International Journal of Computer Vision
https://doi.org/10.1007/s11263-019-01288-9

Exploiting Semantics for Face Image Deblurring

Ziyi Shen[1,3] · Wei-Sheng Lai[2] · Tingfa Xu[3] · Jan Kautz[4] · Ming-Hsuan Yang[2,5,6]

Received: 2 October 2018 / Accepted: 27 December 2019

Abstract
In this paper, we propose an effective and efficient face deblurring algorithm by exploiting semantic cues via deep convolutional neural networks. As the human faces are highly structured and share unified facial components (e.g., eyes and mouths), such semantic information provides a strong prior for restoration. We incorporate face semantic labels as input priors and propose an adaptive structural loss to regularize facial local structures within an end-to-end deep convolutional neural network. Specifically, we first use a coarse deblurring network to reduce the motion blur on the input face image. We then adopt a parsing network to extract the semantic features from the coarse deblurred image. Finally, the fine deblurring network utilizes the semantic information to restore a clear face image. We train the network with perceptual and adversarial losses to generate photo-realistic results. The proposed method restores sharp images with more accurate facial features and details. Quantitative and qualitative evaluations demonstrate that the proposed face deblurring algorithm performs favorably against the state-of-the-art methods in terms of restoration quality, face recognition and execution speed.

Keywords Face image deblurring · Semantic face parsing · Deep convolutional neural networks

1 Introduction

Single image deblurring aims to recover a clear image from a single blurred input. Conventional methods formulate the blur process (assuming spatially invariant blur) as the convolution operation between a latent clear image and a blur kernel, and solve this problem based on the maximum a posteriori (MAP) framework. As the problem is ill-posed, the state-of-the-art algorithms typically rely on natural image priors [e.g., L_0 gradient (Xu et al. 2013) and dark channel prior (Pan et al. 2016b)] to constrain the solution space.

While existing image priors are effective for deblurring natural images, the underlying assumption may not hold well for images from specific categories, e.g., text, face and low-light conditions. Numerous approaches exploit domain specific visual information, such as designing L_0 intensity (Pan et al. 2017b) priors for text images or detecting light streaks (Hu et al. 2014a) for extremely low-light images. As face images contain fewer textures and edges for estimating blur kernels, Pan et al. (2014) search for similar face exemplars from an external dataset and extract the contour as reference edges. However, reference images may not always exist for a specific input due to diversity of real-world face images. Furthermore, those methods based on the MAP framework typically entail heavy computational cost due to the iterative optimization process to determine latent images and blur kernels. The long execution time limits the applications on resource-sensitive platforms, e.g., mobile devices.

In this work, we propose an efficient and effective solution to deblur face images via deep convolutional neural

Communicated by Julien Mairal.

✉ Tingfa Xu
ciom_xtf1@bit.edu.cn

Ziyi Shen
ziyishen@bit.edu.cn

Wei-Sheng Lai
wlai24@ucmerced.edu

Jan Kautz
jkautz@nvidia.com

Ming-Hsuan Yang
mhyang@ucmerced.edu

1 Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

2 School of Engineering, University of California at Merced, Merced, CA, USA

3 School of Optics and Photonics, Beijing Institute of Technology, Beijing, China

4 Nvidia, Santa Clara, CA, USA

5 Google, Mountain View, CA, USA

6 Yonsei University, Seoul, Korea

Published online: 30 March 2020

networks (CNNs). Since face images are highly structured and composed of similar components, semantic information can provide a strong prior for restoration. We propose to leverage the face semantic labels as global priors and local constraints to train a deep CNN. The proposed model consists of three sub-networks: a coarse deblurring network, a face parsing network, and a fine deblurring network. The coarse deblurring network first predicts a deblurred image from the given input blurred image. The face parsing network then estimates the semantic labels from the coarse deblurred image. Finally, the fine deblurring network takes the blurred image, coarse deblurred image, and semantic labels to restore a clear face image. To encourage the network to restore fine details, we propose an adaptive local structural loss on important face components (e.g., eyes, noses, and mouths). Finally, we impose a perceptual loss (Johnson et al. 2016) and an adversarial loss (Goodfellow et al. 2014) to generate photo-realistic deblurred results. As our method is end-to-end without any blur kernel estimation or post-processing, the execution time is significantly shorter than the conventional MAP-based approaches.

To handle blurred images caused by unknown blur kernels, we construct a large face blurred image dataset for training and testing. We first synthesize random blur kernels by modeling the camera trajectories (Chakrabarti 2016; Hradiš et al. 2015). Next, we generate blurred face images using the synthesized blur kernels and face images from the Helen (Le et al. 2012), CMU PIE (Sim et al. 2002), and CelebA (Liu et al. 2015) datasets. We show that the proposed model trained on synthetic images generalizes well to images generated by unseen blur kernels as well as real blurred images. The proposed method reconstructs better facial details and achieves higher accuracy on face detection and recognition than the state-of-the-art face deblurring approaches (Shen et al. 2018; Pan et al. 2014) (see Fig. 1).

In this work, we make the following contributions:

- We propose a deep multi-scale CNN that exploits global semantic priors and local structural constraints for face image deblurring. The proposed local structural loss adaptively adjusts the weights based on the size of each facial component and greatly improves the quantitative and qualitative results.
- We develop a large-scale blurred face image dataset. The training set consists of 130 million blurred images (synthesized from 6464 face images and 20,000 blur kernels) and the test set has 16,000 blurred images (synthesized from 200 face images and 80 blur kernels). Our dataset can serve as a common benchmark for training and evaluating face image deblurring.
- We demonstrate that the proposed method performs favorably against the state-of-the-art deblurring approaches in terms of restoration quality, face detection, recognition and execution speed.

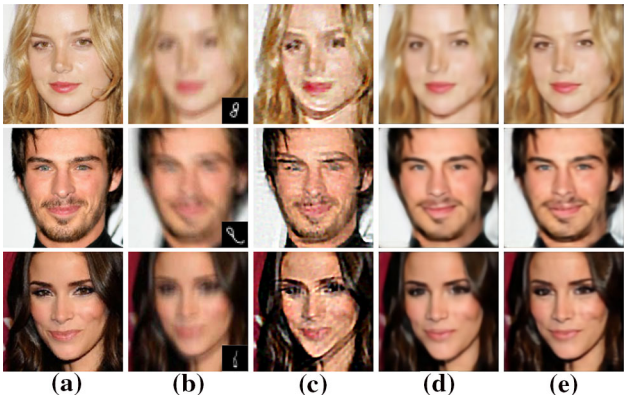


Fig. 1 Visual comparison on face image deblurring. We exploit the semantic information of face images within an end-to-end deep CNN for deblurring. **a** Ground truth images, **b** blurred images **c** Pan et al. (2014) **d** Shen et al. (2018), **e** ours

2 Related Work

Our work belongs to the single-image blind image deblurring problem, where the blur kernel is unknown. In this section, we focus our discussion on generic, domain specific, and recent CNN-based image deblurring approaches.

2.1 Generic Image Deblurring Methods

The recent advances in single image blind deblurring can be attributed to the development of effective natural image priors, including sparse gradient prior (Fergus et al. 2006; Levin et al. 2009), normalized sparsity measure (Krishnan et al. 2011), patch prior (Sun et al. 2013a), L_0 gradient (Xu et al. 2013), color-line prior (Lai et al. 2015), low-rank prior (Ren et al. 2016a), self-similarity (Michaeli and Irani 2014), and extreme channel priors (Pan et al. 2016b; Yan et al. 2017). Recently, a number of approaches learn data fitting functions (Pan et al. 2017a) or image priors with Markov random fields (MRFs) (Liu et al. 2018) to recover latent images. By optimizing the image priors within the MAP framework, those approaches *implicitly* restore strong edges, and therefore, estimate blur kernels and latent sharp images. However, solving complex non-linear priors involves several optimization steps and thus entails high computational loads. Edge-selection based methods (Cho and Lee 2009; Xu and Jia 2010) use simple priors (e.g., L_2 gradients) with image filters (e.g., shock filter) to *explicitly* restore or select strong edges. In addition, a number of approaches use reference images as guidance for non-blind (Sun et al. 2014) and blind

deblurring (Hacohen et al. 2013). However, the performance of such methods hinges on the similarity of reference images and quality of dense correspondence.

While generic image deblurring methods demonstrate the state-of-the-art performance, face images have different statistical properties than natural scenes. Fewer edges or structure on face images can be extracted for blur kernel estimation. The above-mentioned approaches typically cannot deblur face images well and may generate undesired visual artifacts.

Another line of work proposes various motion blur models to handle non-uniform blur (Hirsch et al. 2011; Whyte et al. 2012) and depth variation (Hu et al. 2014b). In this work, we focus on face images caused by uniform motion blur. Our method can also be extended to handle non-uniform blur by synthesizing training data with the non-uniform blur model.

2.2 Domain Specific Image Deblurring Methods

Several domain specific image deblurring approaches have been developed to handle images from different categories. As text images usually contain nearly uniform intensity, Pan et al. (2017b) introduce the L_0-regularized priors on both intensity and image gradients for deblurring text images. To handle extreme cases such as low-light images, Hu et al. (2014a) detect the light streaks in images for estimating blur kernels. Anwar et al. (2015) propose a frequency-domain class-specific prior to restore the band-pass frequency components. Several recent approaches propose outlier detection methods (Pan et al. 2016a) or robust loss functions (Dong et al. 2017) to handle images with non-Gaussian noise.

As face images contain fewer textures and edges, existing algorithms based on implicit or explicit edge restoration are less effective. Pan et al. (2014) search for similar images from a face dataset and extract reference exemplar contours for blur kernel estimation. However, this approach requires manual annotations of the facial contours and involves computationally expensive optimization within the MAP framework. In contrast, we train an end-to-end deep CNN to bypass the blur kernel estimation step, without requiring any reference images or manual annotations for face deblurring.

2.3 CNN-Based Image Deblurring Methods

Deep CNNs have been adopted for several image restoration tasks, such as denoising (Mao et al. 2016), JPEG deblocking (Dong et al. 2015), dehazing (Ren et al. 2016b) and super-resolution (Kim et al. 2016; Lai et al. 2017). Several methods apply deep CNNs for image deblurring in different aspects, including non-blind deconvolution (Schuler et al. 2013; Xu et al. 2014; Zhang et al. 2017), blur kernel estimation (Sun et al. 2015; Schuler et al. 2016; Chakrabarti 2016), and dynamic scene deblurring (Nah et al. 2017; Tao et al. 2018). the state-of-the-art MAP-based approaches, especially in the presence of large motion. Several approaches embed deep CNNs into the conventional MAP-based framework by learning discriminative image priors (Li et al. 2018) or predicting sharp edges (Xu et al. 2018) to achieve the state-of-the-art performance. More recently, Nimisha et al. (2017) and Kupyn et al. (2018) train generative adversarial networks for blind motion deblurring.

A number of methods train end-to-end networks to handle class-specific images, e.g., texts (Hradiš et al. 2015) and faces (Jin et al. 2018; Chrysos et al. 2019). Xu et al. (2017) train generative adversarial networks to jointly deblur and super-resolve low-resolution blurred face and text images, which are typically degraded by Gaussian-like blur kernels. A few face deblurring methods (Jin et al. 2018; Chrysos et al. 2019) based on generic CNNs have recently been developed. Although there are some implementation differences in network architectures and loss functions (e.g., the model of (Jin et al. 2018) is lightweight conditions), these methods do not explore face-related prior information to help the deblurring process. In this work, we focus on deblurring face images affected by complex motion blur. We exploit global and local semantic cues as well as the perceptual (Johnson et al. 2016) and adversarial (Goodfellow et al. 2014) losses to restore photo-realistic face images with fine details.

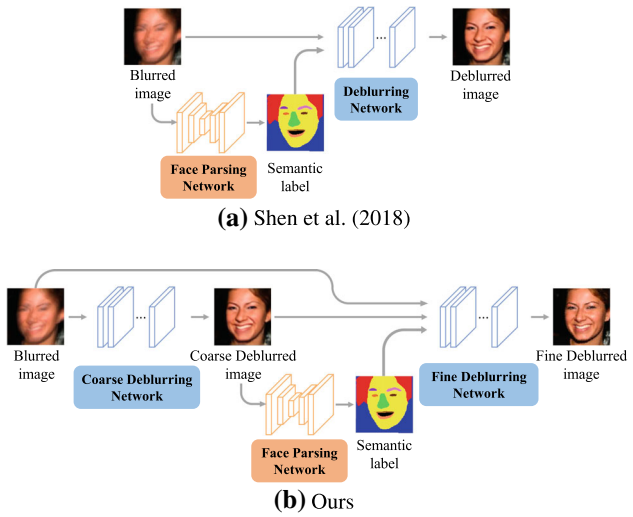
3 Semantic Face Deblurring

In this section, we first give an overview of the proposed face deblurring method. We then describe the design methodology of the network architecture, loss functions, and implementation details.

3.1 Overview

We aim to utilize the face semantic cues to deblur face images. In our preliminary work (Shen et al. 2018), we first apply a face parsing network to extract semantic labels from the input blurred image and then adopt a deblurring network for restoration. We also propose a local structural loss to enforce additional weights on important facial components to recover fine details. However, the labels extracted from the blurred images may be erroneous due to severe motion blur. In this work, we make the following improvements:

– We first construct a coarse deblurring network to reduce the blur in the input image. The face parsing network then extracts semantic labels from the coarse deblurred image. Finally, the fine deblurring network restores a clear face image from the given blurred input image,



(a) Shen et al. (2018)

(b) Ours

Fig. 2 Overview of the proposed model. The state-of-the-art method (Shen et al. 2018) extracts the semantic labels from a blurred image, while we obtain the semantic labels from a coarse deblurred image. The coarse deblurring network reduces the motion blur from the input image and leads to more accurate face parsing results

coarse deblurred image, and corresponding semantic label maps.
- Instead of using a fixed weight for all key components, we propose an adaptive local structural loss which adjusts the weight based on the size of each facial component and restores more fine details.

Figure 2 shows the differences between the method of Shen et al. (2018) and proposed model.

3.2 Network Architecture

Given a blurred face image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ where H and W denotes the height and width of the image, our goal is to recover a clear and sharp face image $\mathbf{y}$ which is as similar as the ground truth image $\mathbf{y}_{\text{GT}}$. To this end, we train an end-to-end deep CNN to deblur the face images efficiently. The proposed face deblurring model consists of three sub-networks: a coarse deblurring network $\mathcal{G}_c$, a face parsing network $\mathcal{P}$, and a fine deblurring network $\mathcal{G}_f$.

Coarse deblurring network To reduce the influence of motion blur on the face parsing, we first use a network to obtain a coarse deblurred image $\mathbf{y}_c$:

$$\mathbf{y}_c = \mathcal{G}_c(\mathbf{x}). \tag{1}$$

We use a multi-scale network similar to the model of Nah et al. (2017), but with several differences. First, as face images typically have smaller spatial resolutions (e.g., 128×128 or less), we use only 2 scales instead of 3 scales for natural images in (Nah et al. 2017). Second, we use fewer ResBlocks (reduce from 19 to 6) and a larger filter size (11×11) at the first convolutional layer to increase the receptive field. The first scale takes as input the $2\times$ downsampled blurred image $\mathbf{x}^{(0.5\times)}$ (3 channels) and generates a deblurred image $\mathbf{y}_c^{(0.5\times)}$. The input to the second scale contains the blurred image $\mathbf{x}$ (3 channels) and the upsampled deblurred image from the first scale $\mathcal{U}_{2\times}(\mathbf{y}_c^{(0.5\times)})$ (3 channels), where $\mathcal{U}_{2\times}$ is $2\times$ bicubic upsampling operator. The output image from the second scale is the coarse deblurred result $\mathbf{y}_c$.

Face parsing network We use an encoder–decoder architecture with skip connections as our face parsing network. The face parsing network takes the coarse deblurred image as input and generates the probability map of face semantic labels $\mathbf{p} \in \mathbb{R}^{H\times W\times K}$:

$$\mathbf{p} = \mathcal{P}(\mathbf{y}_c), \tag{2}$$

where K is the number of semantic classes. The semantic probabilities encode the essential appearance information and approximate locations of the facial components (e.g., eyes, noses and mouths) and serve as a strong global prior for reconstructing the deblurred face image. We extract $K = 14$ semantic labels (see Table 3) for each input image.

Fine deblurring network The fine deblurring network has a similar architecture to the coarse deblurring network. In addition, we take as input the blurred image $\mathbf{x}$, coarse deblurred image $\mathbf{y}_c$, as well as the semantic probability maps $\mathbf{p}$ to recover a clear face image $\mathbf{y}$:

$$\mathbf{y} = \mathcal{G}_f(\mathbf{x}, \mathbf{y}_c, \mathbf{p}). \tag{3}$$

Our fine deblurring network also has a similar two-scale structure to the coarse deblurring network. The input to the first scale includes the $2\times$ downsampled blurred image $\mathbf{x}^{(0.5\times)}$ (3 channels), $2\times$ downsampled coarse deblurred image $\mathbf{y}_c^{(0.5\times)}$ (3 channels), and the $2\times$ downsampled semantic probability maps $\mathbf{p}^{(0.5\times)}$ (11 channels), resulting in a 17-channel input feature. The input to the second scale includes the blurred image $\mathbf{x}$ (3 channels), coarse deblurred image $\mathbf{y}_c$ (3 channels), the upsampled deblurred image from the first scale $\mathcal{U}_{2\times}(\mathbf{y}^{(0.5\times)})$ (3 channels), and the semantic probability maps $\mathbf{p}^{(0.5\times)}$ (11 channels), resulting in a 20-channel input feature. The output image from the second scale of the fine deblurring network is the final deblurred result $\mathbf{y}$. Figure 3 shows an overview of our face parsing and deblurring network.

3.3 Loss Functions

We train the parsing network using a cross-entropy loss and optimize the deblurring networks with a pixel-wise content loss and the proposed adaptive local structural loss. As pixel-wise L_2 or L_1 loss functions typically lead to overly-smooth

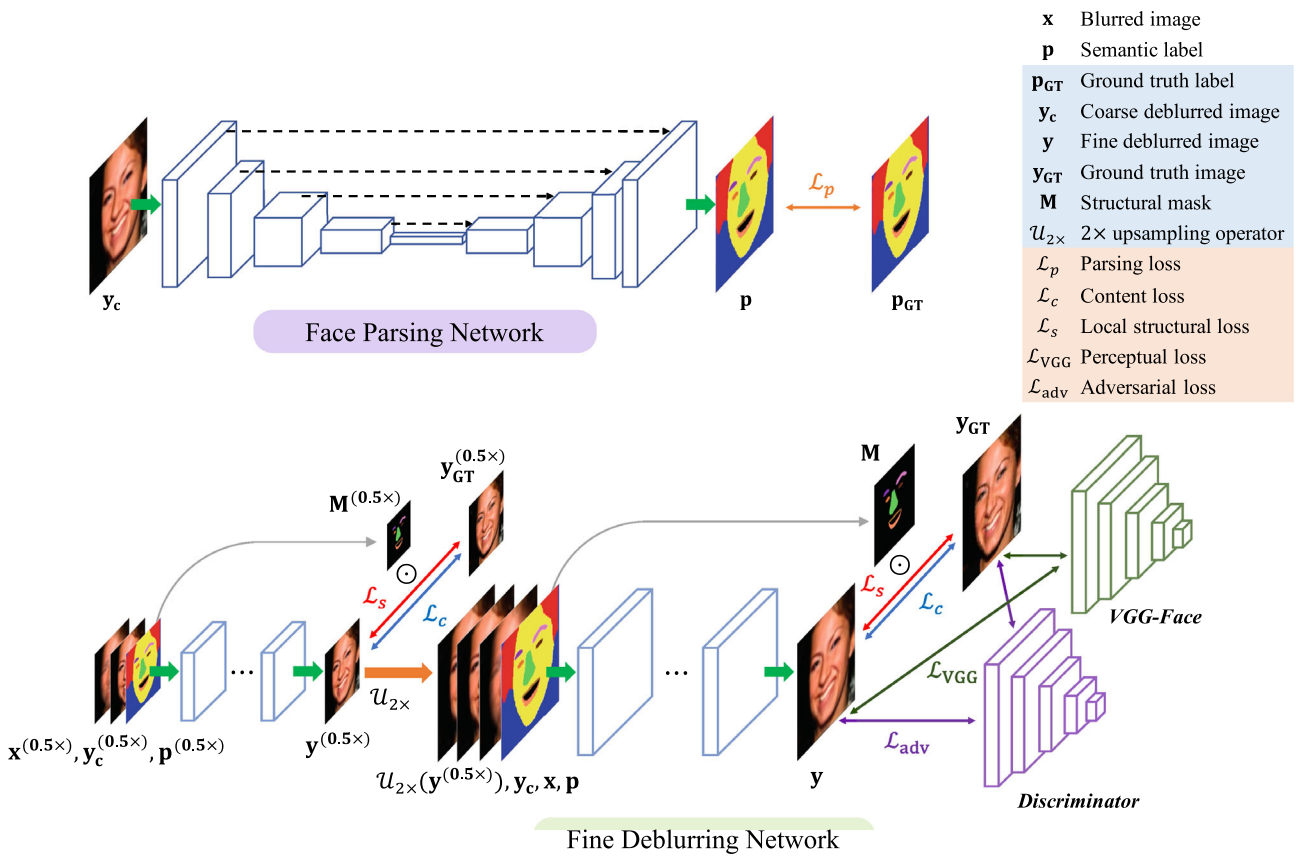


Fig. 3 Architecture of the proposed model. The face parsing network is an encoder–decoder architecture with skip connections from the encoder to the decoder. The fine deblurring network has two scales. The first scale generates a deblurred image with 0.5× spatial resolution, and the second scale generates a full-resolution deblurred image. Each scale of the deblurring network receives the supervision from the pixel-wise content loss and local structural loss. In addition, we impose the perceptual and adversarial losses at the output of the second scale. The coarse deblurring network has a similar architecture to the fine deblurring network but without taking the semantic label as input and only receiving supervision from the content loss

results, we further introduce a perceptual loss (Johnson et al. 2016) and an adversarial loss (Goodfellow et al. 2014) to optimize our deblurring network and generate photo-realistic deblurred results.

Parsing loss We adopt a multi-class cross-entropy loss function to optimize the face parsing network:

$$\mathcal{L}_p = -\sum_{k=1}^{K} \mathbf{p}_{\mathrm{GT}}^{(k)} \log(\mathbf{p}^{(k)}), \tag{4}$$

where $\mathbf{p}_{\mathrm{GT}}^{(k)}$ is the ground truth semantic label for the kth class.

Content loss We adopt the pixel-wise L_1 robust function as the content loss of the coarse and fine deblurring networks:

$$\mathcal{L}_c = \|\mathbf{y}_c - \mathbf{y}_{\mathrm{GT}}\|_1 + \|\mathbf{y} - \mathbf{y}_{\mathrm{GT}}\|_1 . \tag{5}$$

Adaptive local structural loss While the content loss (5) enforces a holistic supervision from the ground truth clear image, the key components (e.g., eyes, lips, and mouths) on faces may be easily ignored as they are typically thin and small. Solely minimizing the content loss on the whole face image cannot guarantee to restore the fine details. Thus, we propose to impose a local structural loss on facial key components:

$$\mathcal{L}_s = \sum_{k=1}^{K} w_k \|\mathbf{M}_k \odot \mathbf{y} - \mathbf{M}_k \odot \mathbf{y}_{GT}\|_1 , \tag{6}$$

where w_k is the weight of each component and $\mathbf{M}_k$ denotes the structural mask of the kth component (extracted from the semantic label $\mathbf{p}$). We apply the local structural losses on eight important components, including left eye, right eye, left eyebrow, right eyebrow, nose, upper lip, lower lip, and teeth, to enhance the local details. We do not apply the local structural loss on textureless regions, such as hair and skin. The local structural losses enforce the deblurring network to restore more details with fewer artifacts on the face images.

In our preliminary work (Shen et al. 2018), we adopt an *equal* weight for all the selected components, i.e., $w_k = 1, \forall k = 1, \ldots, K$. However, tiny components (e.g., eyes) may not be well reconstructed when optimizing the network. In this work, we propose an *adaptive* weighting mechanism based on the size of each component:

$$w_k = c/A_k, \tag{7}$$

where c is a constant and A_k is the size of the kth component. The adaptive local structural loss enforces larger weights on small components and thus helps recover facial details.

Perceptual loss The perceptual loss has been adopted in style transfer (Gatys et al. 2015; Johnson et al. 2016), image super-resolution (Ledig et al. 2017) and image synthesis (Chen and Koltun 2017; Wang et al. 2018b). The perceptual loss aims to measure the similarity in the high dimensional feature space of a pre-trained classification network [e.g., VGG16 (Simonyan and Zisserman 2015)]. Given the input image x, we denote $\phi_l(x)$ as the activation at the l-th layer of the loss network ϕ. The perceptual loss is then defined as:

$$\mathcal{L}_{\text{VGG}} = \sum_l \|\phi_l(\mathbf{y}) - \phi_l(\mathbf{y}_{\text{GT}})\|_1 . \tag{8}$$

We compute the perceptual loss on the `pool2` and `pool5` layers of the pre-trained VGG-Face (Parkhi et al. 2015).

Adversarial loss The adversarial training framework has been effectively applied to synthesize realistic images (Goodfellow et al. 2014; Ledig et al. 2017; Nah et al. 2017). We treat our fine deblurring network as the generator and construct a discriminator based on the DCGAN (Radford et al. 2016) model. The goal of the discriminator $\mathcal{D}$ is to distinguish the real image from the output of the generator. The generator $\mathcal{G}$ aims to generate images as real as possible to fool the discriminator. The adversarial training is formulated as solving the following min-max problem:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}\left[\log \mathcal{D}(\mathbf{y}_{\text{GT}})\right] + \mathbb{E}\left[\log(1 - \mathcal{D}(\mathbf{y}))\right]. \tag{9}$$

When updating the generator, the adversarial loss is:

$$\mathcal{L}_{\text{adv}} = -\log \mathcal{D}(\mathbf{y}). \tag{10}$$

Our discriminator takes an input image with of 128×128 pixels and has 6 strided convolutional layers followed by the ReLU activation function. In the last layer, we use the sigmoid function to output a single scalar as the probability of being a real image. Similar to existing image super-resolution (Ledig et al. 2017) and motion deblurring (Kupyn et al. 2018; Nah et al. 2017) methods, the generator of the proposed model does not take a noise vector as input.

Overall loss function The overall loss function for training our face deblurring model is:

$$\mathcal{L} = \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \tag{11}$$

where λ_s, λ_p, λ_{VGG}, and λ_{adv} are the weights to balance the local structural losses, parsing loss, perceptual loss and adversarial loss, respectively. In this work, we empirically set the weights to $\lambda_s = 50$, $\lambda_p = 1e^{-4}$, $\lambda_{\text{VGG}} = 1e^{-5}$, $\lambda_{\text{adv}} = 5e^{-5}$, and the constant $c = 1$ in (7). We adopt the content and local structural losses at all scales of the deblurring network while only apply the perceptual and adversarial losses to the final output image, i.e., the output of second scale from the fine deblurring network.

3.4 Training Strategy

As our model consists of three sub-networks, it is difficult to jointly optimize the whole model simultaneously. We adopt the following progressive training strategy:

1. We first train the coarse deblurring network $\mathcal{G}_c$ using the content loss (5) on the coarse deblurred image for 200,000 iterations.
2. We then fix $\mathcal{G}_c$ and train the face parsing network $\mathcal{P}$ using the parsing loss (4) for 60,000 iterations.
3. Next, we fix both $\mathcal{G}_c$ and $\mathcal{P}$ and train the fine deblurring network $\mathcal{G}_f$ using the content loss (5), local structural loss (6), perceptual loss (8) and adversarial loss (9) for 200,000 iterations.
4. Finally, we jointly optimize all three sub-networks by minimizing the overall loss (11) for 100,000 iterations.

We demonstrate that such a progressive training strategy can achieve better performance then jointly training the whole model from scratch in Sect. 4.

3.5 Implementation Details

Both the coarse and fine deblurring networks have two scales, where each scale has 6 ResBlock (He et al. 2016) (include two convolutional layers and one activation layer) and 18 convolutional layers. The first convolutional layer at each scale has a kernel size of 11×11, while all other convolutional layers have a kernel size of 5×5 and 64 channels. The upsampling layer uses a 4×4 transposed convolutional layer to upsample the image by $2\times$. We use the ReLU as the activation function and do not use any normalization layer (e.g., batch normalization).

We implement our network using the MatConvNet toolbox (Vedaldi and Lenc 2015). We use a batch size of 16 and set the learning rate to $5e^{-6}$ for the parsing network and $4e^{-5}$

Table 1 Summary of our face deblurring dataset

	Clear images			Blur	Blurred
	Helen	CMU PIE	CelebA	Kernels	Images
Training	2000	2164	2300	20,000	130 M
Testing	100	–	100	80	16,000

We collect clear face images from the Helen (Le et al. 2012), CMU PIE (Sim et al. 2002), and CelebA (Liu et al. 2015) datasets and synthesize blur kernels for generating blurred face images

Table 2 Effect of semantic labels on face image deblurring

Model	PSNR	SSIM	F-score
$\mathcal{G}$	24.85	0.849	N.A.
$\mathbf{p}_{\text{GT}} + \mathcal{G}$	25.85	0.866	1.0
$\mathcal{P}$ (fixed) + $\mathcal{G}$	25.32	0.857	0.615
$\mathcal{G}_c$ (fixed) + $\mathcal{P}$ (fixed) + $\mathcal{G}_f$	25.48	0.860	0.628

We evaluate the average labeling accuracy (i.e., F-score), PSNR and SSIM of the deblurred images on the Helen dataset

for the coarse and fine deblurring networks. During the training process, we apply the following data augmentation: (1) random scaling between $[0.9, 1.1]\times$, (2) random horizontal and vertical shifting within 12 pixels, and (3) random rotating within $\pm 30°$. The whole training process takes about 5 days on an NVIDIA Titan X GPU card.

3.6 Face Deblurring Datasets

We collect clear face images from the Helen (Le et al. 2012), CMU PIE (Sim et al. 2002), and CelebA (Liu et al. 2015) datasets. We align all the face images by first detecting the facial landmarks using the method of Sun et al. (2013b) and warping the images based on the aligned landmarks (Kae et al. 2013). The motion blur kernels are synthesized by modeling random 3D camera trajectories (Boracchi and Foi 2012). We generate blur kernels with 8 different sizes (from 13×13 to 27×27). By convolving the clear images with blur kernels and adding Gaussian noise with $\sigma = 0.01$, we obtain 130 million blurred images for training and 16,000 blurred images for testing. Table 1 summarizes the number of clear face images, motion blur kernels, and synthesized blurred images in the training and testing sets. We note that the 20,000 blur kernels used to generate training images are different from the 80 blur kernels used in the test set. Both the clear faces images and blur kernels are disjoint in the training and testing sets.

4 Analysis and Discussions

In this section, we first demonstrate the effectiveness of using semantic parsing labels for face image deblurring. We then conduct ablation studies to analyze the contribution of each sub-network and loss function.

4.1 Effect of Semantic Parsing

Our key idea is to utilize the face semantic labels as prior information to facilitate the face deblurring. We first validate the idea by using the *ground truth* semantic labels as an additional input to our deblurring network. Since only the Helen dataset contains ground truth face labels, we first train a face parsing network using the clear images and ground truth from the Helen dataset. We then use this face parsing network to generate labels for the clear images in the CMU PIE and CelebA datasets, which are treated as the *pseudo ground-truth* labels to train the proposed face parsing network for deblurring.

We train a baseline model $\mathcal{G}$ using the coarse deblurring network, which does not take any semantic information as input and does not adopt the local structural loss. Then, we concatenate the ground truth semantic labels $\mathbf{p}_{\text{GT}}$ with the blurred image as input to the baseline model. We evaluate the PSNR and SSIM on the Helen test set and present the results in Table 2. The model with prior knowledge from the ground truth labels (2nd row) significantly outperforms the baseline model (1st row), which demonstrates the effect of semantic labels on deblurring face images.

In Shen et al. (2018), the semantic labels are extracted from the *blurred* images. While the parsing network $\mathcal{P}$ is fine-tuned on blurred images for performance gain, the semantic labels of some small components (e.g., eyebrows, lips, and teeth) may not be accurate enough when the input image suffers from large motion blur. In the proposed method, we first apply a coarse deblurring network $\mathcal{G}_c$ to reduce the motion blur and recover a rough structure of the input face image. We then fine-tune the parsing network $\mathcal{P}$ on the coarse deblurred images and train the fine deblurring network $\mathcal{G}_f$ using the labels extracted from the *coarse deblurred* images. Table 2 shows the performance difference between the method of Shen et al. (2018) (3rd row) and the proposed model (4th row). The proposed method achieves higher label accuracy and obtains better deblurring results. We note that we only use the content loss (5) to train the models in Table 2. We also fix the coarse deblurring network and parsing network when training the fine deblurring network to rule out the influence of model parameters.

Figure 4 shows the deblurred images by the models listed in Table 2. Table 3 shows the parsing accuracy (in terms of the F-score) of each component, and Fig. 5 visualizes the parsing results. It is clear that more accurate semantic labels provide stronger priors to achieve better deblurring results.

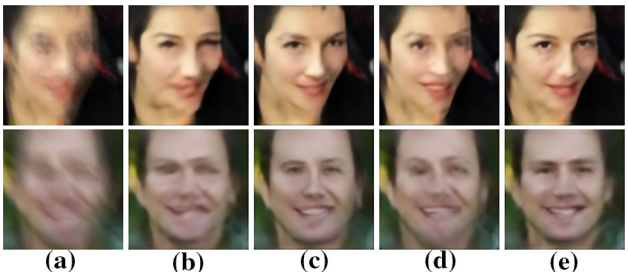


Fig. 4 Deblurred results using different semantic labels. **a** Blurred images, **b** baseline (w/o semantic labels), **c** using ground truth semantic labels, **d** using labels from blurred images, **e** using labels from coarse deblurred images

Table 3 Performance of face parsing network

Input image	Clear	Blurred	Deblurred
Face	0.915	0.886	0.881
Left eyebrow	0.733	0.587	0.640
Right eyebrow	0.721	0.596	0.642
Left eye	0.741	0.679	0.655
Right eye	0.774	0.601	0.665
Nose	0.899	0.864	0.872
Upper lip	0.653	0.477	0.502
Lower lip	0.733	0.632	0.625
Teeth	0.397	0.325	0.337
Hair	0.566	0.499	0.466
Average	0.713	0.615	0.628

We measure the F-score for each facial component on the Helen dataset

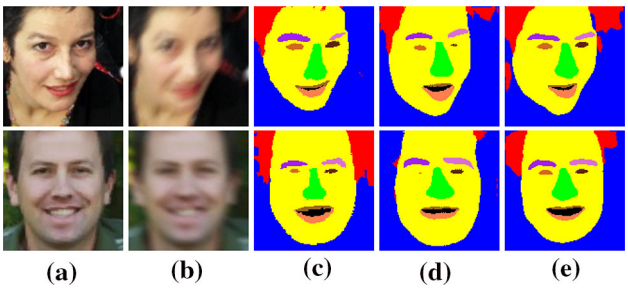


Fig. 5 Labeling results of face parsing network. **a** Ground truth clear images, **b** input blurred images, **c** ground truth semantic labels, **d** semantic labels from blurred images, **e** semantic labels from coarse deblurred images

4.2 Ablation Study

In this section, we analyze the contribution of loss functions, training strategy, and several design choices of the proposed model, including the kernel size, multi-stage deblurring, and effective range of hyper-parameters.

Local structural loss Shen et al. (2018) adopt an *equal* weight in the local structural loss $\mathcal{L}_s$ for all the key components, while we apply *adaptive* weights based on the size of each component. Here we train our fine deblurring network (freezing the coarse deblurring network and parsing network) using the content loss as well as local structural loss, and present the results in Table 4. We note that the model trained solely on the content loss $\mathcal{L}_c$ considers all the pixels, including hair, skin, and background, equally. The equal-weight local structural loss significantly improves the performance by encouraging the network to enhance details on eight key components, including left eye, right eye, left eyebrow, right eyebrow, nose, upper lip, lower lip, and teeth. The proposed adaptive local structural loss further adjusts the weights by considering the size of key components to prevent the model from sacrificing some tiny components, e.g., lips and teeth. Figure 6 shows the deblurred results by the models listed in Table 4.

Table 4 Analysis on loss functions

Losses	Helen		CelebA	
	PSNR	SSIM	PSNR	SSIM
$\mathcal{L}_c$	25.48	0.860	24.51	0.868
$\mathcal{L}_c$ + equal-weight $\mathcal{L}_s$	25.72	0.863	24.72	0.869
$\mathcal{L}_c$ + adaptive $\mathcal{L}_s$	25.80	0.866	24.86	0.874

We fix the parsing network and coarse deblurring network and train the fine deblurring network using the content loss $\mathcal{L}_c$ and local structural loss $\mathcal{L}_s$

Fig. 6 Effects of loss functions. **a** Ground truth images, **b** blurred images, **c** $\mathcal{L}_c$, **d** $\mathcal{L}_c$ + equal-weight $\mathcal{L}_s$, **e** $\mathcal{L}_c$ + adaptive $\mathcal{L}_s$

Training strategy Since the proposed model consists of three sub-networks, the cascade of all sub-networks becomes a very deep model. As such, it is not easy to training such a deep model from scratch. The last row of Table 5 shows that the model trained from scratch does not perform well. Thus, we train our model stage-by-stage using the training strategy described in Sect. 3.4. We show the evaluation results of each stage in Table 5. With the proposed training strategy, our model gradually achieves better performance. Figure 7 shows the deblurred results of the models listed in Table 5. The model using the progressive training strategy recovers

Table 5 Analysis on training strategy

Model	Helen		CelebA	
	PSNR	SSIM	PSNR	SSIM
$\mathcal{G}_c$	25.26	0.855	24.58	0.869
$\mathcal{G}_c$ (fixed) + $\mathcal{P}$ (fixed) + $\mathcal{G}_f$	25.80	0.866	24.86	0.874
$\mathcal{G}_c + \mathcal{P} + \mathcal{G}_f$ (fine-tuned)	25.92	0.868	24.89	0.875
$\mathcal{G}_c + \mathcal{P} + \mathcal{G}_f$ (scratch)	24.74	0.845	24.08	0.860

We progressively train the coarse deblurring network $\mathcal{G}_c$, face parsing network $\mathcal{P}$, and the fine deblurring network $\mathcal{G}_f$. Finally, we jointly fine-tune all three sub-networks. The proposed training strategy achieves better performance than the training the whole model from scratch

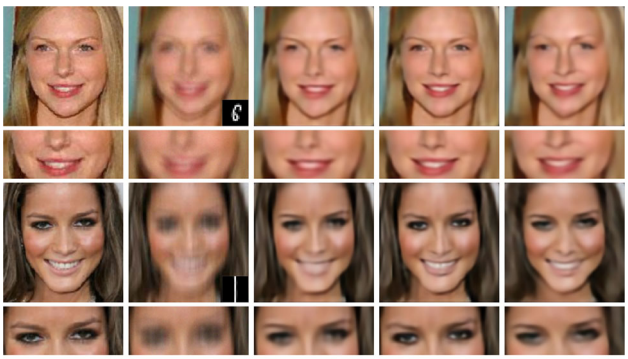


Fig. 7 Effects of training strategy. **a** Ground truth images, **b** blurred images, **c** $\mathcal{G}_c$, **d** $\mathcal{G}_c + \mathcal{P} + \mathcal{G}_f$ (fine-tuned), **e** $\mathcal{G}_c + \mathcal{P} + \mathcal{G}_f$ (from scratch)

Table 6 Analysis on perceptual and adversarial losses

$\mathcal{L}_{\text{VGG}}$	$\mathcal{L}_{\text{adv}}$	Helen		CelebA	
		PSNR	SSIM	PSNR	SSIM
		25.92	0.868	24.89	0.875
✓		26.28	0.877	25.14	0.881
✓	✓	26.34	0.876	25.33	0.881

Both perceptual and adversarial losses further improve the performance by restoring more faithful details

better content and more facial details than the model trained from scratch.

Perceptual and adversarial losses We compare the deblurring results with and without using the perceptual and adversarial losses in Table 6 and Fig. 8. The perceptual loss encourages the images to match the high-level activations of the VGG-Face network and makes the output look more photo-realistic. The adversarial loss further introduces more details on hairs and beards, which cannot be reconstructed well using the pixel-wise L_2 or L_1 loss. As shown in Table 6, both the perceptual and adversarial losses improve the average PSNR and SSIM on both test sets as more faithful details are recovered.

Kernel size We use a larger kernel size at the first convolutional layer of our coarse and fine deblurring networks. Here

Fig. 8 Effects of perceptual and adversarial functions. **a** Ground truth images, **b** blurred images, **c** ours w/o $\mathcal{L}_{\text{VGG}}$ and $\mathcal{L}_{\text{adv}}$, **d** ours w/ $\mathcal{L}_{\text{VGG}}$, **e** ours w/ $\mathcal{L}_{\text{VGG}}$ and $\mathcal{L}_{\text{adv}}$

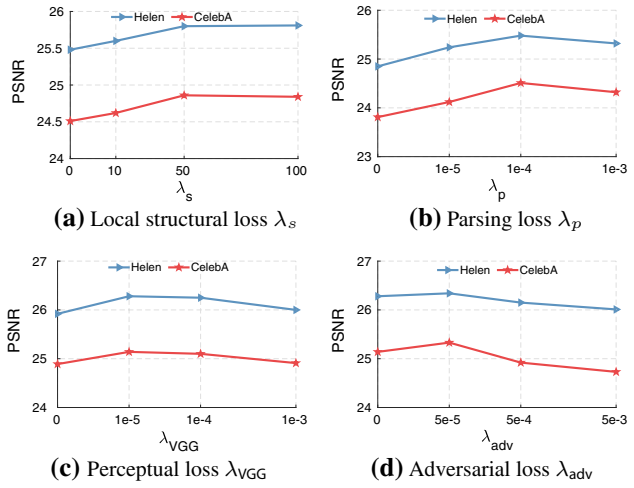
Table 7 Analysis on kernel size

Kernel size	Helen		CelebA		#Parameters
	PSNR	SSIM	PSNR	SSIM	
5×5	25.63	0.860	24.65	0.868	14.56 M
9×9	25.75	0.864	24.72	0.868	14.70 M
11×11	25.80	0.866	24.86	0.874	14.80 M
13×13	25.80	0.867	24.87	0.876	14.92 M

We evaluate the model performance by changing the kernel size at the first convolutional layer

we evaluate the performance of the proposed model with different kernel sizes in Table 7. Consistent performance gain can be achieved when using a larger filter size up to the kernel of 11×11 pixels. Therefore, we choose to use 11×11 filter at the first convolutional layer to have a larger receptive field for the whole model. In addition, as the first convolutional layer only contains 64 feature channels, using a larger filter size does not significantly increase the number of model parameters.

Hyper-parameters We analyze the effective range of the hyper-parameters, λ_s, λ_p, λ_{VGG}, and λ_{adv} by changing one of the hyper-parameters and fixing the others. In Fig. 9a, we show that the local structural loss effectively improves the PSNR but saturates at $\lambda_s = 50$. As shown in Fig. 9b, without the parsing loss (i.e., $\lambda_p = 0$), the face parsing network cannot learn meaningful semantic labels as the facial priors. However, a larger λ_p does not further improve the face restoration performance as the gradient of the parsing loss is back-propagated to the coarse deblurring network, which may introduce additional artifacts. Therefore, setting $\lambda_p = 1e-4$ achieves a good balance for the whole model. In Fig. 9c, we show that the proposed model obtains plausible results when choosing $1e-5 \leq \lambda_p \leq 1e-4$. Using a larger weight for the perceptual loss introduces more checkerboard artifacts and harms the restoration per-



(a) Local structural loss λ_s **(b)** Parsing loss λ_p

(c) Perceptual loss λ_{VGG} **(d)** Adversarial loss λ_{adv}

Fig. 9 Effective range of hyper-parameters. We plot the average PSNR on both the CelebA and Helen datasets

formance. Finally, in Fig. 9d, we show that using a smaller weight for the adversarial loss, i.e., $\lambda_{adv} \leq 1e-4$, does not affect the PSNR too much. However, the model can generate more facial details to improve visual quality. When increasing λ_{adv}, the model generates noise-like artifacts, resulting in a performance drop. Therefore, we choose $\lambda_{adv} = 5e-5$.

Multi-stage deblurring Due to our architecture design, we are able to extend the proposed model by cascading multiple fine deblurring networks. Here we construct our model with one coarse deblurring network, one face parsing network, and N fine deblurring networks. The Nth fine deblurring network takes as input the blurred image, deblurred image from the $N-1$th fine deblurring network, and the semantic labels from the face parsing network. We compare the performance, model parameters, and execution time in Table 8. The performance of our model saturates at $N = 2$. When using three fine deblurring networks, the model only slightly improves the performance but uses 160% more parameters and runs $1.6\times$ slower than the model with $N = 1$.

In the last two rows of Table 8, we show the performance of the proposed model by sharing the weight of the fine deblurring network. The experimental results show that the models with shared weights do not perform well. As the fine deblurring network is already a deep sub-network, sharing the weight of a large sub-module is not guaranteed to improve the performance. Instead, sharing the weight of a single convolutional layer or a small block (e.g., a residual block) might be a more reasonable way to design a recurrent structure. As our goal is to utilize the semantic labels for face deblurring instead of exploring a better network architecture, we leave this issue as future work. Overall, the proposed model with a single fine deblurring network already achieves state-of-the-art performance.

5 Evaluation Against with the State-of-the-Art Methods

In this section, we present evaluations against the state-of-the-art deblurring approaches in terms of the restoration quality, face detection, face recognition, and execution time. We also provide visual comparisons on synthetic datasets and real blurred images. Finally, we discuss the limitation and failure cases of the proposed method.

5.1 Restoration Quality

We compare the proposed method with the state-of-the-art deblurring algorithms, including MAP-based methods (Cho and Lee 2009; Krishnan et al. 2011; Shan et al. 2008; Xu et al. 2013; Zhong et al. 2013; Pan et al. 2014, 2017a; Li et al. 2018) and CNN-based methods (Nah et al. 2017; Tao et al. 2018; Kupyn et al. 2018; Jin et al. 2018; Shen et al. 2018). We evaluate all the algorithms on both the Helen and CelebA test sets. Table 9 presents the average PSNR and SSIM for different sizes of blur kernels, and Table 10 shows the average and the worst PSNR/SSIM on the entire datasets for each method. We note that the optimization-based methods (Shan et al., 2008; Cho and Lee, 2009; Krishnan et al., 2011; Xu et al., 2013; Zhong et al., 2013; Pan et al., 2014, 2017a) may generate severe visual artifacts when the blur kernel is not estimated well and achieve significant lower PSNR/SSIM values. The proposed method performs favorably against existing deblurring approaches and our preliminary method (Shen et al. 2018) on both datasets.

We show the results of the Helen dataset in Fig. 10 and the CelebA dataset in Fig. 11. Conventional MAP-based approaches (Cho and Lee 2009; Krishnan et al. 2011; Shan et al. 2008; Xu et al. 2013; Zhong et al. 2013; Pan et al. 2017a) do not estimate blur kernels well and therefore generate more ringing artifacts. The face deblurring approach (Pan et al. 2014) is not robust to noise and the performance depends heavily on the similarity of the reference image. There are several ringing artifacts in the deblurred images by Pan et al. (2014). The method of Li et al. (2018) generates sharp debluured images, but the faces do not look realistic. The CNN-based methods (Nah et al. 2017; Kupyn et al. 2018; Tao et al. 2018; Jin et al. 2018) do not consider the face semantic information and thus cannot effectively reduce the motion blur.

Both the method by Shen et al. (2018) and the proposed model obtain visually pleasing results. However, the method by Shen et al. (2018) is not robust to the error on semantic labels (which is predicted from blurred images) and less effective in restoring facial details (e.g., the mouth of the first and second rows in Fig. 11). In contrast, the proposed method extracts more accurate semantic priors and restores

Fig. 10 Visual comparison on Helen dataset. The results from the proposed method contain fewer visual artifacts and more details on key face components (e.g., eyes and mouths)

better facial structures and details (e.g., the eyes of the second and third rows in Fig. 10).

In Fig. 12, we show the deblurring results from images with specific attributes, such as occlusion, mustaches, saturation, and people with different skin colors. As our test set does not contain images with significant saturation, we adjust the intensity of the blurred images (row 5 and 6 of Fig. 12) by multiplying the Y-channel by $1.5\times$. The proposed method can still recover more facial details than existing approaches from such an input. Overall, our method performs well in real-world scenarios.

5.2 Face Recognition

We also demonstrate the performance of the proposed method by evaluating the face identity distance, face detection, and recognition accuracy.

Identity distance We use the FaceNet (Schroff et al. 2015) to extract face features and compute the identity distance with the L_2 loss and cosine loss (Wang et al. 2018a) between the ground truth image and deblurred image. Figure 13 shows that the deblurred images from the proposed method have the lowest identity distance on both measurements, which demonstrates that the proposed method preserves the face identity well.

Face detection We use the OpenFace toolbox (Amos et al. 2016) to detect the face for each image in the CelebA test set. We show the success rate of the face detection for blurred images and the state-of-the-art deblurring approaches in Table 11. The clear images have a success rate of 96%, while the success rate on blurred images drops to 77.4% due to motion blur. The deblurred images from some of the evaluated methods have an lower success rate as the images contain severe ringing artifacts. In contrast, the proposed method has 95.3% success rate, which is close to the upper bound of the clear images.

Face recognition As the CelebA dataset contains identity labels, we conduct another experiment on the identity recognition. We consider our CelebA test images as a probe set, which has 100 different identities. For each identity, we collect additional 9 clear face images as a gallery set. For each image in the probe set, our goal is to find the most similar face image from the gallery set and identify whether they belong to the same identity (Fig. 14).

Table 8 Multi-stage deblurring

#Stages	Helen		CelebA		#Parameters	Time (s)
	PSNR	SSIM	PSNR	SSIM		
1	25.80	0.866	24.86	0.874	14.80 M	0.08
2	25.87	0.869	24.88	0.878	26.66 M	0.11
3	25.89	0.866	24.86	0.875	38.52 M	0.13
2 (shared)	25.78	0.864	24.74	0.870	14.80 M	0.11
3 (shared)	25.74	0.861	24.76	0.871	14.80 M	0.13

We apply the fine deblurring network for multiple times and compare the restoration performance, model parameters, and execution time

Table 9 Quantitative comparison with the state-of-the-art methods

Method	13 × 13		15 × 15		17 × 17		19 × 19		21 × 21		23 × 23		25 × 25		27 × 27	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Helen																
Shan et al. (2008)	20.65	0.726	20.01	0.693	20.66	0.718	19.16	0.653	19.75	0.675	19.07	0.645	19.51	0.665	17.79	0.587
Cho and Lee (2009)	17.46	0.610	17.16	0.590	17.58	0.613	16.58	0.563	16.82	0.574	16.52	0.554	16.88	0.575	15.58	0.509
Krishnan et al. (2011)	18.93	0.674	19.47	0.685	20.10	0.707	18.99	0.660	19.98	0.693	19.02	0.648	19.75	0.682	18.12	0.608
Xu et al. (2013)	20.85	0.744	20.51	0.730	21.12	0.751	19.43	0.686	20.38	0.723	19.65	0.691	20.24	0.714	18.69	0.648
Zhong et al. (2013)	17.33	0.646	16.83	0.631	16.78	0.629	16.38	0.613	15.96	0.598	16.07	0.600	16.02	0.603	15.98	0.593
Pan et al. (2017a)	23.00	0.797	22.01	0.769	22.87	0.792	20.35	0.711	20.99	0.733	19.90	0.691	20.32	0.708	18.04	0.617
Pan et al. (2014)	19.74	0.682	16.88	0.680	20.37	0.695	18.83	0.632	19.32	0.650	18.43	0.604	19.14	0.634	16.59	0.528
Nah et al. (2017)	26.02	0.878	25.43	0.858	25.84	0.862	23.37	0.805	24.45	0.835	23.10	0.793	23.80	0.815	20.98	0.734
Tao et al. (2018)	24.51	0.834	24.23	0.811	23.98	0.798	22.04	0.735	23.12	0.775	22.04	0.728	22.78	0.756	20.25	0.661
Li et al. (2018)	22.94	0.811	22.25	0.778	23.28	0.813	21.00	0.726	21.51	0.751	20.33	0.697	20.86	0.727	18.14	0.594
Kupyn et al. (2018)	25.91	0.861	24.88	0.828	25.12	0.828	23.36	0.776	23.43	0.779	22.65	0.743	23.07	0.763	20.64	0.670
Jin et al. (2018)	26.75	0.897	26.09	0.880	*27.03*	*0.894*	24.64	0.847	25.38	0.863	*24.89*	*0.854*	*25.21*	*0.859*	*22.76*	*0.796*
Shen et al. (2018)	*27.15*	*0.896*	*26.43*	*0.882*	26.74	0.885	*25.41*	*0.859*	*25.61*	*0.864*	24.78	0.843	25.18	0.854	22.52	0.788
Ours	**27.58**	**0.905**	**26.87**	**0.891**	**27.31**	**0.895**	**26.00**	**0.872**	**25.99**	**0.871**	**25.14**	**0.853**	**25.48**	**0.861**	**22.96**	**0.799**
CelebA																
Shan et al. (2008)	19.38	0.685	18.85	0.663	19.25	0.675	18.05	0.629	18.62	0.653	17.98	0.624	18.50	0.646	16.83	0.579
Cho and Lee (2009)	13.72	0.481	13.41	0.467	13.46	0.472	13.04	0.449	12.87	0.438	12.73	0.425	12.84	0.435	12.19	0.395
Krishnan et al. (2011)	18.23	0.679	18.57	0.685	19.18	0.705	18.04	0.661	19.18	0.699	17.98	0.648	18.77	0.684	17.12	0.617
Xu et al. (2013)	19.44	0.695	19.14	0.691	19.59	0.709	18.28	0.659	19.32	0.705	18.54	0.669	19.15	0.696	17.95	0.653
Zhong et al. (2013)	18.10	0.723	17.48	0.703	17.63	0.710	17.08	0.692	16.80	0.681	17.17	0.689	16.95	0.685	16.90	0.679
Pan et al. (2017a)	21.08	0.763	19.61	0.717	20.07	0.732	18.12	0.665	18.41	0.677	17.63	0.641	17.86	0.651	15.94	0.572
Pan et al. (2014)	20.47	0.712	20.60	0.709	21.16	0.730	19.36	0.655	19.89	0.677	18.82	0.628	19.51	0.664	16.78	0.540
Nah et al. (2017)	24.23	0.879	23.49	0.862	23.87	0.863	21.90	0.820	22.65	0.839	21.53	0.806	22.12	0.824	19.71	0.759
Tao et al. (2018)	26.10	0.881	25.49	0.853	25.7	0.856	23.13	0.771	24.54	0.826	23.05	0.768	23.89	0.801	21.04	0.694
Li et al. (2018)	20.36	0.724	18.92	0.664	18.68	0.663	17.19	0.589	16.79	0.575	16.35	0.541	16.49	0.556	14.90	0.461
Kupyn et al. (2018)	24.34	0.802	23.36	0.765	23.54	0.764	22.31	0.726	22.31	0.727	21.68	0.696	22.06	0.715	20.01	0.639
Jin et al. (2018)	26.08	0.897	25.50	0.843	*26.01*	0.880	24.02	0.821	*24.79*	*0.871*	23.79	0.831	24.10	0.836	21.42	0.747
Shen et al. (2018)	*26.19*	*0.900*	*25.52*	*0.887*	25.77	*0.889*	*24.58*	*0.867*	24.61	0.868	*23.87*	*0.852*	*24.25*	*0.861*	*21.93*	*0.807*
Ours	**26.25**	**0.905**	**25.71**	**0.894**	**26.14**	**0.898**	**25.01**	**0.878**	**25.04**	**0.879**	**24.24**	**0.862**	**24.56**	**0.870**	**22.19**	**0.814**

We compute the average PSNR and SSIM on the Helen and CelebA test sets. Each dataset has 8000 blurred images synthesized from 100 clear face images and 80 blur kernels (10 blur kernels for each size). The bold and italic emphasis texts indicate the best and second best performance

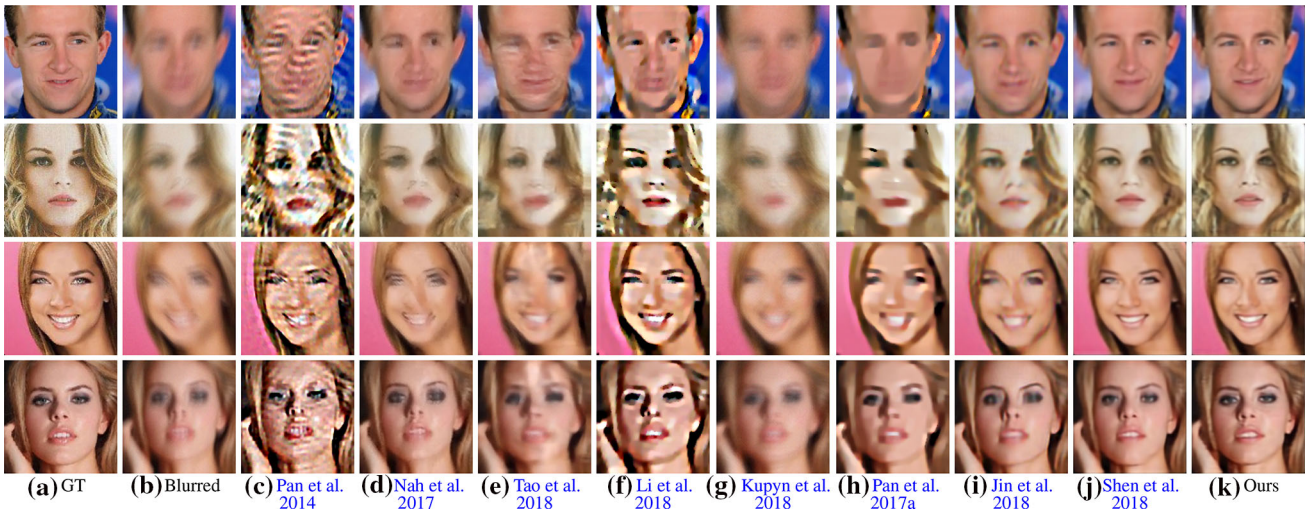


Fig. 11 Visual comparison on CelebA dataset. The results from the proposed method contain fewer visual artifacts and more details on key face components (e.g., eyes and mouths)

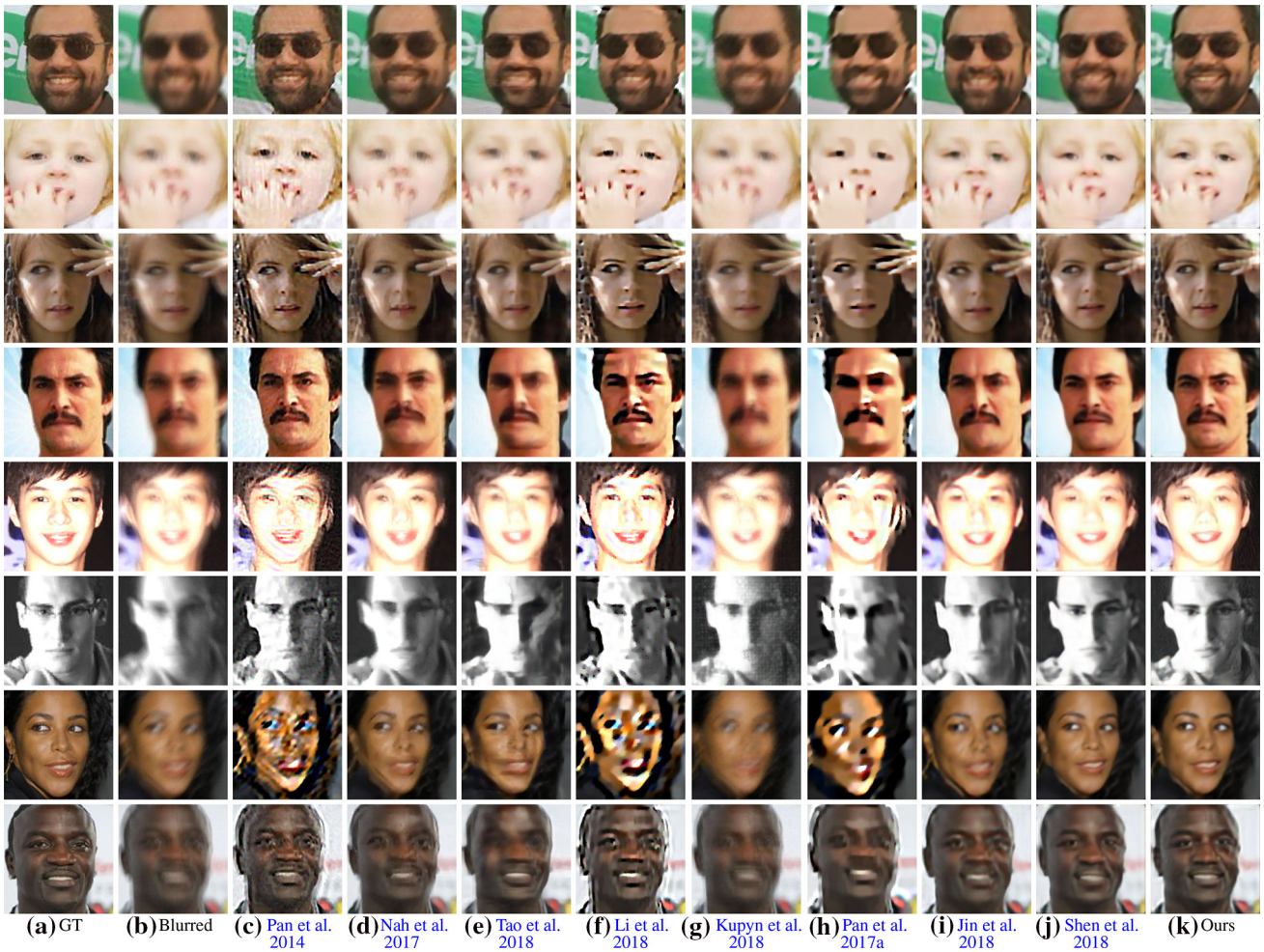


Fig. 12 Visual comparison on images with different attributes. We show that the proposed method is able to generate sharp images and robust to several scenarios, e.g., occlusion with sunglass or hands (row 1 to 3), faces with mustaches (row 1 and 4), over-exposed images (row 5 to 6), and people with different skin colors

Table 10 Quantitative comparison with the state-of-the-art methods

Method	Helen				CelebA			
	Average PSNR	Worst PSNR	Average SSIM	Worst SSIM	Average PSNR	Worst PSNR	Average SSIM	Worst SSIM
Shan et al. (2008)	19.57 ± 2.72	9.97	0.670 ± 0.137	0.109	18.43 ± 2.20	9.72	0.644 ± 0.119	0.040
Cho and Lee (2009)	16.82 ± 2.79	7.83	0.574 ± 0.126	0.215	13.03 ± 1.74	8.21	0.445 ± 0.098	0.097
Krishnan et al. (2011)	19.30 ± 3.42	5.91	0.670 ± 0.167	0.137	18.38 ± 2.74	6.90	0.672 ± 0.146	0.108
Xu et al. (2013)	20.11 ± 3.18	9.45	0.711 ± 0.147	0.075	18.93 ± 2.55	9.92	0.685 ± 0.124	0.106
Zhong et al. (2013)	16.41 ± 3.13	8.25	0.614 ± 0.142	0.140	17.26 ± 2.66	10.41	0.695 ± 0.115	0.278
Pan et al. (2014)	18.66 ± 3.95	8.82	0.677 ± 0.175	0.167	18.59 ± 3.59	9.47	0.677 ± 0.183	0.117
Pan et al. (2017a)	20.93 ± 4.27	7.21	0.727 ± 0.168	0.120	19.57 ± 4.02	8.07	0.664 ± 0.160	0.118
Nah et al. (2017)	24.12 ± 3.46	11.59	0.823 ± 0.107	0.240	22.43 ± 2.82	12.11	0.832 ± 0.103	0.277
Tao et al. (2018)	22.86 ± 3.51	11.68	0.762 ± 0.109	0.259	24.11 ± 2.67	11.21	*0.862 ± 0.091*	0.245
Li et al. (2018)	21.28 ± 3.65	9.24	0.737 ± 0.143	0.159	17.46 ± 3.39	8.86	0.596 ± 0.166	0.129
Kupyn et al. (2018)	23.63 ± 2.90	11.98	0.781 ± 0.094	0.267	22.45 ± 2.21	12.29	0.729 ± 0.080	0.283
Jin et al. (2018)	25.34 ± 3.17	11.49	*0.861 ± 0.078*	0.252	*24.46 ± 2.77*	12.54	0.841 ± 0.075	0.308
Shen et al. (2018)	*25.58 ± 2.94*	12.45	*0.861 ± 0.070*	0.403	24.34 ± 2.46	12.03	0.860 ± 0.066	0.303
Ours	**25.91 ± 2.91**	13.55	**0.869 ± 0.062**	0.480	**24.89 ± 2.32**	12.46	**0.875 ± 0.063**	0.310

We compute the average PSNR and SSIM on the Helen and CelebA test sets. The bold and italic emphasis texts indicate the best and second best performance

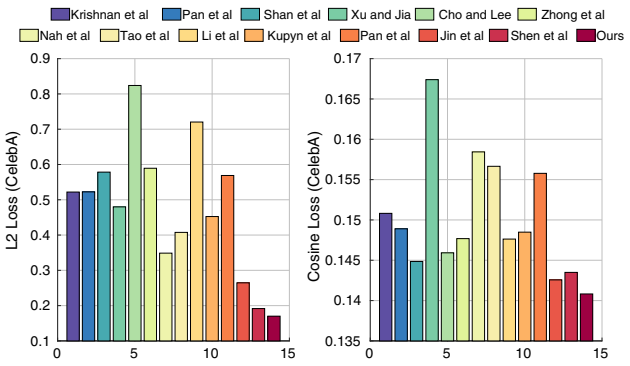


Fig. 13 Quantitative evaluation on face identity. We compute the L2 and cosine losses on the features extracted from the FaceNet (Schroff et al. 2015). The proposed method has the lowest values on the CelebA test sets

Given a blurred or deblurred image from the probe set, we compute the identity distance with all images in the gallery set and select the top-K nearest matches. Table 11 shows the top-1, top-3 and top-5 accuracy. The proposed method generates fewer artifacts and thus achieves the highest recognition accuracy against other evaluated approaches.

Table 11 Face detection and recognition on the CelebA dataset

Method	Detection (%)	Top-1 (%)	Top-3 (%)	Top-5 (%)
Clear images	96.0	74.0	86.4	90.0
Blurred images	77.4	29.1	43.4	51.3
Shan et al. (2008)	76.0	32.4	46.9	54.0
Cho and Lee (2009)	52.2	17.2	27.3	32.5
Krishnan et al. (2011)	80.0	33.8	48.9	56.6
Xu et al. (2013)	82.5	41.1	55.4	62.1
Zhong et al. (2013)	69.5	27.6	41.6	48.5
Pan et al. (2014)	78.9	42.0	55.7	62.2
Pan et al. (2017a)	74.3	40.9	48.2	58.3
Nah et al. (2017)	86.0	40.1	55.3	62.4
Tao et al. (2018)	80.5	36.8	53.6	59.8
Li et al. (2018)	78.2	40.7	49.8	56.2
Kupyn et al. (2018)	88.4	43.5	59.6	65.3
Jin et al. (2018)	89.8	42.3	60.2	67.7
Shen et al. (2018)	*94.8*	*48.3*	*63.2*	*70.0*
Ours	**95.3**	**53.8**	**68.7**	**74.2**

We show the success rate of face detection and top-1, top-3 and top-5 accuracy of face recognition. The bold and italic emphasis texts indicate the best and second best performance

5.3 Real-World Blurred Images

We evaluate the proposed method on face images collected from the real blurred dataset of Lai et al. (2016). As real images usually contain outliers that cannot be modeled well by Gaussian distributions, conventional methods fail to estimate the blur kernel and generate serious ringing artifacts. The CNN-based generic deblurring method (Nah et al. 2017) generates overly smooth results. In contrast, both the method of Shen et al. (2018) and the proposed model restore sharp and visually pleasing face images.

Fig. 14 Visual comparison on real blurred images. The proposed method generates visually pleasing deblurred results with fewer artifacts

Table 12 Comparison of execution time and model size

Method	Implementation	Seconds	Parameters
Shan et al. (2008)	C++ (CPU)	16.32	–
Cho and Lee (2009)	C++ (CPU)	0.41	–
Krishnan et al. (2011)	MATLAB (CPU)	2.52	–
Xu et al. (2013)	C++ (CPU)	0.31	–
Zhong et al. (2013)	MATLAB (CPU)	8.07	–
Pan et al. (2014)	MATLAB (CPU)	8.11	–
Pan et al. (2017a)	MATLAB (CPU)	10.55	–
Nah et al. (2017)	MATLAB (GPU)	0.09	303.6 M
Tao et al. (2018)	Python (GPU)	0.15	32.2 M
Li et al. (2018)	MATLAB (CPU)	18.53	558 K
Kupyn et al. (2018)	Python (GPU)	0.05	45.5 M
Jin et al. (2018)	Torch (GPU)	0.01	1.4 M
Shen et al. (2018)	MATLAB (GPU)	0.05	14.8 M
Ours	MATLAB (GPU)	0.08	26.6 M

We report the average execution time on 10 images with the size of 128×128

5.4 Execution Time

We evaluate the execution time of the state-of-the-art approaches and the proposed model on a machine with a 3.4 GHz Intel i7 CPU (64G RAM) and an NVIDIA Titan X GPU card (12G memory). Table 12 shows the average execution time based on 10 images with a size of 128×128. Most conventional approaches require solving several iterative optimization problems and therefore are computationally expensive. Since we use only two scales and fewer residual blocks, our model is more efficient than the model of Nah et al. (2017). The proposed model is slightly slower than the model of Shen et al. (2018) as there is an additional coarse deblurring network.

5.5 Limitations and Discussions

Our method is likely to fail in two situations. First, when the input image contains severe non-uniform blur or non-Gaussian noise, our model may not be able to reduce the blur effectively, as shown in Fig. 15. A potential solution is to synthesize more training data with complex motion models or realistic noise (Foi et al. 2008). Second, when the face cannot be well aligned (e.g., profile faces in Fig. 15 bottom), the face parsing network may not estimate accurate semantic labels to guide the deblurring network. To further analyze the performance of the proposed model on profile faces, we evaluate the face images from the FEI face database (Thomaz and Giraldi 2010), where each face is captured under different rotation angles. As shown in Fig. 16, our model performs well on frontal faces and profile faces which are rotated by about 60° (i.e., 2nd to 6th columns of Fig. 16). For extreme cases (e.g., rotated by about 90° as shown in

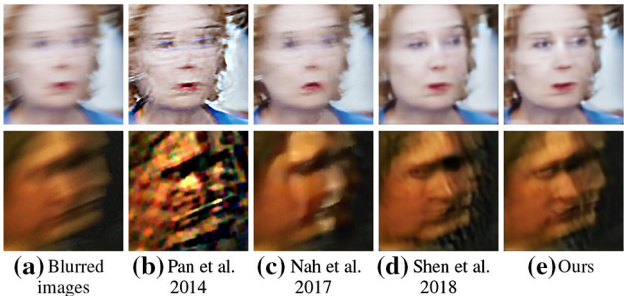


Fig. 15 Failure cases. Our method fails when the input image suffers from extremely large motion blur and the semantic labels cannot be estimated well

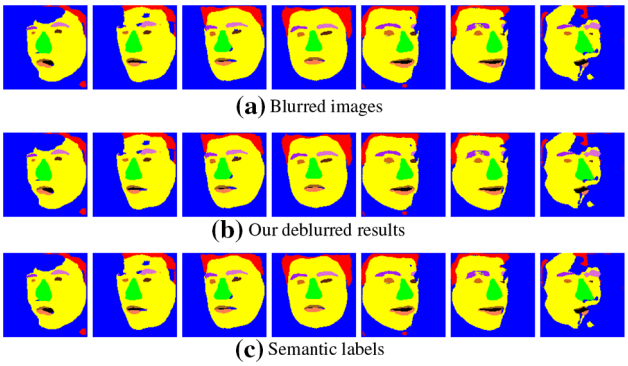


Fig. 16 Deblurring profile faces. We evaluate our model on the FEI face database (Thomaz and Giraldi 2010). The proposed model becomes less effective when a face is rotated by 90°

the 1st and 7th columns of Fig. 16), our deblurred results contain some visual artifacts around the nose and mouth. The eyes are not restored well due to the inaccurate semantic labels.

6 Conclusions

In this work, we propose a multi-scale deep convolutional neural network for face image deblurring. We exploit the face semantic information as global priors and local structural constraints to better restore the shape and detail of face images. Compared with the preliminary work (Shen et al. 2018) which obtains the semantic labels from the input blurred image, we show that the semantic information extracted from a coarse deblurred image is more accurate and leads to better performance on deblurring images. Furthermore, we propose an adaptive local structural loss to balance the weights of facial key components and restore better content and details. Experimental results on image deblurring, execution time and face recognition demonstrate that the proposed method performs favorably against our preliminary method (Shen et al. 2018) and the state-of-the-art deblurring algorithms.

Acknowledgements This work was supported by the Major Science Instrument Program of the National Natural Science Foundation of China under Grant 61527802, the General Program of National Nature Science Foundation of China under Grants 61371132 and 61471043, NSF CAREER (No. 1149783) and gifts from Adobe and Nvidia.

References

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). *Openface: A general-purpose face recognition library with mobile applications*. Technical report, CMU-CS-16-118.

Anwar, S., Phuoc Huynh, C., & Porikli, F. (2015). Class-specific image deblurring. In *IEEE international conference on computer vision*.

Boracchi, G., & Foi, A. (2012). Modeling the performance of image restoration from motion blur. *IEEE Transactions on Image Processing*, *21*(8), 3502–3517.

Chakrabarti, A. (2016). A neural approach to blind motion deblurring. In *European conference on computer vision*.

Chen, Q., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *IEEE international conference on computer vision*.

Cho, S., & Lee, S. (2009). Fast motion deblurring. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, *28*(5), 145:1–145:8.

Chrysos, G. G., Favaro, P., & Zafeiriou, S. (2019). Motion deblurring of faces. *International Journal of Computer Vision*, *127*(6–7), 801–823.

Dong, C., Deng, Y., Change Loy, C., & Tang, X. (2015). Compression artifacts reduction by a deep convolutional network. In *IEEE international conference on computer vision*.

Dong, J., Pan, J., Su, Z., & Yang, M.H. (2017). Blind image deblurring with outlier handling. In *IEEE international conference on computer vision*.

Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., & Freeman, W. T. (2006). Removing camera shake from a single photograph. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, *25*, 787–794.

Foi, A., Trimeche, M., Katkovnik, V., & Egiazarian, K. (2008). Practical Poissonian–Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, *17*, 1737–1754.

Gatys, L.A., Ecker, A.S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. In *Neural information processing systems*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Neural information processing systems*.

Hacohen, Y., Shechtman, E., Lischinski, D. (2013). Deblurring by example using dense correspondence. In IEEE *International conference on computer vision*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*.

Hirsch, M., Schuler, C. J., Harmeling, S., & Schölkopf, B. (2011). Fast removal of non-uniform camera shake. In *IEEE international conference on computer vision*.

Hradiš, M., Kotera, J., Zemcík, P., & Sroubek, F. (2015). Convolutional neural networks for direct text deblurring. In *British machine vision conference*.

Hu, Z., Cho, S., Wang, J., & Yang, M. H. (2014a). Deblurring low-light images with light streaks. In *IEEE conference on computer vision and pattern recognition*.

Hu, Z., Xu, L., & Yang, M. H. (2014b). Joint depth estimation and camera shake removal from single blurry image. In *IEEE conference on computer vision and pattern recognition*.

Jin, M., Hirsch, M., & Favaro, P. (2018). Learning face deblurring fast and wide. In *CVPR workshops* (pp. 745–753)

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*

Kae, A., Sohn, K., Lee, H., & Learned-Miller, E. G. (2013). Augmenting CRFs with Boltzmann machine shape priors for image labeling. In *IEEE conference on computer vision and pattern recognition*

Kim, J., Lee, J. K., & Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. In *IEEE conference on computer vision and pattern recognition*.

Krishnan, D., Tay, T., & Fergus, R. (2011). Blind deconvolution using a normalized sparsity measure. In *IEEE conference on computer vision and pattern recognition*.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., & Matas, J. (2018). Deblurgan: Blind motion deblurring using conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition* (pp. 8183–8192).

Lai, W. S., Ding, J. J., Lin, Y. Y., & Chuang, Y. Y. (2015). Blur kernel estimation using normalized color-line prior. In *IEEE conference on computer vision and pattern recognition*.

Lai, W. S., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Deep Laplacian pyramid networks for fast and accurate super-resolution. In *IEEE conference on computer vision and pattern recognition*.

Lai, W. S., Huang, J. B., Hu, Z., Ahuja, N., & Yang, M. H. (2016). A comparative study for single image blind deblurring. In *IEEE conference on computer vision and pattern recognition*.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. In *European conference on computer vision*.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE conference on computer vision and pattern recognition*.

Levin, A., Weiss, Y., Durand, F., & Freeman, W. T. (2009). Understanding and evaluating blind deconvolution algorithms. In *IEEE conference on computer vision and pattern recognition*.

Li, L., Pan, J., Lai, W. S., Gao, C., Sang, N., & Yang, M. H. (2018). Learning a discriminative prior for blind image deblurring. In *IEEE conference on computer vision and pattern recognition*.

Liu, Y., Dong, W., Gong, D., Zhang, L., & Shi, Q. (2018). Deblurring natural image using super-Gaussian fields. In *European conference on computer vision* (pp. 467–484).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *IEEE international conference on computer vision*.

Mao, X., Shen, C., & Yang, Y. B. (2016). Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections. In *Neural information processing systems*.

Michaeli, T., & Irani, M. (2014). Blind deblurring using internal patch recurrence. In *European conference on computer vision*.

Nah, S., Hyun Kim, T., & Mu Lee, K. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE conference on computer vision and pattern recognition*.

Nimisha, T. M., Singh, A. K., & Rajagopalan, A. N. (2017). Blur-invariant deep learning for blind-deblurring. In *IEEE international conference on computer vision* (pp. 4762–4770).

Pan, J., Dong, J., Tai, Y., Su, Z., & Yang, M. (2017a). Learning discriminative data fitting functions for blind image deblurring. In *IEEE international conference on computer vision* (pp. 1077–1085)

Pan, J., Hu, Z., Su, Z., & Yang, M. (2014). Deblurring face images with exemplars. In *European conference on computer vision*.

Pan, J., Hu, Z., Su, Z., & Yang, M. (2017b). L_0-regularized intensity and gradient prior for deblurring text images and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(2), 342–355.

Pan, J., Lin, Z., Su, Z., & Yang, M. H. (2016a). Robust kernel estimation with outliers handling for image deblurring. In *IEEE conference on computer vision and pattern recognition*.

Pan, J., Sun, D., Pfister, H., & Yang, M. (2016b). Blind image deblurring using dark channel prior. In *IEEE conference on computer vision and pattern recognition*.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British machine vision conference*.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.

Ren, W., Cao, X., Pan, J., Guo, X., Zuo, W., & Yang, M. H. (2016a). Image deblurring via enhanced low-rank prior. *IEEE Transactions on Image Processing*, *25*(7), 3426–3437.

Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., & Yang, M. H. (2016b). Single image dehazing via multi-scale convolutional neural networks. In *European conference on computer vision*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *IEEE conference on computer vision and pattern recognition*.

Schuler, C. J., Christopher Burger, H., Harmeling, S., & Scholkopf, B. (2013). A machine learning approach for non-blind image deconvolution. In *IEEE conference on computer vision and pattern recognition*.

Schuler, C. J., Hirsch, M., Harmeling, S., & Schölkopf, B. (2016). Learning to deblur. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(7), 1439–1451.

Shan, Q., Jia, J., & Agarwala, A. (2008). High-quality motion deblurring from a single image. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, *27*(3), 73:1–73:10.

Shen, Z., Lai, W., Xu, T., Kautz, J., & Yang, M. (2018). Deep semantic face deblurring. In *IEEE conference on computer vision and pattern recognition*.

Sim, T., Baker, S., & Bsat, M. (2002). The CMU pose, illumination, and expression (pie) database. In *IEEE international conference on automatic face and gesture recognition*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations*.

Sun, J., Cao, W., Xu, Z., & Ponce, J. (2015). Learning a convolutional neural network for non-uniform motion blur removal. In *IEEE conference on computer vision and pattern recognition*.

Sun, L., Cho, S., Wang, J., & Hays, J. (2013a). Edge-based blur kernel estimation using patch priors. In *IEEE international conference on computational photography*.

Sun, L., Cho, S., Wang, J., & Hays, J. (2014). Good image priors for non-blind deconvolution. In *European conference on computer vision*.

Sun, Y., Wang, X., & Tang, X. (2013b). Deep convolutional network cascade for facial point detection. In *IEEE conference on computer vision and pattern recognition*.

Tao, X., Gao, H., Shen, X., Wang, J., & Jia, J. (2018). Scale-recurrent network for deep image deblurring. In *IEEE conference on computer vision and pattern recognition*.

Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, *28*(6), 902–913.

Vedaldi, A., & Lenc, K. (2015). MatConvNet: Convolutional neural networks for matlab. In *ACM international conference on multimedia*.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018a). Cosface: Large margin cosine loss for deep face recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 5265–5274).

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., Catanzaro, B. (2018b). High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE conference on computer vision and pattern recognition*.

Whyte, O., Sivic, J., Zisserman, A., & Ponce, J. (2012). Non-uniform deblurring for shaken images. *International Journal of Computer Vision*, *98*(2), 168–186.

Xu, L., & Jia, J. (2010). Two-phase kernel estimation for robust motion deblurring. In *European conference on computer vision*.

Xu, L., Ren, J. S. J., Liu, C., & Jia, J. (2014). Deep convolutional neural network for image deconvolution. In *Neural information processing systems*.

Xu, L., Zheng, S., & Jia, J. (2013). Unnatural L0 sparse representation for natural image deblurring. In *IEEE conference on computer vision and pattern recognition*.

Xu, X., Pan, J., Zhang, Y., & Yang, M. (2018). Motion blur kernel estimation via deep learning. *IEEE Transactions on Image Processing*, *27*(1), 194–205.

Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., & Yang, M. H. (2017). Learning to super-resolve blurry face and text images. In *IEEE international conference on computer vision*.

Yan, Y., Ren, W., Guo, Y., Wang, R., & Cao, X. (2017). Image deblurring via extreme channels prior. In *IEEE conference on computer vision and pattern recognition*.

Zhang, J., Pan, J., Lai, W. S., Lau, R. W. H., & Yang, M. H. (2017). Learning fully convolutional networks for iterative non-blind deconvolution. In *IEEE conference on computer vision and pattern recognition*.

Zhong, L., Cho, S., Metaxas, D. N., Paris, S., & Wang, J. (2013). Handling noise in single image deblurring using directional filters. In *IEEE conference on computer vision and pattern recognition*.