also worth noting that all of the aforementioned face deblurring approaches ignore facial identity information.

In contrast to previous methods, we propose a face video deblurring method by predicting facial structure and identity information from the blurry face using a deep 3D face reconstruction and rendering branch. Specifically, we first generate a textured 3D face for the central frame using the a 3D face reconstruction network, which provides both image-level (e.g., intensity with sharp edges) and perception-level (e.g., identity) information. Then the face deblurring network applies the rendered, pose-aligned face image as guidance to restore a sharp face. In addition, to encourage generating identity-related face details during the deblurring process, we further embed the identity descriptor extracted by the 3D reconstruction network into the deblurring network. We show that the 3D facial priors we exploit in this paper can significantly facilitate face deblurring.

The main contributions of this paper are summarized as follows:

- We propose a face deblurring method from videos by explicitly exploiting 3D facial priors. Our 3D facial priors provides not only sharp facial structures and detailed intensity information as a reference but also a face identity feature representation.
- We present a loss function for 3D reconstruction learning on blurred faces to adapt the face reconstruction and rendering branch to the deblurring task.
- Compared with the state-of-the-art face deblurring methods, the proposed network achieves superior visual quality and identity recognizability on both synthetic and real face videos.

2. Related Work

Face video deblurring relates closely to natural image and video deblurring. In this section, we review related work on generic image/video and face deblurring to help place our contribution in the proper context.

Generic image and video deblurring. Although numerous image deblurring algorithms have emerged in the last decades, success is still very dependent on the scene. Traditional methods often assume that blur is spatially uniform and leverage various image priors, such as the total variation regularizer [28], sparsity [3, 17], color-line [19], and L_0 gradient based regularizers [43], to tackle the ill-posedness of the problem, Although these priors work well on some benchmarks, they are often characterized by restrictive assumptions that limit their practical applicability. Besides, uniform kernel based methods are less effective for complex scenes with spatially-variant blurs [41].

To handle such spatially-variant blur, Gupta et al. [10] model the camera motion as a motion density function for non-uniform deblurring. In [49], a projective motion path model is used to estimate blur kernels by exploiting inter-frame misalignments. However, the global homography projection model cannot well handle object motion and depth variation [5]. To solve this problem, Kim and Lee [13] proposed a segmentation-free algorithm by using bidirectional optical flow to model motion blurs for dynamic scene deblurring. This method is extended to generalized video deblurring in [14] by alternatively estimating optical flow and latent frames. However, the assumption that motion blur is same as optical flow does not hold for complex motions in the real world.

To address these issues, deblurring algorithms based on deep learning have been proposed recently. Sun et al. [37] learn blur kernels via a classification and regression network. Several approaches train deep CNNs [20, 31, 51, 50] as an image prior for uniform deconvolution, which cannot be directly applied in dynamic scenes. To deal with complex motion blurs, Nah et al. [23] proposed a deep multi-scale network which progressively recovers the sharp image from a coarse scale until the full resolution. However, this method may overfit to a specific image resolution or motion scale. Tao et al. [39] adopt a scale-recurrent network to remove blur by sharing network weights across scales so that it can be applied to arbitrary image resolutions. To aggregate information across multiple video frames, Su et al. [36] apply an encoder-decoder network to learn video deblurring by stacking consecutive frames as input. However, all these networks are designed for natural images or videos and cannot be easily modified to leverage facial priors for face deblurring.

Face deblurring. While most deblurring methods work well on natural scenes, they often do not generalize well to face images. To explicitly handle face images, several class-specific image deblurring approaches have been developed. HaCohen et al. [11] use additional reference images with shared content to guide face deblurring. Anwar et al. [1] proposed a frequency-domain class-specific prior to restore the band-pass frequency components for face images. In [24], Pan et al. presented a face image deblurring method by matching the blurred image with a sharp face from an external exemplar dataset. However, searching for a reference from a large exemplar dataset is time-consuming.

Since CNNs have been widely used in natural image deblurring, several methods also employ CNNs to learn a mapping from blurry faces to their sharp counterparts. Xu et al. [45] use a Generative Adversarial Network (GAN) to

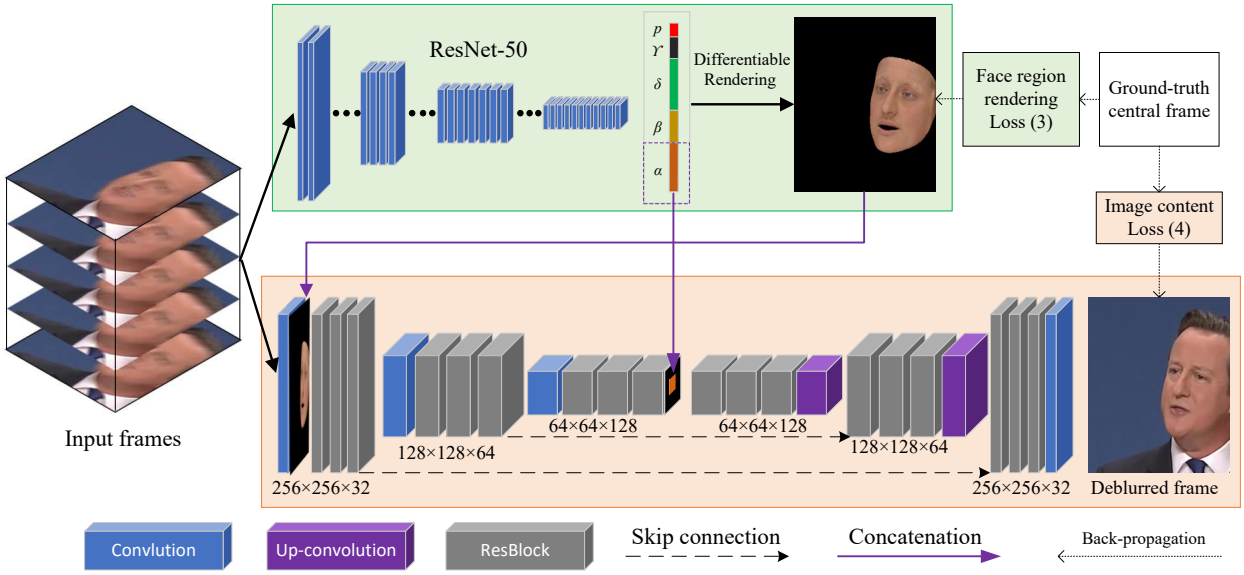


Figure 2. The proposed face video deblurring framework. Our model consists of two branch. The top green block is a ResNet-50 network which aims to reconstruct a 3D face by regressing 3DMM coefficients (as well as pose and illumination parameters) and render a sharp face image. The bottom orange block focuses on face deblurring guided by the extracted identity vector and the rendered sharp face structure from the 3D face reconstruction branch. The residual block is constructed by Conv (k5s1), ReLU, Conv (k5s1), and Relu layers, where k5s1 indicates the convolution kernel is 5×5 with stride 1.

deblur face images. However, without exploiting the unique structure of human faces, this approach is not able to well handle the face restoration problem especially for regions around facial components. Shen et al. [35] exploit semantic labels of faces as a global prior for restoration. Nevertheless, it relies on accurate face segmentation. In addition, all these face deblurring methods do not take the recovery of identity information into consideration and cannot generalize well to non-uniform blur.

Different from these methods, we take both face structure and identity into account. We use a 3D face reconstruction network to extract the face structure and spatial information to guide face blur removal. The identity descriptor from the 3D face reconstruction network is also incorporated to retain identity-aware facial details.

3D face reconstruction. 3D face reconstruction aims to recover the 3D shapes (and textures) of human faces from 2D images. In the literature, the widely-used method for parametric 3D face modeling is 3D Morphable Models (3DMM) [2, 27]. With a 3DMM, face reconstruction can be achieved using an analysis-by-synthesis optimization scheme. The Morphable face model is based on the combination of parametric descriptions of 3D face geometry and texture with PCAs build from a collection of real scans. The reconstructed face will always be "sharp" with clear components (or at least motion-blur-free), since the base textures are all sharp. We exploit this form of 3D reconstruction as the basis of our face deblurring priors.

3. Approach

Given a blurred face video, our network first reconstruct a 3D face based on the 3D Morphable Model (3DMM) [27]. Capitalizing on the regressed 3DMM coefficients, we can generate a high-quality textured 3D face and render it to a reference image to guide face deblurring. Moreover, to provide richer identity information, we take the extracted identity information into consideration such that the deblurring network can better recover identity-related facial details. Figure 2 illustrates the architecture of the proposed face video deblurring network, which consists of two branch: a 3D face reconstruction and rendering branch (green block) and a face video deblurring branch (orange block).

3.1. 3D Face Reconstruction Branch

When the captured face video contains camera shake and/or face motion, it is very difficult for state-of-the-art edge prediction methods to localize sharp edge in blurry frames. In this paper, we consider predicting sharp facial structures using 3D face reconstruction (and 2D rendering), motivated by the following intuitions. First, 3D face reconstruction, especially with a 3DMM representation, can produce sharp, or at least motion-blur-free reference face images via rendering. This is because the PCA models of a 3DMM are built with high-quality face scans captured in controlled static environments. Second, the recent work of [6] has shown that 3DMM fitting using a CNN can produce remarkable reconstruction results that are robust to mod-

est image degradations such as blur and occlusion and can well handle large poses. Finally, the fitted 3DMM coefficients bear the subject's identity information which also can be leveraged to restore a clear image of the subject using a CNN.

To this end, we apply a CNN for 3DMM-based face reconstruction. We follow [6] to use a ResNet-50 Network [12] to regress the 3DMM coefficients together with face pose and environment illumination. The output of the ResNet-50 is a vector $\boldsymbol{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{p}) \in \mathbb{R}^{239}$, 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 $\boldsymbol{p} \in \mathbb{R}^{6}$ are the parameters of face identity, expression, texture, illumination, and pose, respectively. With the extracted 3DMM coefficients, we can easily construct a 3D face shape $\mathbf{S}$ and texture $\mathbf{T}$ for the input face image as

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

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

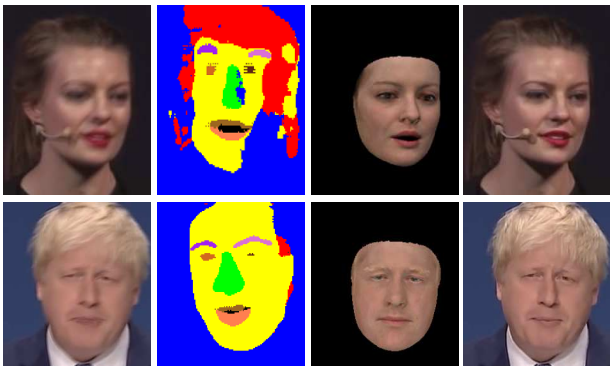
where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the mean face shape and texture, respectively. $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_{t}$ represent the PCA bases of identity, expression, and texture, respectively. We can then project the constructed 3D face onto 2D image plane with the regressed pose and illumination, and obtain a rendered face image as illustrated in Figure 1-3. For more details regarding the 3D face model and the rendering process, we refer the readers to [6].

With a differentiable mesh renderer, the 3D face reconstruction network can be trained in an unsupervised/weakly-supervised fashion on natural face images without the need for ground truth labels [40, 7, 6]. To further improve the performance on blurry faces, we apply a new rendering loss function to finetune the pre-trained network on our paired training frames so that the 3D reconstruction branch better adapts to our deblurring task:

$$\mathcal{L}_r = \frac{1}{pq}\sum_{v=1}^{p}\sum_{f=1}^{q}\frac{\sum_{i\in M_{v,f}} A^i_{v,f}\|I^i_{v,f} - R^i\big(\boldsymbol{x}(B_{v,f})\big)\|_2}{\sum_{i\in M_{v,f}} A^i_{v,f}}, \quad (3)$$

where f is the frame index, v denotes the video index, and p, q are the total number of training videos and frames in each training video, respectively. In addtion, I and B are sharp and blurry image pairs in the training data, i denotes pixel index, M is reprojected face region, A is a skin color based attention mask for the training image [6], $\boldsymbol{x}(B)$ denotes the regressed coefficients by the network with B as input, and finally R denotes the image rendered with $\boldsymbol{x}$.

As shown in Figure 3(c), given blurred frames our 3D face reconstruction branch can generate rendered faces with clear face components which are visually quite similar to the ground-truths in Figure 3(d). In contrast to the state-of-the-art face deblurring method [35] which fails to locate facial components as shown in Figure 3(b), our face reconstruction branch is more robust to motion blur and can reconstruct facial components well. The rendered face image is well-aligned with the blurry input, providing sharp intensity refernece for the facial components. Next, we present the face debluring branch which will leverage the rendered image as well as the identity information for debluring.



(a) blurred inputs (b) Semantic [35] (c) Our rendered (d) Ground truth

Figure 3. Intermediate rendered results by our method. (a) Blurred frames. (b) Semantic labels predicted by the face parsing network in [35]. (c) Our rendered face structures. (d) Ground-truth sharp images. As shown, the semantic labeled from [35] do not have corresponding position and shape of faces. By contrast, the rendered face images by our face reconstruction branch provide clear spatial positions and intensity information of the facial components.

3.2. Deblurring Branch

The debluring branch takes the blurry face video frames and the 3D reconstruction results as input and predicts clear face images. Similar to [36], we perform an early fusion of the consecutive frames by concatenating them in the input layer. Multi-frame input will provide not only motion cues but also complementary information across frames, thus leading to superior performance compared to single-image input (see the supplementary material for an experimental comparison). The output of the debluring branch is the predicted clear face content of the central frame.

We employ an encoder-decoder structure for the deblurring network, which has been shown to produce remarkable results for a number of generative tasks [21, 30, 32, 39, 44]. In particular, we choose a variation of the residual encoder-decoder network model for face video deblurring. We use skip connections between encoder and decoder which can significantly accelerate the convergence [21]. We implement stacked convolution and residual blocks (ResBlock) [12] in the debluring network, as shown in the orange block of Figure 2. More details about the network structure can be found in the supplementary material.

To leverage the facial priors generated by the 3D face reconstruction branch, we incorporate the rendered sharp face as an additional feature map into the face video deblur-

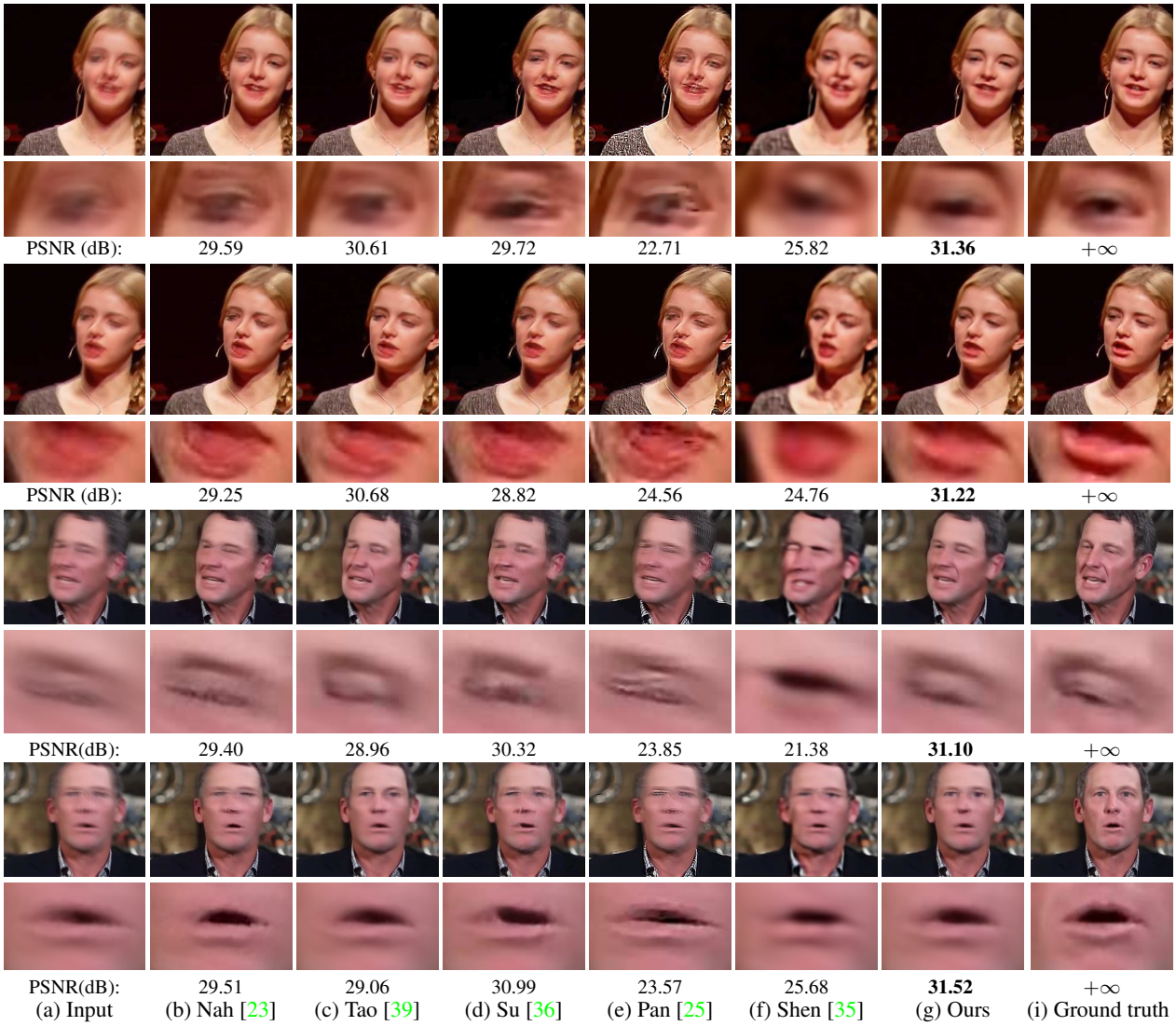


Figure 4. Face debluring results on the testing set with PSNR and SSIM relative to the ground truth. Here we compare our algorithm with single image deblurring approaches [23, 39], video deblurring [36], and face deblurring methods [25, 35].

ring branch. As shown in the orange block of Figure 2, the rendered face structure is concatenated with the first convolutional layer to provide the reference spatial positions and intensity information of facial components. To impose the identity information, we concatenate the identity vector $\boldsymbol{\alpha}$ to the last layer of the encoder network. As shown in Figure 2, we first reshape $\boldsymbol{\alpha}$ to a 9×9 matrix by setting the last element as zero, then we expand it to the size of the 64×64 by zero-padding and concatenate it with the last-layer output of encoder.

We note that previous image deblurring networks [23, 39] (including the state-of-the-art face deblurring approach of [35]) often adopt a multi-scale debluring strategy to recover sharp images which mimics the traditional coarse-to-fine optimization scheme. By contrast, benefited from the estimated sharp facial structure from the 3D face rendering, our algorithm acts on the original scale only and performs well without any coarse-to-fine strategy, which simplifies the face video deblurring process significantly.

The training losses of the our deblurring branch is the Euclidean difference between the image content of the network output and the ground truth central frame,

$$\mathcal{L}_c = \frac{1}{pq} \sum_{v=1}^{p} \sum_{f=1}^{q} \|I_{v,f} - I'_{v,f}\|_2. \tag{4}$$

where I and I' are the ground-truths and deblurred results, respectively. Note that in this work, we do not use other sophisticated loss functions such as adversarial loss [18, 23] and motion flow loss [9]. We show that simply using the naive Euclidean image intensity loss $\mathcal{L}_c$ can already achieve very competitive results, as will be demonstrated next.

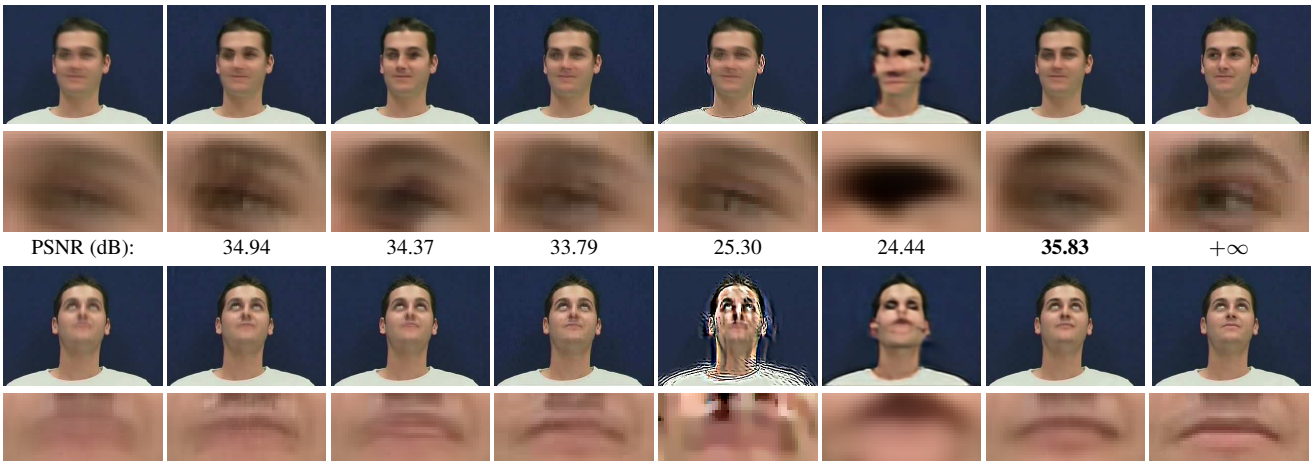


Figure 5. Face debluring results on the testing data, with PSNR and SSIM relative to the ground truth. Here we compare our algorithm with single image deblurring approaches [23, 39], video deblurring [36], and face deblurring methods [25, 35].

Table 1. Quantitative PSNR and SSIM results on the synthetic datasets using different deblurring methods.

	Pan et al. [25]	Shen et al. [35]	Su et al. [36]	Nah et al. [23]	Tao et al. [39]	Ours
9 synthetic testing videos from the 300VW dataset [34]						
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 [33]						
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**

4. Experiments

In this section, we evaluate the proposed method on both synthetic datasets and real-world face videos with comparisons to state-of-the-art image/video deblurring methods.

4.1. Implementation Details

Our method is implemented with Tensorflow. We use a batch size 16 for training and image patches of $256 \times 256 \times 15$ where 15 is the total number of RGB channels stacked from the crops of 5 neighboring frames. The Adam optimizer [16] is applied with decay rates β_1 and β_2 set as 0.9 and 0.999, respectively. The initial learning rate is 0.0001 and we decrease the learning rate by 0.3 every 50K iterations. For all the results reported in the paper, we train the network for 400K iterations.

4.2. Training Data

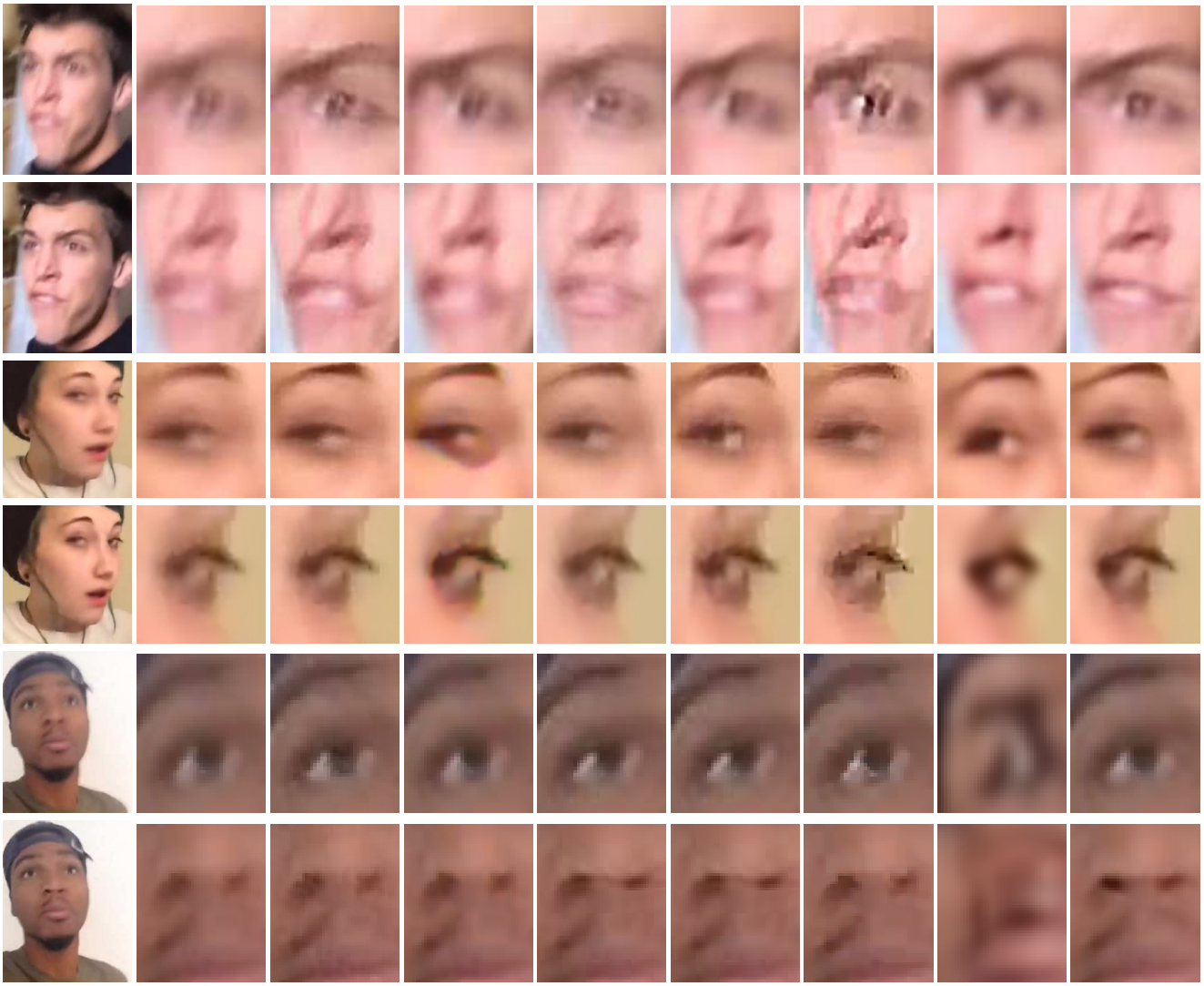
To create a large face deblurring training dataset, Shen et al. [35] synthesize blurred images by convolving sharp images with generated uniform blur kernels. However, images with uniform blur are different from real cases captured by cameras. Similar to [36], we opt for generating blurred images through averaging 5 consecutive frames from sharp videos to approximate motive blur. The generated face frames are more realistic since they can simulate complex camera shake and face motion.

In this paper, we use the 300VW face dataset [34] to synthesize our training videos since most faces therein are sharp with high resolutions. Since these videos are captured with general commodity cameras, there are still some low-quality videos inappropriate for our synthesis purpose. Therefore, we remove videos that are already blurred and/or of low resolutions and use the remaining for data generation. Specifically, we select 83 videos as our training data and 9 videos as testing data from the 114 videos in the 300VW dataset.

4.3. Quantitative Evaluation

In this section, we compare the proposed algorithm with following six methods of video deblurring [36], natural image deblurring [23, 39], and face image deblurring [25, 35]. For fair comparisons, we fine-tuned the image deblurring network of [39] with another 50K iterations and re-trained the video deblurring network of [36] using the same training data in this work. We evaluate the results of different methods by PSNR and SSIM metrics.

300VW dataset. Table 1 reports the average PSNR and SSIM values of the deblurred frames on the test data. The results generated by the proposed algorithm have much higher PSNR and SSIM values than all other competitors. Figure 4 shows four examples from the test set synthesized by the 300VW dataset [34]. The single image deblurring



(a) Input / Ours (b) Input (zoom) (c) Nah [23] (d) Tao [39] (e) Kim [14] (f) Su [36] (g) Pan [25] (h) Shen [35] (i) Ours

Figure 6. Qualitative results for real-world blurry face videos. Here we compare our algorithm with single image deblurring approaches [23, 39], video deblurring algorithms [36, 14], and face deblurring methods [25, 35].

methods of [23, 39] fail to generate sharp face components as shown in Figure 4(b) and (c). The deep video deblurring method of Su et al. [36] is developed dynamic scenes. However, the final recovered frames contain some artifacts as shown in Figure 4(d). Compared with the face image deblurring methods of [25, 35], our method generates much sharper images with clearer structures.

VidTIMIT dataset. In addition to the testing data from the 300VW dataset, we further synthesize more testing videos using the VidTIMIT dataset [33] to evaluate the generalization ability of the proposed method. We randomly select 11 videos from VidTIMIT, synthesize the corresponding testing data, and directly run the different methods on them. Table 1 shows that our method generalizes well to these new face videos and yields higher PSNR and SSIM values than other methods again. The examples shown in Figure 5 also demonstrate the superiority of our results.

4.4. Qualitative Evaluation

To further evaluate the proposed method on real data, we collect a suite of videos from YouTube containing blurred face frames that caused by camera shake and human motion, and compare against video deblurring [36, 14], natural image deblurring [23, 39], and face image deblurring [25, 35]. Although there is no ground truth for quantitative analysis, the difference in visual quality is clearly visible for the restored facial components, as shown in Figure 6. The uniform face deblurring methods of [25, 35] failed to generate clear results as these methods focus on blur caused by camera shake. The CNN-based methods of [23, 39] are

(a) Input (b) Baseline (c) w/o rend. (d) w/o iden. (e) Ours

Figure 7. Comparisons between our proposed face video deblurring model with different configurations. (a) Input. (b) The baseline method without using the 3D face rendering sub-network. (c) The deblurred result without using the rendered face image. (d) Deblurred result without the concatenated identity vector. (e) Our deblurred result.

Table 2. Quantitative results with different configurations on 9 synthetic testing videos generated from the 300VW dataset.

		PSNR/SSIM		Identity similarity	
ident.	rend.	Su [36]	Ours	Su [36]	Ours
×	×	34.40/0.9218	36.74/0.9817	0.8352	0.8352
√	×		36.82/0.9812		0.8335
×	√		37.53/**0.9869**		0.8364
√	√		**37.70**/0.9849		**0.8373**

designed for dynamic scene deblurring. However, they are not able to remove face blur as shown in Figure 6(c) and (d). The video deblurring method of [36] is also less effective for face deblurring as shown in Figure 6(f). By contrast, the proposed method produces higher-fidelity faces with finer facial components details as shown in Figure 6(i).

4.5. Ablation Study

In this section, we compare the proposed network with and without using the 3D facial priors provided by the 3D face reconstruction branch. Here we evaluate not only PSNR and SSIM values, but also the identity similarity between the deburred result and the ground truth computed as the cosine distance of deep face features extracted by a face recognition network from [47].

As shown in Figure 7, the baseline method without using rendered face and identity information (vector $\boldsymbol{\alpha}$) tends to generate some artifacts around the facial components, while without the rendered face some details are lost in the deblurred result and the edges are not sharp enough, as shown in Figure 7(c). By adding the rendered face in the face deblurring branch, the result exhibits clearer and sharper facial structures in Figure 7(d) and (e), which demonstrates that the proposed 3D face rendering module could help the network understand the spatial and intensity information of the face components thus generating better deblurring results. The quantitative results on in Table 2 also demonstrate the effectiveness of our 3D facial priors. As shown, with the rendered facial structure, the proposed algorithm obtains highest SSIM value on the test data, while adding both facial structure and identity knowledge obtains optimal performance in terms of PSNR and identity similarity scores. The results from the deep video deblurring method of [36] are also presented in Table 2 for reference.

Table 3. Average running time on testing video frames.

Image size	Nah [23]	Tao [39]	Su [36]	Ours
360×450	2.03s	1.81s	**0.31s**	0.67s
580×610	2.56s	1.98	0.77s	**0.70s**

We also conducted ablation studies to analyze the effectiveness of our blurry-image face rendering loss L_r in Equation 3 as well as the multi-frame input scheme (vs. single-image input). They are presented in the supplementary material due to space limitation.

4.6. Running Time

In terms of running time, the proposed algorithm performs favorably against the state-of-the-art image and video deblurring methods [23, 39, 24, 14, 36]. The average running times for two image resolutions are shown in Table 3. 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 [23, 39] use a multi-scale strategy which inevitably increases the computational cost. They are clearly less efficient than our method. The video deblurring approach of [36] runs faster than our algorithm on low-resolution frames. However, our method tends to have better scalability and is slightly faster than [36] on higher resolutions.

5. Conclusion

We have presented a face video deblurring network which incorporates 3D facial priors. Our network exploits the face rendering loss to estimate a high-quality rendered image as guidance, which provides clear spatial location of facial components and their intensity information. We also embed the estimated identity vector from the 3DMM face reconstruction into the deblurring branch to better recover identity-related facial details. Quantitative and qualitative results show that our proposed network performs favorably against the state-of-the-art deblurring methods and can generate visually-pleasing results on real-world blurred face videos. We believe 3D information is valuable for low-level vision and image processing tasks, and foresee more applications of face image/video processing and other tasks using 3D priors.

Acknowledgments. This work is supported in part by the National Key R&D Program of China (Grant No. 2018YFB0803701), National Natural Science Foundation of China (No. 61802403, U1605252, U1736219, U1803264, 61861166002), Beijing Natural Science Foundation (No. L182057), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

[1] Saeed Anwar, Cong Phuoc Huynh, and Fatih Porikli. Image deblurring with a class-specific prior. *TPAMI*, 2018. 2

[2] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 3

[3] Xiaochun Cao, Wenqi Ren, Wangmeng Zuo, Xiaojie Guo, and Hassan Foroosh. Scene text deblurring using text-specific multiscale dictionaries. *TIP*, 24(4):1302–1314, 2015. 2

[4] Grigorios G Chrysos and Stefanos Zafeiriou. Deep face deblurring. In *CVPRW*, 2017. 1

[5] Mauricio Delbracio and Guillermo Sapiro. Hand-held video deblurring via efficient fourier aggregation. *TCI*, 1(4):270–283, 2015. 2

[6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshop on Analysis and Modeling of Faces and Gestures*, 2019. 3, 4

[7] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018. 4

[8] Dong Gong, Mingkui Tan, Yanning Zhang, Anton Van den Hengel, and Qinfeng Shi. Blind image deconvolution by automatic gradient activation. In *CVPR*, 2016. 1

[9] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *CVPR*, 2017. 1, 5

[10] Ankit Gupta, Neel Joshi, C Lawrence Zitnick, Michael Cohen, and Brian Curless. Single image deblurring using motion density functions. In *ECCV*, 2010. 2

[11] Yoav Hacohen, Eli Shechtman, and Dani Lischinski. Deblurring by example using dense correspondence. In *ICCV*, 2013. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[13] Tae Hyun Kim and Kyoung Mu Lee. Segmentation-free dynamic scene deblurring. In *CVPR*, 2014. 2

[14] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, 2015. 2, 7, 8

[15] Meiguang Jin, Michael Hirsch, and Paolo Favaro. Learning face deblurring fast and wide. In *CVPRW*, 2018. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[17] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, 2011. 2

[18] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 5

[19] Wei-Sheng Lai, Jian-Jiun Ding, Yen-Yu Lin, and Yung-Yu Chuang. Blur kernel estimation using normalized color-line prior. In *CVPR*, 2015. 2

[20] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Learning a discriminative prior for blind image deblurring. In *CVPR*, 2018. 2

[21] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NeurIPS*, 2016. 4

[22] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *ECCV*, 2014. 1

[23] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2, 5, 6, 7, 8

[24] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring face images with exemplars. In *ECCV*, 2014. 1, 2, 8

[25] Jinshan Pan, Wenqi Ren, Zhe Hu, and Ming-Hsuan Yang. Learning to deblur images with exemplars. *TPAMI*, 41(6):1412–1425, 2018. 5, 6, 7

[26] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, 2016. 1

[27] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, 2009. 3

[28] Daniele Perrone and Paolo Favaro. Total variation blind deconvolution: The devil is in the details. In *CVPR*, 2014. 2

[29] Wenqi Ren, Xiaochun Cao, Jinshan Pan, Xiaojie Guo, Wangmeng Zuo, and Ming-Hsuan Yang. Image deblurring via enhanced low-rank prior. *TIP*, 25(7):3426–3437, 2016. 1

[30] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, 2018. 4

[31] Wenqi Ren, Jiawei Zhang, Lin Ma, Jinshan Pan, Xiaochun Cao, Wangmeng Zuo, Wei Liu, and Ming-Hsuan Yang. Deep non-blind deconvolution via generalized low-rank approximation. In *NeurIPS*, 2018. 2

[32] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *TIP*, 28(4):1895–1908, 2018. 4

[33] Conrad Sanderson and Kuldip K Paliwal. Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449–480, 2004. 6, 7

[34] 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. 6

[35] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7

[36] 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, 8

[37] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, 2015. 1, 2

[38] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *ICCP*, 2013. 1

[39] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 2, 4, 5, 6, 7, 8

[40] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. MoFa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 4

[41] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *IJCV*, 98(2):168–186, 2012. 2

[42] Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, and Hendrik Lensch. Learning blind motion deblurring. In *ICCV*, 2017.

[43] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, 2013. 1, 2

[44] Xiangyu Xu, Deqing Sun, Sifei Liu, Wenqi Ren, Yu-Jin Zhang, Ming-Hsuan Yang, and Jian Sun. Rendering portraitures from monocular camera and beyond. In *ECCV*, 2018. 4

[45] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 2

[46] Yanyang Yan, Wenqi Ren, Yuanfang Guo, Rui Wang, and Xiaochun Cao. Image deblurring via extreme channels prior. In *CVPR*, 2017. 1

[47] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 8

[48] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *ECCV*, 2018. 1

[49] Haichao Zhang and Jianchao Yang. Intra-frame deblurring by leveraging inter-frame camera motion. In *CVPR*, 2015. 2

[50] Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson WH Lau, and Ming-Hsuan Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *CVPR*, 2017. 2

[51] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, 2017. 2