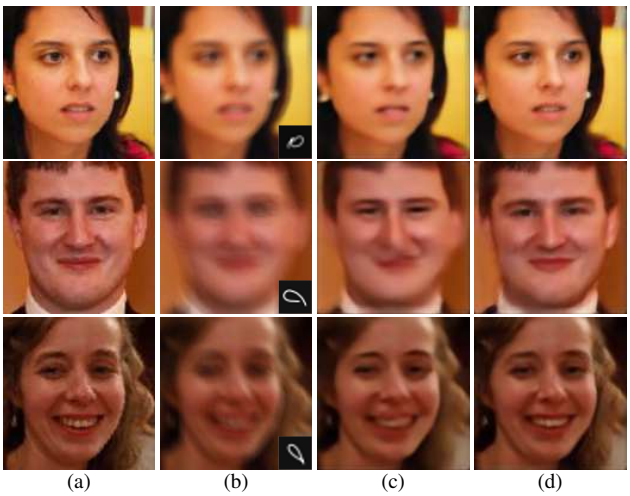
Deep Semantic Face Deblurring

Ziyi Shen[1] Wei-Sheng Lai[2] Tingfa Xu[1*] Jan Kautz[3] Ming-Hsuan Yang[2,4]
[1]Beijing Institute of Technology [2]University of California, Merced
[3]Nvidia [4]Google Cloud

https://sites.google.com/site/ziyishenmi/cvpr18_face_deblur

Abstract

In this paper, we present an effective and efficient face deblurring algorithm by exploiting semantic cues via deep convolutional neural networks (CNNs). As face images are highly structured and share several key semantic components (e.g., eyes and mouths), the semantic information of a face provides a strong prior for restoration. As such, we propose to incorporate global semantic priors as input and impose local structure losses to regularize the output within a multi-scale deep CNN. We train the network with perceptual and adversarial losses to generate photo-realistic results and develop an incremental training strategy to handle random blur kernels in the wild. Quantitative and qualitative evaluations demonstrate that the proposed face deblurring algorithm restores sharp images with more facial details and performs favorably against state-of-the-art methods in terms of restoration quality, face recognition and execution speed.



(a) (b) (c) (d)

Figure 1. **Face deblurring results**. We exploit the semantic information of face within an end-to-end deep CNN for face image deblurring. (a) Ground truth images (b) Blurred images (c) Ours w/o semantics (d) Ours w/ semantics.

1. Introduction

Single image deblurring aims to recover a clear image from a single blurred input image. Conventional methods model the blur process (assuming spatially invariant blur) as the convolution operation between a latent clear image and a blur kernel, and formulate this problem based on the maximum a posteriori (MAP) framework. As the problem is ill-posed, the state-of-the-art algorithms rely on natural image priors (e.g., L_0 gradient [48] and dark channel prior [31]) to constrain the solution space.

While existing image priors are effective for deblurring natural images, the underlying assumption may not hold for images from specific categories, e.g., text, face and low-light conditions. Therefore, numerous approaches exploit domain-specific priors or strategies, such as L_0 intensity [30] for text images and light streaks [13] for extremely low-light images. As face images typically have fewer textures and edges for estimating blur kernels, Pan et al. [29] propose to search a similar face exemplar from an external dataset and extract the contour as reference edges. However, a similar reference image may not always exist to cover the diversity of face images in the wild. Furthermore, those methods based on the MAP framework typically entail heavy computational cost due to the iterative optimization of latent images and blur kernels. The long execution time limits the applications on resource-sensitive platforms, e.g., cloud and mobile devices.

In this work, we focus on deblurring face images and propose an efficient as well as effective solution using deep CNNs. Since face images are highly structured and composed of similar components, the semantic information serves as a strong prior for restoration. Therefore, we propose to leverage the face semantic labels as global priors and local constraints for deblurring face images. Specifically, we first generate the semantic labels of blurred input images using a face parsing network. The face deblurring network then takes the blurred image and semantic labels as input to restore a clear image in a coarse-to-fine man-

*Corresponding author

ner. To encourage the network for generating fine details, we further impose a local structure loss on important face components (e.g., eyes, noses, and mouths). Figure 1 shows deblurred examples with and without the proposed semantic priors and losses. The proposed method is able to reconstruct better facial details than the network trained with only the pixel-wise L_1 loss function (i.e., without using semantics). As our method is end-to-end without any blur kernel estimation or post-processing, the execution time is much shorter than the state-of-the-art MAP-based approaches.

To handle blurred images produced by unknown blur kernels, existing methods typically synthesize blur kernels by modeling the camera trajectories [4, 12] and generate a large number of blurred images for training. Instead of simultaneously using all synthetic blurred images for training, we propose an incremental training strategy by first training the network on a set of small blur kernels and then incorporating larger blur kernels sequentially. We show that the proposed incremental training strategy facilitates the convergence and improves the performance of our deblurring network on various sizes of blur kernels. Finally, we impose a perceptual loss [14] and an adversarial loss [10] to generate photo-realistic deblurred results.

We make the following contributions in this work:

- We propose a deep multi-scale CNN that exploits global semantic priors and local structural constraints for face image deblurring.
- We present an incremental strategy to train CNNs to better handle unknown motion blur kernels.
- We demonstrate that the proposed method performs favorably against state-of-the-art deblurring approaches in terms of restoration quality, face recognition and execution speed.

2. Related Work

Single image deblurring can be categorized into non-blind and blind deblurring based on whether the blur kernel is available or not. We focus our discussion on blind image deblurring in this section.

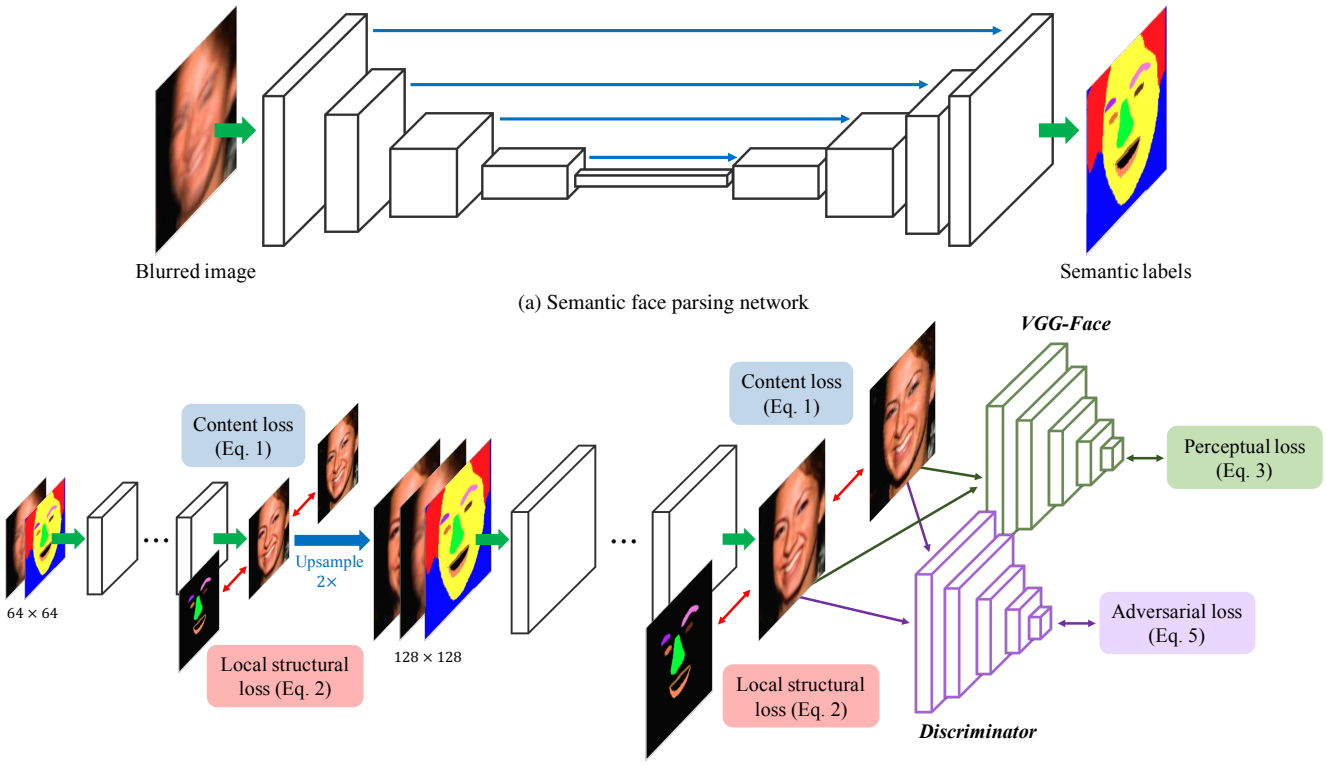
Generic methods. The recent progress in single image blind deblurring can be attributed to the development of effective natural image priors, including sparse image gradient prior [8, 23], normalized sparsity measure [17], patch prior [42], L_0 gradient [48], color-line model [18], low-rank prior [34], self-similarity [27] and dark channel prior [31]. Through optimizing the image priors within the MAP framework, those approaches implicitly restore strong edges for estimating the blur kernels and latent sharp images. However, solving complex non-linear priors involve several optimization steps and entail high computational loads. As such, edge-selection based methods [6, 46] adopt simple image priors (e.g., L_2 gradients) with image filters (e.g., shock filter) to explicitly restore or select strong edges. While generic image deblurring methods demonstrate state-of-the-art performance, face images have different statistical properties than natural scenes and cannot be restored well using the above approaches.

Domain-specific methods. To handle images from specific categories, several domain-specific image deblurring approaches have been developed. Pan et al. [30] introduce the L_0-regularized priors on both intensity and image gradients for text image deblurring as text images usually contain nearly uniform intensity. To handle extreme cases such as low-light images, Hu et al. [13] detect the light streaks in images for estimating blur kernels. Anwar et al. [2] propose a frequency-domain class-specific prior to restore the band-pass frequency components. In addition, a number of approaches use reference images as guidance for non-blind [43] and blind deblurring [11]. However, the performance of such methods hinges on the similarity of the reference images and quality of dense correspondence.

As face images have fewer textures and edges, existing algorithms based on implicit or explicit edge restoration are less effective. Pan et al. [29] search for similar faces from a face dataset and extract reference exemplar contours for estimating blur kernels. However, this approach requires manual annotations of the face contours and involves computationally expensive optimization processes of blur kernels and latent images in the MAP framework. In contrast, we train an end-to-end deep CNN to bypass the blur kernel estimation step and do not use any reference images or manual annotations when deblurring an image.

CNN-based methods. Deep CNNs have been adopted for several image restoration tasks, such as denoising [26], JPEG deblocking [7], dehazing [35] and super-resolution [16, 19]. Recent approaches apply deep CNNs for image deblurring in several aspects, including non-blind deconvolution [37, 47, 50], blur kernel estimation [38] and dynamic scene deblurring [28]. Chakrabarti et al. [4] train a deep network to predict the Fourier coefficients of a deconvolution filter. Despite computational efficiency, these CNN-based methods do not perform as well as the state-of-the-art MAP-based approaches, especially on large motion kernels.

Since text images usually contain uniform intensities with fewer texture regions, an end-to-end deep network [12] performs well, especially under large noise levels. Xu et al. [49] aim to jointly deblur and super-resolve low-resolution blurred face and text images, which are typically degraded by Gaussian-like blur kernels. In this work, we focus on deblurring face images from complex motion blur. We exploit global and local semantic cues as well as perceptual [14] and adversarial [10] losses to restore photo-realistic face images with fine details.



(a) Semantic face parsing network

(b) Face deblurring network

Figure 2. **Overview of the proposed semantic face deblurring network.** The proposed network consists of two sub-networks: a semantic face parsing network and a multi-scale deblurring network. The face parsing network generates the semantic labels of the input blurred images. The multi-scale deblurring network has two scales. We concatenate the blurred image and semantic labels as the input to the first scale. At the second scale, the input is the concatenation of the upsampled deblurred image from the first scale, the blurred image and the corresponding semantic labels. Each scale of the deblurring network receives the supervision from the pixel-wise content loss and local structural losses. We impose the perceptual and adversarial losses at the output of the second scale.

3. Semantic Face Deblurring

In this section, we describe the design methodology of the proposed semantic face deblurring approach. We exploit the semantic labels from a face parsing network as the global semantic priors and local structural losses within a deep multi-scale CNN. We then train the proposed network jointly with perceptual and adversarial losses to generate photo-realistic deblurred results.

3.1. Face deblurring network

We use a multi-scale network similar to that from Nah et al. [28], but with several differences. First, as face images typically have a spatial resolution of 128×128 or less, we use only 2 scales instead of 3 scales for natural images in [28]. Second, we use fewer ResBlocks (reduce from 19 to 6) and larger filter size (11×11) at the first convolutional layer to increase the receptive field. Finally, we introduce additional inputs from semantic face parsing as global priors and impose local structural constraints on the output at each scale.

3.2. Global semantic priors

We propose to utilize the semantic parsing information as a global prior for face deblurring. Given a blurred image, we first use a face parsing network [24] to extract the semantic labels. We then concatenate the probability maps of the semantic labels with the blurred face image as the input to our deblurring network. The input to the first scale of the deblurring network has a spatial resolution of 64×64 and a total of 14 channels (3-channel RGB image and 11-channel semantic probabilities). The deblurred image of the first scale is then upsampled by $2\times$ through a transposed convolutional layer. The input of the second scale has a spatial resolution of 128×128 and a total of 17 channels, including the upsampled deblurred image, the blurred image, and the corresponding semantic probabilities. Figure 2 shows an overview of our face parsing and deblurring network. The semantic labels encode the essential appearance information and rough locations of the facial components (e.g., eyes, noses and mouths) and serve as a strong global prior for reconstructing the deblurred face image.

3.3. Local structural constraints

We use the pixel-wise L_1 robust function as the content loss of our face deblurring network:

$$\mathcal{L}_c = \|\mathcal{G}(B, \mathcal{P}(B)) - I\|_1 , \tag{1}$$

where $\mathcal{P}$ and $\mathcal{G}$ denote the face parsing and deblurring networks. In addition, B and I are the blurred and ground truth clear images, respectively. However, the key components (e.g., eyes, lips and mouths) on faces are typically small and cannot be well reconstructed by solely minimizing the content loss on the whole face image. As human vision is more sensitive to the artifacts on key components, we propose to impose local structural losses:

$$\mathcal{L}_s = \sum_{k=1}^{K} \|\mathbb{M}_k(\mathcal{P}(B)) \odot (\mathcal{G}(B, \mathcal{P}(B)) - I)\|_1 , \tag{2}$$

where $\mathbb{M}_k$ denotes the structural mask of the k-th component and $\odot$ is the element-wise multiplication. We apply the local structural losses on eyebrows, eyes, noses, lips and teeth. The local structural losses enforce the deblurring network to restore more details on those key components.

3.4. Generating photo-realistic face images

As pixel-wise L_2 or L_1 loss functions typically lead to overly-smooth results, we introduce a perceptual loss [14] and an adversarial loss [10] to optimize our deblurring network and generate photo-realistic deblurred results.

Perceptual loss. The perceptual loss has been adopted in style transfer [9, 14], image super-resolution [22] and image synthesis [5]. The perceptual loss aims to measure the similarity in the high dimensional feature space of a pre-trained loss network (e.g., VGG16 [41]). 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:

$$\mathcal{L}_p = \sum_{l} \|\phi_l(\mathcal{G}(B)) - \phi_l(I)\|_2^2 . \tag{3}$$

We compute the perceptual loss on the Pool2 and Pool5 layers of the pre-trained VGG-Face [32] network.

Adversarial loss. The adversarial training framework has been shown effective to synthesize realistic images [10, 22, 28]. We treat our face deblurring network as the generator and construct a discriminator based on the architecture of DCGAN [33]. 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:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}\left[\log \mathcal{D}(I)\right] + \mathbb{E}\left[\log(1 - \mathcal{D}(\mathcal{G}(B)))\right] . \tag{4}$$

When updating the generator, the adversarial loss is:

$$\mathcal{L}_{\text{adv}} = -\log \mathcal{D}(\mathcal{G}(B)). \tag{5}$$

Our discriminator takes an input image with a size of 128×128 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 to be a real image.

Overall loss function. The overall loss function for training our face deblurring network consists of the content loss, local structural losses, perceptual loss and adversarial loss:

$$\mathcal{L} = \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}, \tag{6}$$

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

3.5. Implementation details

We use a variant of Liu et al. [24] as our semantic face parsing network, which is an encoder-decoder architecture with skip connections from the encoder to the decoder (see Figure 2(a)). Our face deblurring network has two scales, where each scale has 6 ResBlocks and a total of 18 convolutional layers. All convolutional layers except the first layer have the kernel size of 5×5 and 64 channels. The upsampling layer uses a 4×4 transposed convolutional layer to upsample the image by $2\times$. The detailed architecture of our face deblurring network is described in the supplementary material.

We implement our network using the MatConvNet toolbox [45]. We use a batch size of 16 and set the learning rate to $5e^{-6}$ when training the parsing network and $4e^{-5}$ when training the deblurring network. The parsing network converges within 60,000 iterations and the training takes less than one day. We train the deblurring network for 17 million iterations, which takes about 5 days on an NVIDIA Titan X GPU. We note that we first train the semantic face parsing network until convergence. We then fix the parsing network while training the deblurring network.

4. Experimental Results

In this section, we first describe the training and test datasets used in our experiments. We then analyze the performance of the semantic face parsing network and face deblurring network, describe our incremental training strategy to handle random blur kernels, and finally compare with state-of-the-art deblurring algorithms.

4.1. Datasets

We use the Helen dataset [21], which has ground truth face semantic labels, for training our semantic face parsing network. The Helen dataset consists of 2,000 training images and 330 validation images. We use the method of Sun et al. [44] to detect the facial key points and align all face images using the method of Kae et al. [15]. During training, we apply data augmentation using affine transformations to avoid over-fitting.

To train the deblurring network, we collect training images from the Helen dataset [21] (2,000 images), CMU PIE dataset [40] (2,164 images) and CelebA dataset [25] (2,300 images) as our training data. We synthesize 20,000 motion blur kernels from random 3D camera trajectories [3]. The size of blur kernels range 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.

In addition to the training set, we synthesize another 80 random blur kernels, which are different from the 20,000 blur kernels used for training. We collect 100 clear face images from the validation set of the Helen and CelebA datasets, respectively. There are a total of 16,000 blurred images for testing.

4.2. Semantic face parsing

We first validate the performance of our semantic face parsing network. We use the images from the Helen dataset for training and evaluate the F-scores of each facial component on the Helen validation set. We report the performance on clear and blurred images in Table 1. Due to motion blur, the face parsing network does not perform well on blurred images, especially for small and thin components, e.g., eyebrows, lips, and teeth. We further fine-tune the parsing network on blurred images to improve the performance. Figure 3 shows the parsing results before and after fine-tuning on blurred images. The fine-tuned model is more robust to motion blur and parses facial components well.

4.3. Face image deblurring

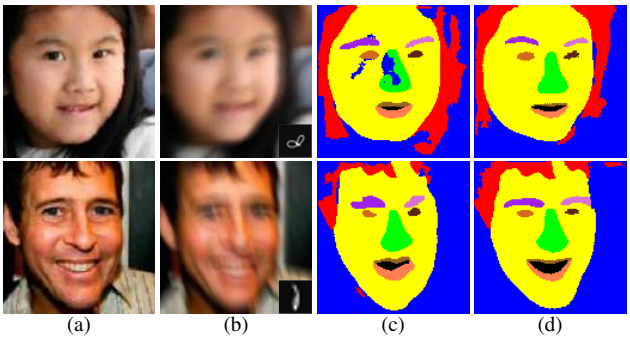
In this section, we evaluate the effect of using semantic information on face image deblurring, describe our incremental training strategy for handling random blur kernels, and compare with state-of-the-art deblurring methods.

Effect of semantic parsing. We train a baseline model using only the content loss function (1). We then train another two models by first introducing the semantic labels as input priors and then including the local structural losses (2).

Figure 4 shows two deblurred results from our Helen test set. The network optimized solely from the content loss produces overly smooth deblurred results. The shape of the faces and lips cannot be well recovered as in Figure 4(c). By introducing the semantic labels as the global priors, the network better reconstructs the outline of faces. However, the results may not contain fine details in several key components, such as teeth and eyes. The network with the additional local structural losses restores more details and textures as shown in Figure 4(e). Table 2 shows the performance contribution of each component on both the Helen and CelebA test sets.

Table 1. **Performance of our semantic face parsing network.** We measure the F-score on each facial component. "Pre-trained" model denotes the network trained on clear images. "Fine-tuned" model is the network fine-tuned on blurred images.

Input image	Clear	Blurred	
Evaluated model	Pre-trained	Pre-trained	Fine-tuned
face	0.923	0.891	0.896
left eyebrow	0.730	0.574	0.596
right eyebrow	0.731	0.581	0.618
left eye	0.748	0.602	0.677
right eye	0.786	0.630	0.608
nose	0.893	0.875	0.855
upper lip	0.645	0.489	0.477
lower lip	0.744	0.605	0.650
teeth	0.451	0.303	0.369
hair	0.557	0.481	0.499
average	0.721	0.603	0.625

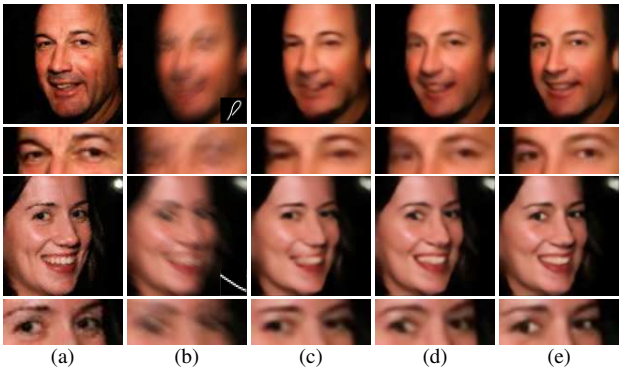


(a) (b) (c) (d)

Figure 3. **Labeling results of our semantic face parsing network.** (a) Ground truth images (b) Blurred images (c) Results from pre-trained model (trained on clear images) (d) Results from fine-tuned model (fine-tuned on blurred images).

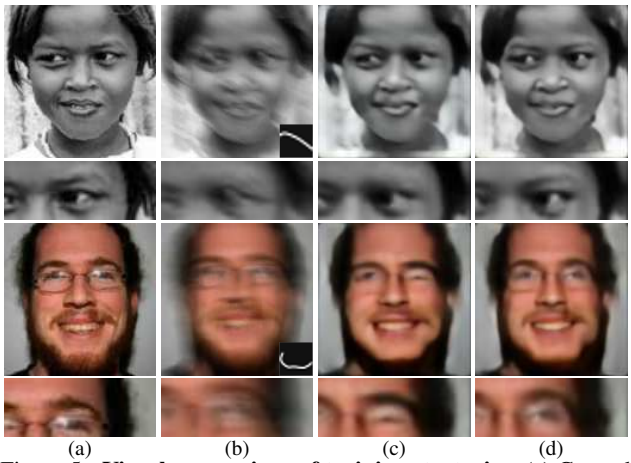
Incremental training. Real-world blurred images are likely formed by a large diversity of camera motion. In order to handle random blur kernels in the wild, a simple strategy is to synthesize a large number of blur kernels and blurred images for training. However, it is difficult to train a deep network from scratch using all blurred images simultaneously as the network has to learn N-to-1 mapping where N is the number of blur kernels. The network may converge to a bad local minimum and cannot restore images well especially for large blur kernels.

To address this issue, we propose a simple yet effective incremental training strategy by incorporating more blur kernels sequentially during training. We first train the net-



(a) (b) (c) (d) (e)

Figure 4. **Effects of semantic face parsing on image deblurring.** (a) Ground truth images (b) Blurred images (c) Content loss (d) Content loss + global semantic priors (e) Content loss + global semantic priors + local structural losses.



(a) (b) (c) (d)

Figure 5. **Visual comparison of training strategies.** (a) Ground truth images (b) Blurred images (c) Direct training (d) Incremental training.

work on smaller blur kernels (i.e., 13×13). We then gradually expand the training set by increasing the size of blur kernels. Specifically, we train the network for K iterations before introducing new blur kernels. While incorporating new blur kernels, we still sample the existing blur kernels for training until all blur kernels are included. We set $K = 30000$ iterations in our experiments and train the network for a total of 17 million iterations.

We provide a comparison of the direct training (i.e., training all blurred kernels simultaneously) and the proposed incremental training in Figure 5 and Table 2. Figure 6 shows the quantitative comparison on different sizes of blur kernels. The proposed incremental training strategy performs better on all sizes of blur kernels and restores the images well.

Comparisons with state-of-the-arts. We provide qualitative and quantitative comparisons with 7 state-of-the-art deblurring algorithms, including MAP-based methods [6, 17, 39, 48, 51], a face deblurring method [29] and a CNN-based method [28]. We denote our method with all the losses and semantic priors as "ours w/ semantics" and our method using only the content loss as "ours w/o semantics".

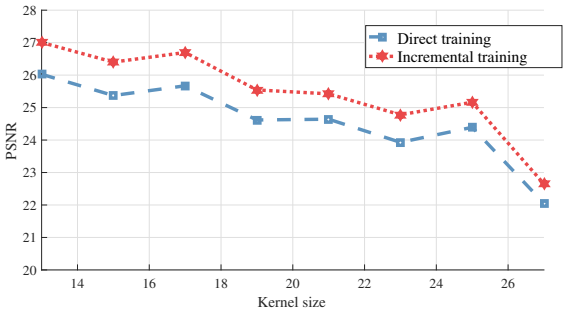


Figure 6. **Quantitative evaluation of training strategies.** We compare the direct training (blue curve) and the proposed incremental training (red curve) strategies on the Helen test set.

Table 2. **Ablation study**. While the model with the perceptual loss achieves the highest PSNR/SSIM, including the adversarial loss produces more realistic face images.

Approach	Helen		CelebA	
	PSNR	SSIM	PSNR	SSIM
Content loss	24.85	0.849	24.23	0.864
+ Global semantic priors	25.32	0.857	24.32	0.864
+ Local structural loss	25.48	0.859	24.58	0.866
+ Incremental training	25.55	0.860	24.61	0.869
+ Perceptual loss	**25.99**	**0.871**	**25.05**	**0.879**
+ Adversarial loss	25.58	0.861	24.34	0.860

Table 3. **Quantitative comparison with state-of-the-art methods.** We compute the average PSNR and SSIM on two test sets.

Method	Helen		CelebA	
	PSNR	SSIM	PSNR	SSIM
Krishnan et al. [17]	19.30	0.670	18.38	0.672
Pan et al. [29]	20.93	0.727	18.59	0.677
Shan et al. [39]	19.57	0.670	18.43	0.644
Xu et al. [48]	20.11	0.711	18.93	0.685
Cho and Lee [6]	16.82	0.574	13.03	0.445
Zhong et al. [51]	16.41	0.614	17.26	0.695
Nah et al. [28]	24.12	0.823	22.43	0.832
Ours	**25.58**	**0.861**	**24.34**	**0.860**

We evaluate the PSNR and SSIM on both the Helen and CelebA datasets in Table 3. Figure 7 shows quantitative comparisons on different sizes of blur kernels. The proposed method performs favorably against the state-of-the-art approaches on both datasets and all blur kernel sizes. We present visual comparisons in Figure 8. Conventional MAP-based methods [6, 17, 39, 48, 51] are less effective on deblurring face images and lead to more ringing artifacts. The MAP-based face deblurring approach [29] is not robust to noise and highly relies on the similarity of the reference image. The CNN-based method [28] does not consider the face semantic information and thus produces overly smooth results. In contrast, the proposed method utilizes the global and local face semantics to restore face images with more fine details and less visual artifacts. We provide more visual comparisons in the supplementary material.

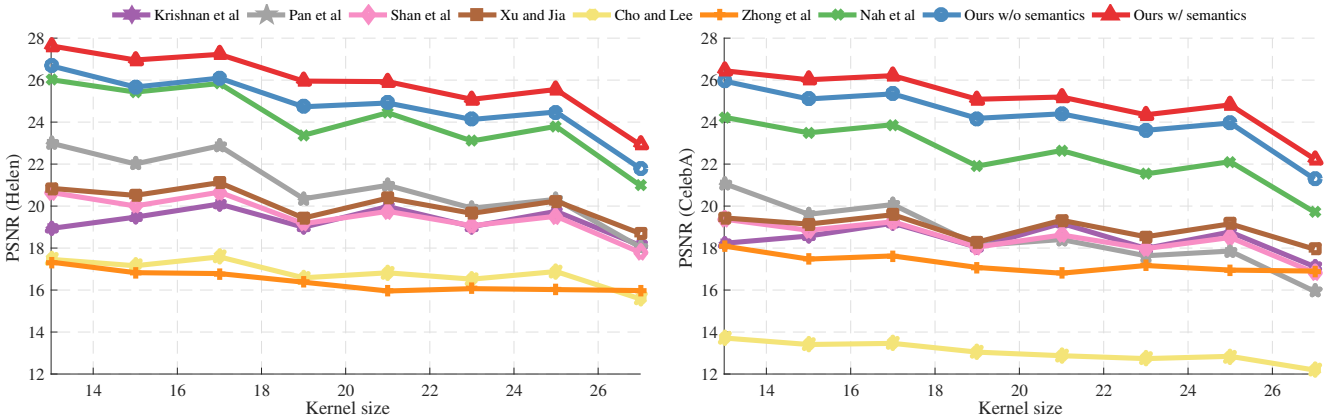


Figure 7. **Quantitative evaluation on different sizes of blur kernels.** There are 100 latent clear images, 80 blur kernels and a total of 8,000 blurred images in the Helen and CelebA test sets, respectively. Our method performs well on all sizes of blur kernels.

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

Table 4. **Comparison of execution time.** We report the average execution time on 10 images with the size of 128 × 128.

Method	Implementation	CPU / GPU	Seconds
Krishnan et al. [17]	MATLAB	CPU	2.52
Pan et al. [29]	MATLAB	CPU	8.11
Shan et al. [39]	C++	CPU	16.32
Xu et al. [48]	C++	CPU	0.31
Cho and Lee [6]	C++	CPU	0.41
Zhong et al. [51]	MATLAB	CPU	8.07
Nah et al. [28]	MATLAB	GPU	0.09
Ours	MATLAB	GPU	**0.05**

Execution time. We analyze the execution time on a machine with a 3.4 GHz Intel i7 CPU (64G RAM) and an NVIDIA Titan X GPU (12G memory). Table 4 shows the average execution time based on 10 images with a size of 128 × 128. The proposed method is more efficient than the state-of-the-art deblurring algorithms.

Face recognition. We first use the FaceNet [36] to compute the identity distance (i.e., the L_2 distance on the outputs of FaceNet) between the ground truth face image and deblurred results. Figure 9 shows that the deblurred images from the proposed method have the lowest identity distance, which demonstrates that the proposed method preserves the face identity well.

As the CelebA dataset contains identity labels, we conduct another experiment on face detection and 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. Given an image from 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.

We use the OpenFace toolbox [1] to detect the face for each image in the probe set. However, due to the motion

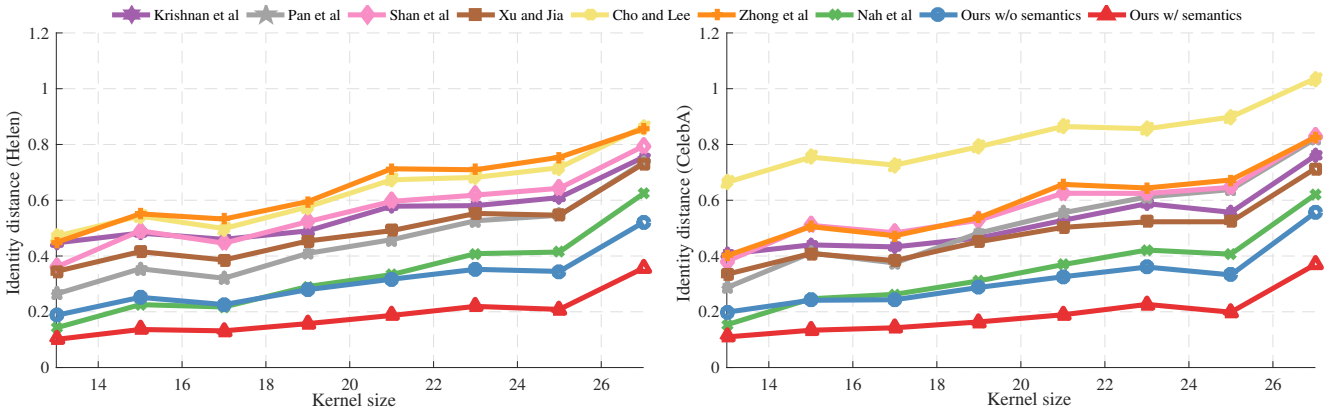


Figure 9. **Quantitative evaluation on face identity distance.** We use the FaceNet [36] to compute the identity distance between the clear and deblurred face images. The proposed method achieves the lowest identity distance on both the Helen and CelebA test sets.

Table 5. **Face detection and recognition on the CelebA dataset.** We show the success rate of face detection and top-1, top-3 and top-5 accuracy of face recognition.

Method	Detection	Top-1	Top-3	Top-5
Clear images	100%	71%	84%	89%
Blurred images	81%	31%	46%	53%
Krishnan et al. [17]	84%	36%	51%	59%
Pan et al. [29]	82%	44%	57%	64%
Shan et al. [39]	80%	34%	49%	56%
Xu et al. [48]	86%	43%	57%	64%
Cho and Lee [6]	56%	21%	31%	37%
Zhong et al. [51]	73%	30%	44%	51%
Nah et al. [28]	90%	42%	57%	64%
Ours w/o semantics	95%	42%	55%	62%
Ours w/ semantics	**99**%	**54**%	**68**%	**74**%

blur and the ringing artifacts, faces in some of the blurred and deblurred images cannot be well detected. We then compute the identity distance with all images in the gallery set and select the top-K nearest matches. We show the success rate of the face detection for blurred images and state-of-the-art deblurring approaches in Table 5. Furthermore, we compute the recognition accuracy on those successfully detected face images and show the top-1, top-3 and top-5 accuracy. The proposed method produces fewer artifacts and thus achieves the highest success rate as well as recognition accuracy against other evaluated approaches.

Real-world blurred images. We also test the proposed method on face images collected from the real blurred dataset of Lai et al. [20]. As shown in Figure 10, our method restores more visually pleasing faces than state-of-the-art approaches [29, 48]. We provide more deblurred results of real-world blurred images in the supplementary material.

4.4. Limitations and discussions

Our method may fail when the input face image cannot be well aligned, e.g., side faces or extremely large motion blur. Future work includes improving the performance on handling large and non-uniform blur kernels and relieving the requirement of face alignment.

(a) (b) (c) (d) (e)

Figure 10. **Visual comparison of real-world blurred images.** (a) Blurred images (b) Xu et al. [48] (c) Pan et al. [29] (d) Nah et al. [28] (e) Ours

5. Conclusions

In this work, we propose a 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. In addition, we optimize the network with perceptual and adversarial losses to produce photo-realistic results. We further propose an incremental training strategy for handling random and unknown blur kernels in the wild. Experimental results on image deblurring, execution time and face recognition demonstrate that the proposed method performs favorably against state-of-the-art deblurring algorithms.

Acknowledgments

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

[1] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, 2016. 7

[2] S. Anwar, C. Phuoc Huynh, and F. Porikli. Class-specific image deblurring. In *ICCV*, 2015. 2

[3] G. Boracchi and A. Foi. Modeling the performance of image restoration from motion blur. *TIP*, 21(8):3502–3517, 2012. 5

[4] A. Chakrabarti. A neural approach to blind motion deblurring. In *ECCV*, 2016. 2

[5] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 4

[6] S. Cho and S. Lee. Fast motion deblurring. *ACM TOG (Proceedings of SIGGRAPH Asia)*, 28(5):145:1–145:8, 2009. 2, 6, 7, 8

[7] C. Dong, Y. Deng, C. Change Loy, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, 2015. 2

[8] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. *ACM TOG (Proceedings of SIGGRAPH)*, pages 787–794, 2006. 2

[9] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, 2015. 4

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 4

[11] Y. Hacohen, E. Shechtman, and D. Lischinski. Deblurring by example using dense correspondence. In *ICCV*, 2013. 2

[12] M. Hradiš, J. Kotera, P. Zemcík, and F. Sroubek. Convolutional neural networks for direct text deblurring. In *BMVC*, 2015. 2

[13] Z. Hu, S. Cho, J. Wang, and M.-H. Yang. Deblurring low-light images with light streaks. In *CVPR*, 2014. 1, 2

[14] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2, 4

[15] A. Kae, K. Sohn, H. Lee, and E. G. Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. In *CVPR*, 2013. 5

[16] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[17] D. Krishnan, T. Tay, and R. Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, 2011. 2, 6, 7, 8

[18] W.-S. Lai, J.-J. Ding, Y.-Y. Lin, and Y.-Y. Chuang. Blur kernel estimation using normalized color-line prior. In *CVPR*, 2015. 2

[19] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 2

[20] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016. 8

[21] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012. 5

[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 4

[23] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009. 2

[24] S. Liu, J. Yang, C. Huang, and M. Yang. Multi-objective convolutional learning for face labeling. In *CVPR*, 2015. 3, 4

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[26] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NIPS*, 2016. 2

[27] T. Michaeli and M. Irani. Blind deblurring using internal patch recurrence. In *ECCV*, 2014. 2

[28] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[29] J. Pan, Z. Hu, Z. Su, and M. Yang. Deblurring face images with exemplars. In *ECCV*, 2014. 1, 2, 6, 7, 8

[30] J. Pan, Z. Hu, Z. Su, and M. Yang. L_0-regularized intensity and gradient prior for deblurring text images and beyond. *TPAMI*, 39(2):342–355, 2017. 1, 2

[31] J. Pan, D. Sun, H. Pfister, and M. Yang. Blind image deblurring using dark channel prior. In *CVPR*, 2016. 1, 2

[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 4

[33] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 4

[34] W. Ren, X. Cao, J. Pan, X. Guo, W. Zuo, and M.-H. Yang. Image deblurring via enhanced low-rank prior. *TIP*, 25(7):3426–3437, 2016. 2

[35] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, 2016. 2

[36] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 7, 8

[37] C. J. Schuler, H. Christopher Burger, S. Harmeling, and B. Scholkopf. A machine learning approach for non-blind image deconvolution. In *CVPR*, 2013. 2

[38] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf. Learning to deblur. *TPAMI*, 38(7):1439–1451, 2016. 2

[39] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *ACM TOG (Proceedings of SIGGRAPH)*, 27(3):73:1–73:10, 2008. 6, 7, 8

[40] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002. 5

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[42] L. Sun, S. Cho, J. Wang, and J. Hays. Edge-based blur kernel estimation using patch priors. In *ICCP*, 2013. 2

[43] L. Sun, S. Cho, J. Wang, and J. Hays. Good image priors for non-blind deconvolution. In *ECCV*, 2014. 2

[44] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 5

[45] A. Vedaldi and K. Lenc. MatConvNet: Convolutional neural networks for matlab. In *ACM MM*, 2015. 4

[46] L. Xu and J. Jia. Two-phase kernel estimation for robust motion deblurring. In *ECCV*, 2010. 2

[47] L. Xu, J. S. J. Ren, C. Liu, and J. Jia. Deep convolutional neural network for image deconvolution. In *NIPS*, 2014. 2

[48] L. Xu, S. Zheng, and J. Jia. Unnatural L0 sparse representation for natural image deblurring. In *CVPR*, 2013. 1, 2, 6, 7, 8

[49] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang. Learning to super-resolve blurry face and text images. In *ICCV*, 2017. 2

[50] J. Zhang, J. Pan, W.-S. Lai, R. W. H. Lau, and M.-H. Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *CVPR*, 2017. 2

[51] L. Zhong, S. Cho, D. N. Metaxas, S. Paris, and J. Wang. Handling noise in single image deblurring using directional filters. In *CVPR*, 2013. 6, 7, 8