Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09460-x

Blind face images deblurring with enhancement

Qing Qi[1,2] · Jichang Guo[1] · Chongyi Li[3] · Lijun Xiao[1]

Received: 2 April 2019 / Revised: 7 March 2020 / Accepted: 28 July 2020 /
Published online: 18 September 2020

Abstract
Face images deblurring has achieved advanced development; however, existing methods involve high computational cost problems. Furthermore, the recovered face images by current methods have the problems of over-smooth textures, ringing artifacts, and poor details. We consider the problem of face images deblurring as a semantic generation task. In this paper, we propose a generative adversarial network (GAN), which includes a perception-inspired blurry removal generator and a discriminator. The proposed generator reconstructs the latent deblurred image by a U-net based network that contains an enhancement module. Face images are highly structured, and thus can be served as a class-specific prior. Considering this, we propose a perceptual loss function to regularize the recovery of face images, which introduces more clear details and reduces the effects of artifacts. The proposed method has a robust capability of generating realistic face images with pleasant visual effects. Extensive experiments on both synthetic and real-world face images demonstrate that the proposed method is comparable with state-of-the-art methods.

Keywords Deep learning · Face images deblurring · Enhancement module

✉ Jichang Guo
jcguo@tju.edu.cn

Qing Qi
qiqing@tju.edu.cn

Chongyi Li
lichongyi25@gmail.com

Lijun Xiao
xljtju@163.com

[1] School of Electrical and Information Engineering, Tianjin University, Weijin Road 92, 300072, Tianjin, China

[2] School of Physics and Electronic Information Engineering, Qinghai Nationalities University, Bayi Middle Road 3, 810007, Xining, China

[3] Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, 999077, Hong Kong, China

1 Introduction

Blind images deblurring task is an ill-posed problem where both sharp latent image and kernels need to be recovered from the single degraded observation. Therefore, additional feature priors are required to constrain image recovery. Single image deblurring task has benefited from hand-crafted priors which are usually developed by natural images and have made advanced progress. Our main focus is on the task of the face images deblurring, the proposed method is potentially applicable in the other types of image. Tackling the face image deblurring problem by directly exploiting classic deconvolution methods which designed generic images, including [51] and the methods of heuristic edge selections are less effective. In other words, these methods mentioned above are modeled on the statistics of natural images and fail to characterize the properties of face images. Furthermore, the strategy of extracting obvious edges to solve deblurring problems is not always effective. In the case that images have large motion blur or highly structured, it is difficult for edges methods to restore salient edges for kernel estimation.

Therefore, several class-specific images deblurring algorithms are proposed. For the face images deblurring task, [2, 11, 34, 35] have emerged in recent years. Anwar et al. [2] propose a class-specific prior for face categories and it is less effective for scenes with complex background. Pan et al. [35] aim to extract similar contours from the external dataset as a reference for kernel estimation. Due to the diversity of facial expressions and posture changes, their proposed external exemplars fail to cover the complexity of face images. Hacohen et al. [11] and Nishiyama et al. [34] have limited applications, due to the strict requirements for references or the known blur kernels. Recently, deep learning is widely used in the domain of computer vision. Chrysos et al. [6] develop a modified version of residual network [12] to deblur face images in a weakly supervised fashion. This algorithm requires the preprocessing of face detection. Although these methods can solve the image deblurring task according to their applicability, the recovered face images have the problems of overly smooth textures, ringing artifacts and poor details. Moreover, these methods mentioned above involve high computational cost problems which lead to the limited practical application.

In this work, we propose an efficient and effectiveness method to recover details of face images. Inspired by GANs [9], we address the face images deblurring problem by learning a class-specific semantic prior. In the context of GANs, the blurry and clear face images are treated as two distinct types of images. This paper takes advantage of adversarial learning to generate similar image statistics to those of sharp face images. In addition, in order to generate realistic face images with pleasant visual effects, we propose a perceptual loss function to regularize the images recovery. Fig. 1 shows the deblurred face images results with and without the proposed loss function. The results of our method have significant visually pleasant effect, especially in the parts of eyes, mouth, and noses. Our model is different from those image deblurring methods which usually require specifically designed priors or rely on salient edges for kernel estimation.

The contributions of this paper are listed as follows: First, we introduce the perceptual loss function to promote the face images to be more clear details. Secondly, we propose a U-net based generator which includes an enhancement module. The enhancement module is aiming at reinforcing feature representations and constructing complicated high-order maps. Finally, we show that the model has a robust capability of generating realistic and artifact-less latent images with pleasant visual effects.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 presents the problem formulation and depicts the proposed method. Section 4 conducts experiments to demonstrate the proposed method. Section 5 analyses and

Fig. 1 The visualization of the deblurred face images with and without the proposed perceptual loss functions. The introduction of perceptual loss functions leads to clearer and more natural-realistic results with fewer artifacts. **a** Ground truth **b** Blurry inputs **c** Our results without perceptual losses **d** Our results with perceptual losses

discusses loss functions and network architecture utilized in the proposed method. Section 6 concludes the proposed method.

2 Related work

In recent years, single image deblurring issue has made significant progress. In this section, we focus on generic image deblurring methods and face class-specific image deblurring methods. Specifically, according to the principle of generic image deblurring algorithms, we classify these methods into three categories: statistics priors, hand-crafted priors, and explicit edge extraction.

2.1 Image deblurring methods

Statistical Priors Since the image deblurring is a highly ill-posed problem, most of image deblurring methods utilize various natural image priors to regularize image reconstruction.

Generally, exploiting statistical properties of generic natural image as constraint terms is a common choice. Fergus et al. [8] construct the mixture of Gaussians to estimate image gradient prior for blind image deblurring. They make single image deblurring possible through constructing a mathematical optimization model based on variational Bayesian inference. Levin et al. [23] point out that instead of simple MAP (Maximum a posteriori)-based methods, methods based on variational Bayesian inference [8] can avoid trivial solutions. However, the computational burden of variational Bayesian inference is enormous.

Hand-Crafted Priors Image deblurring methods with different likelihood functions and image priors based on MAP-framework are proposed and make excellent progress. Hence, the image deblurring problem is also formulated by various priors. The priors include low-rank prior [37], normalized sparsity [19], L_0 gradients [51] and sparse prior [53]. Zhang et al. [53] consider the sparsity of clear images. Image patches can be represented by an over-complete dictionary, which can be used as a regular term for deblurring images. These methods which are based on hand-crafted priors make considerable progress of image deblurring. However, these hand-crafted priors are inevitably involved in the process of kernel estimation. Once the estimated kernel is incorrect, the deconvolution operation results in deviation.

Explicit Edge Extraction In addition to the statistical priors and hand-crafted priors for restricting the solution space of images, the algorithms which explicitly employed salient edges [5, 17, 43, 49] from the blurry observations to estimate kernels are also proposed. In order to estimate kernel, Cho et al. [5] utilize the bilateral filter and the shock filter to suppress the noise and recover the edges respectively. Sun et al. [43] estimate kernel by predicting the edge of each image patch. Once the restored edges are not structural ones, it would result in erroneous results due to the edges extracted from the patches which represent local information rather than global structure. Xu et al. [49] propose a strategy to select edges and note that not all of them extracted from the observed image are beneficial to blur estimation. In the case that images have the characteristics of large motion blur or highly structured, it is difficult for adopting these methods to restore sharp and salient edges directly from the inputs for kernels estimation. Therefore, methods of edge extraction do not perform decent on class-specific images deblurring.

2.2 Face images deblurring

Face Class-Specific Methods Anwar et al. [2] focus on specific images deblurring problem, they design a class-specific image prior by means of aggregation of Fourier magnitude spectrum across all frequency bands. It has been shown to be effective for certain object categories and less effective for scenes with complex background. Recently, exemplar-based methods [11, 35] are proposed. Pan et al. [35] develop a face images deblurring method by utilizing external examples datasets. They manually extract the overall structure of the external exemplar datasets, and then they match the structure of the input observation with those of exemplar to find the optimal structure. Finally, the matched structure is employed to estimate kernels. Although the method performs decently on face images deblurring problem, the matching process of the overall structure is time-consuming and the real-time performance is poor. HaCohen et al. [11] propose a face images deblurring method with the assistance of reference images, which have the same scene with the original blurred image. The functions of reference images are two-fold. On one hand, more information from the reference image can promote kernel estimation. On the other hand, powerful image local

prior is provided for the non-blind deconvolution. The algorithm performs well on deblurring class-specific images. However, reference images with the same contents as input observation have certain limitations in practical applications. Recently, Wen et al. [47] propose an image deblurring method based on the prior of patch-pise pinimal pixels. It is a generic image-oriented approach and meanwhile extends to general restoration, such as face image.

Learning-based Methods Recently, deep learning is widely applied and achieve promising performance in many fields, such as image style transferring [14, 16], underwater image processing [24, 25, 27], salient object detection [26] and image dehazing [28]. It also brings a direct inspiration to the image deblurring task, and make excellent progress. These methods can be roughly divided three categories. (i) Taking advantage of networks to acquire initial estimation, and then iteratively refine it in the traditional image deblurring framework; (ii) Combining multi-network to achieve core components (kernel/latent image estimation) in the traditional methods; (iii) Utilizing the end-to-end leaning manner to model the whole recovery process. In the first category, to obtain an initial estimation of the sharp image, Ayan [3] computes the coefficients of deconvolution filters of each image patch by utilizing a CNN. To address the non-uniform blur, Cronje [7] estimates non-uniform motion vectors through combining image patches which are learned by a CNN. In the second category, Schuler et al. [39] employ the stacked convolution layers to perform kernel estimation and deconvolution iteratively. This algorithm adopts the coarse-to-fine strategy in traditional deblurring methods. However, the network does not have good generalization to varying kernel sizes. Yan et al. [52] introduce a pre-trained neural network and a regression neural network to complete identification and restoration of blurry images respectively. Xu et al. [50] achieve image deconvolution and remove artifacts through separately training two convolution neural networks. These methods bridge the gap between the traditional and learning-based image deblurring methods, however they still involved kernel estimation. If the estimated kernels are incorrect, the latent images will have significant artifacts. For the last category of methods, Mao et al. [32] propose a symmetric encoding-decoding framework for image restoration, combining the detailed and abstract information with skip connections. Hradiš et al. [13] develop a blind deconvolution model based on CNN. The model consists of 15-layer fully convolutional networks, which is free of kernel estimation. Nah et al. [33] exploite a multi-scale CNN to decompose the complexity of the problem by ensuring that each level receives an image from the previous level with blur as input is enough to be processed by the CNN. Chrysos et al. [6] present a modified ResNet and implement it in a weakly supervised manner for face images deblurring. However, the blurry input requires an off-the-shelf face detector for pre-processing. In addition, the face detector is designed for sharp face images. Once detection mistakes occurred, it would affect the subsequent deblurring process. Inspired by the method [40] for solving the problem of image super-resolution, Jin et al. [15] propose a face image deblurring method by expanding the receptive fields based on CNN. Instead of adopting dilated convolution networks, they rely on resampling to achieve wide receptive fields. Li et al. [29] propose a two-step method for solving the problem of face blind restoration. First they utilize the WarpNet to predict the dense flow field, and then the RecNet is adopted for recover latent image. However, the image restoration method emphasizes on the problem of image degradation which caused by image compression, low resolution, noise and defocus. In conclusion, previous face image deblurring methods which usually require specifically designed priors or rely on salient edges for kernel estimation and the semantic priors are rarely considered.

3 Proposed method

In this section, we first review the basic formulation of GANs and then introduce the proposed loss functions and network architecture.

3.1 Overview of GANs

Generally, a GAN framework consists of two components: the generation model G, and the discrimination model D. The generator G receives a noise vector z as input and learns to generate synthesized images $G_\varepsilon(z)$ that can cheat the discriminator. The discriminator D takes the synthesized images $G_\varepsilon(z)$ or real data x with a distribution $P_{data}(x)$ as input and outputs a classification probability. In theory, this is a minimization problem. The D and the G are in a competitive relationship which makes the training process quite challenging. Both G and D are trained simultaneously, expecting to achieve the global optimal solution. The mathematical expression of the objective function is as follows:

$$\min_G \max_D E_{x\sim P_{data}(x)}[log D_\theta(x)] + E_{z\sim P_z(z)}[log(1 - D_\theta(G_\varepsilon(z)))], \tag{1}$$

ε and θ are parameters of G and D respectively. Although the formulation of GAN is difficult to be directly applied to the face images deblurring task, we modify these items in the above GANs expression to fit face images deblurring problem. First, instead of the random noise, we consider the blurry face images as inputs of G to generate realistic images. Second, we train G via fully supervised learning using pairs of pixel-aligned face images. Third, the experimental results are conducted by the loss functions of G and D cannot regularize the image recovery. Therefore through the analysis of GANs framework, we require extra constraint terms for training G so that the latent results having high-quality.

3.2 Loss functions

The loss function $\mathcal{L}(G, D)$ of the proposed network in (2) is composed of three parts: (1) the content loss function $\mathcal{L}_{content}$, which enforces the content similarity between the ground truth and the learned face images; (2) the perceptual loss $\mathcal{L}_{perceptual}$, which preserves the image semantic content coherence during the process of discriminative learning; (3) the adversarial loss $\mathcal{L}_{adv}$, which propels the generator to reach the desired manifold. The loss function is expressed by a simple additive form:

$$\mathcal{L}(G, D) = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{perceptual} + \sigma\mathcal{L}_{adv}, \tag{2}$$

where α, β, and σ are weights of content loss, perceptual loss, and adversarial loss respectively. A Larger value indicates the corresponding component is more significant. Generally, the task of transforming the images of the blurry manifold to the sharp one expects to recover the faithful characteristics of sharp face images. In all our experiments, according to method [20], we empirically set $\alpha = 1$, $\beta = 10$ and $\sigma = 1$ respectively.

3.2.1 Content loss

The purpose of the content loss function is that enforcing the content similarity between the ground truth and the learned face images. Therefore, our network is trained via fully supervised manner which uses pairs of pixel-aligned clear/blurry face images to force the

content similarity between them. The content loss for the generator is defined as the MSE (Mean Square Error) between the generated image $G_\varepsilon(b^i)$ and the ground truth image s^i:

$$\mathcal{L}_{\text{content}} = \frac{1}{C_k H_k W_k} \left\| s^i - G_\varepsilon\left(b^i\right) \right\|, \tag{3}$$

where k stands for the k-th layer, and C_k, H_k and W_k denote the number, height, and width of the features maps. Thus, encouraging the generator to reconstruct detailed and structural features of the clean face images, the perceptual loss function is developed.

3.2.2 Perceptual loss

Inspiring by the idea of being closer to perceptual similarity, Johnson et al. [16] propose a perceptual loss and extend it in image super-resolution. Based on this, in order to make the generated images keep semantical correspondences with the ground truths. The mathematical expression of the perceptual loss function is defined as follows:

$$\mathcal{L}_{deep} = \frac{1}{C_m H_m W_m} \left\| \psi_m\left(s^i\right) - \psi_m(G_\varepsilon(b^i)) \right\|, \tag{4}$$

where m stands for the m-th layer, and C_m, H_m and W_m denote the number, height, and width of the features maps. $\psi_m(s^i)$ represents the feature response of input s^i at the m-th layer, $\psi_m(G_\varepsilon(b^i))$ represents the feature response to the input b^i at the m-th layer. The perceptual loss for the generator is defined by punishing the l_2 loss between the generated image $G_\varepsilon(b^i)$ and the ground truth image s^i. Typically, the pre-trained VGG-19 network [41] is exploited for feature extraction and applied in many low-level computer vision problems. In the proposed method, we use the layer $Relu3\_3$ of the pre-trained VGG-19 network to extract features.

3.2.3 Adversarial loss

The adversarial loss is applied to both generator and discriminator. Its probability value represents to what extent the deblurred image generated by the generator looks like a real face image. The competitive relationship between the generator and the discriminator, correctly driving the generator to transform the blurry face image into a high-quality one. In the proposed model, we use WGAN-GP [10] as the discriminator loss function, which can stabilize the training process of the network. The mathematical expression of the WGAN-GP objective function is as

$$\begin{aligned} L_{GAN} \;=\; & E[D_\theta(s^i)] - E[D_\theta(G_\varepsilon(b^i))] \\ & -\lambda E_{\hat{x}\sim\ p_{\hat{x}}}[(\left\| \nabla \hat{x} D_\theta(\hat{x}) \right\|_2 - 1)^2] \end{aligned} \tag{5}$$

where $G_\varepsilon(b^i)$ is the generated image, s^i is the real image, $E[D_\theta(G_\varepsilon(b^i))]$ and $E[D_\theta(s^i)]$ implicate the expectation of the discriminator assigns correct labels to the generated image $G_\varepsilon(b^i)$ and the clean image s^i, respectively. $\lambda E_{\hat{x}\sim\ p_{\hat{x}}}[(\left\| \nabla \hat{x} D_\theta(\hat{x}) \right\|_2 - 1)^2]$ represents samples as a linear combination of the real image and the generated image points.

3.3 Network architecture

Encoder-decoder network architectures are successfully applied in the community of computer vision, such as video deblurring [42] [48], biomedical image segmentation [38]. The encoder-decoder network usually consists of symmetric multi-level stacked convolution and

deconvolution layers. The benefits of using U-Net are as follows. Firstly, the input images can be represented as feature maps at multi-level progressively, and the encoder extracts features ranging from figuration to abstraction. Secondly, the decoder symmetrically restores detailed information from the corresponding encoder. Thirdly, the related detailed and abstract features are widely combined with skip connections. The skip connections benefit gradients propagation and facilitate convergence, not only the abstract information but the detailed features need to be incorporated. Therefore, the U-Net network which includes convolution and deconvolution layers is beneficial for image deblurring.

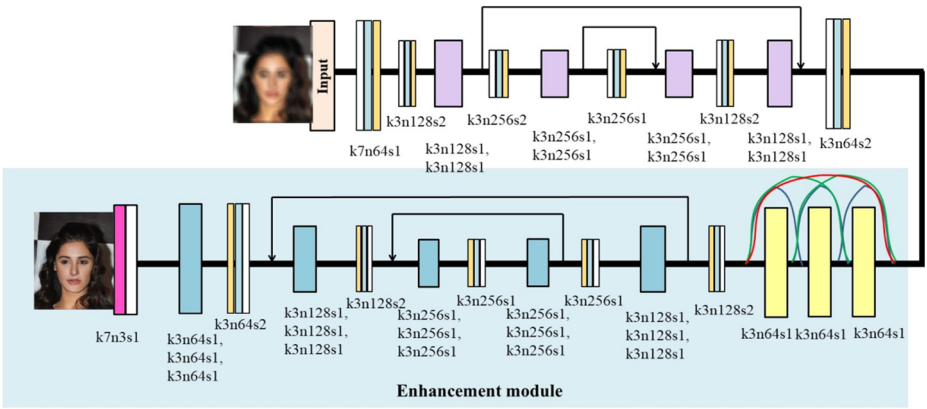
However, we found that directly using the U-net for face images deblurring could not generate decent performance. The potential reasons for this phenomenon are illustrated as follows. First, since the task of face images deblurring is a highly ill-posed problem, it requires large receptive fields to capture more features for latent images recovery. Thus, a straightforward approach is to stack more convolution layers. However, this idea is less feasible in practice, as the overly deep network discourages the parameter configuration and convergence. Secondly, though the method of [38] has achieved amazing results in image recognition, as for image deblurring task, the image features need to be closely related to those features learned before. Therefore, according to the above analysis, we modify the U-Net model to handle the problem of face images deblurring.

Generator Instead of directly employing the original U-Net architecture, we propose a modified version of the U-net based network. An overview of the generator as shown in Fig. 2. Unlike previous U-net based network, we elaborately design a two-step generator which contains an enhancement module. At the first stage the generator achieves the encoding and the decoding process, at the second the reconstructed feature representations obtained by the decoder are delivered in the enhancement module for reinforcing feature representations and constructing complicated high-order maps. It should be noted that the enhancement module is inserted in the generator rather than another subnetwork. There are several advantages to this: first, instead of downsampling the results from the first stage, the reconstructed feature representations can directly reinforce; second, repeated downsampling operations result in loss of details; third, simultaneous training strategy can reduce the difficulty of training network and facilitate network convergence.

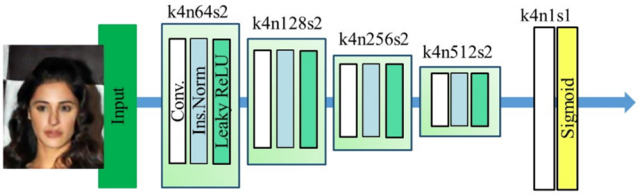
Specifically, at the sides of the encoder path, each level contains one convolution layer followed by three RRRB. The convolution layer doubles the number of feature maps from the previous layer and downsamples the resolution to the half size with a stride of 2. All convolution layers have the same number of kernels. At the sides of the decoder path, each scale is symmetric to that of encoder. It contains three RRRB followed by a deconvolution layer. The deconvolution layer is used to double the spatial size of resolution and recover the details of image content. Furthermore, we introduce an enhancement module to enhance the feature representations. The enhancement module is consists of two parts, the first part is aiming at reinforcing the reconstructed features by the improved Dense-in-Residual Dense Blocks Unit (**DRDBU**), which includes three **RRDB** blocks connected in the form of dense connections [12]. The second part is delivering the strengthened features to the second-stage of encoding and decoding processing. It should be pointed out that the network structure keeps identical parameters except for the Residual-in-Recurrent Residual Block (**RRRB**). Considering the memory requirements, we replace the **RRRB** as **Residual Block** (**ResBlock**). Finally, a convolution layer is attached after the U-Net model, and a Tanh activation function activates the last convolution layer. There are some skip connections between deconvolution layers and corresponding convolution layers. These connections facilitate gradient propagation and convergence as well as benefit to the restoration of

face images. The synthesized images which are learned from the generator are considered deblurred images. The structure and parameters configuration details of the generator subnetwork are shown in Fig. 2. Next, we introduce its basic building blocks.

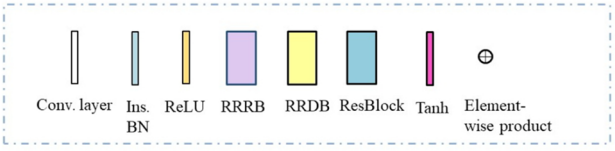
(1) **Residual Block (ResBlock)** To keep and enhance the feature maps obtained by downsample and upsample operations, we incorporate a ResBlock after downsample operations and before upsample operations, the proposed ResBlock is depicted in Fig. 3. Each ResBlock consists of three Conv-instance normalization [44]-ReLU [31] groups which arranged in residual form. Using ResBlock enables deeper architecture compared to a plain



(a) Network architecture of generator subnetwork

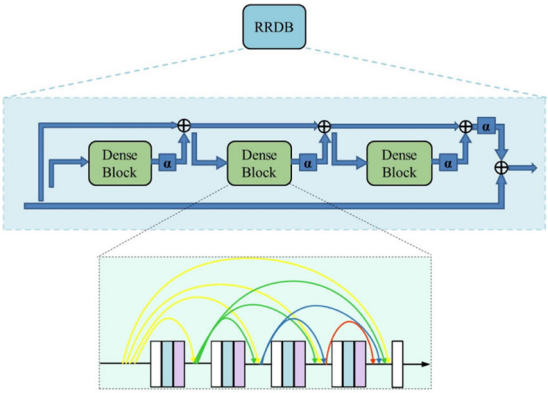


(b) Network architecture of discriminator subnetwork

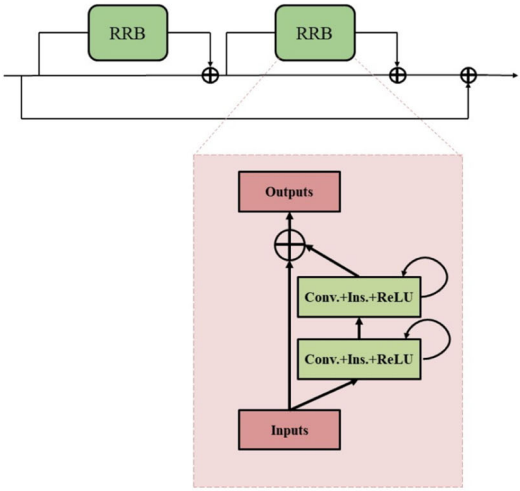


(c) Illustration of network layers

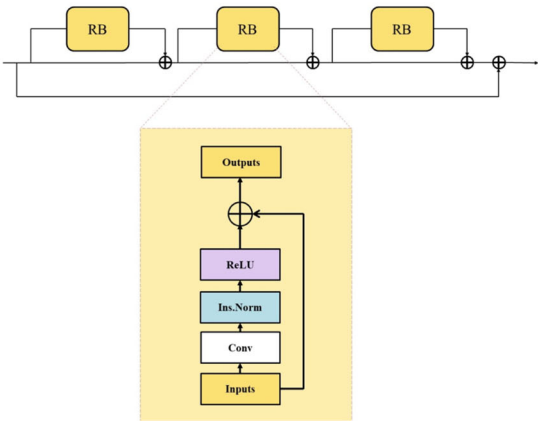
Fig. 2 Overview of the proposed GANs-based semantic face images deblurring network. The proposed network consists of two sub-networks: a U-Net based generator network and a discriminator network. The U-Net generator network plays the role of face images deblurring, and the generator network receives the supervision from the pixel-wise content loss. In addition, the discrimination network is used to judge the input image and output a probability for indicating whether the image is a real image



(a) DRDBU



(b) RRRB



(c) ResBlock

Fig. 3 DRDBU block, RRRB block and ResBlock block utilized in the generator. **a** DRDBU block is adopted in the proposed generator and α is the residual scaling parameter. **b** RRRB block is employed in the proposed generator. **c** ResBlock block is exploited in the proposed generator

convolutional layer. Each residual block uses the instance normalization (Ins.Norm), where the reparametrization on the input sample rather than the batch.

(2) **Dense-in-Residual Dense Blocks Unit (DRDBU)** Based on the experimental observations that more layers and connections always facilitate image deblurring performance. Although the RRDB has the ability to construct complex features in image super-resolution [46], we further improve it by densely connecting each RRDB. Specifically, as shown in Fig. 3a, the improved **DRDBU** has a dense-in-residual-in-residual structure, where dense connections are used in local and global. Both in the local dense blocks and the global dense connections, all intermediate features are encouraged to reuse and strength as well we alleviate the vanishing gradient problem. Therefore, benefiting from the dense connections in the local paths and the global scope, the performance of the network is facilitated. In addition, residual scaling is employed to scale the residuals by multiplying a constant α ranges from 0 to 1 before they are incorporated to the next RRDB for facilitating stability.

(3) **Residual-in-Recurrent Residual Block (RRRB)** As shown in Fig. 3b, the **RRRB** consists of two recurrent residual blocks arranged in residual connectivity. The recurrent residual block operations can be demonstrated as the improved-residual networks in [1]. More importantly, the recurrent convolutional layer inside the recurrent residual block can be considered as the iterative enhancement operation. Therefore, to enhance the feature representations, we replace the convolutional layers with recurrent convolutional layers. In the proposed method, in each recurrent residual block, every recurrent convolutional layer recycles twice.

Discriminator Our discriminator is shown in Fig. 2b, the generated image or a real image as input is randomly fed into the discriminator, and the discriminator outputs a probability in the range of [0, 1] that indicates the input image is real face images or not. Different from the high-level image processing task (object classification), image sharpness discrimination relies on local features of the image. Instead of a general full-image discriminator, we adopt PatchGAN [14] for the discriminator D. The D starts with a flat convolutional layer, after which adopts three stride convolutional layers to reduce the resolution and encode essential local features for classification. Afterward, a convolutional layer and a Sigmoid activate function are employed to obtain the classification response. Instance normalization and Leaky ReLU are attached after each convolutional layer.

4 Experiments

In this section, we first describe the training details and parameter settings which are utilized in our experiments. Then, we compare our proposed methods with state-of-the-art deblurring methods by qualitative and quantitative evaluation.

4.1 Training details and parameter settings

Training datasets We collect 8000 face images from the dataset of CelebA [30]. The method of [4] is employed to generate 800 blur kernels with different exposure vectors, the kernel size ranging from 13×13 to 27×27. In each kernel size, every 400 images are convolved with 5 different kernels, totally yielding 320000 pairs of synthetic blurry face images and corresponding sharp face images for training. Then the blurry and sharp face images are randomly cropped into 150×150 patches as the training dataset and ground truth respectively. We set the patch size based on empirical experiments on the training process, and

larger patch size requires considerable computational budget. There is no overlap between the training dataset and the test dataset.

Parameter settings We implement the developed network using the TensorFlow framework with Python interface. The network is trained on NVIDIA Titan 1080 Ti GPU for 3000K iterations using a batch size of 16, and the learning rate is 0.0001. The slope of the Leaky ReLU is 0.2. The parameters of the networks are optimized using the Adam optimizer [18] with momentum terms β_1=0.5 and β_2=0.999. The growth rate of the dense blocks in the RRDBs is set as 8, α is 0.2, and 3 RRDBs are adopted. We set the trade-off weights based on empirical experiments on the training process. The experimental parameters are identical in all experiments.

4.2 Comparisons with the state-of-the-art methods

In order to verify the effectiveness and generalization of the proposed method, we make inference on different synthetic face image datasets and real-world inputs. Besides, we also compare our method with state-of-the-art methods. The source codes of these methods are provided by their authors, and we recurrent these methods according to the original configuration parameters.

Synthetic Blurry Face Images We validate our method on synthetic images by quantitative evaluation manner. First, we randomly select 200 images from CelebA and Helen [22] respectively. Each of 25 images is convolved with 8 different kernels in kernel size from 13 $\times$ 13 to 27 $\times$ 27, totally 1600 synthesized blurry face images as test dataset. We compare our model with the methods of the MAP-based framework [35, 36, 47, 51, 54], and deep-learning based method [33][21][15].

We report the PSNR and SSIM [45] values on the datasets of CelebA and Helen in Table 1, results about the specific kernel size and the corresponding performance on different blurry face images are shown in Fig. 4. Compared with other methods, our proposed method achieves better performance on different kernel sizes and datasets. Visual results comparisons on dataset CelebA and Helen are shown in Figs. 5 and 6 respectively. Since the methods of hand-crafted priors [51], [36], [54] are designed for generic images rather than face images, the results obtained by these methods always arise ringing artifacts. [35]

Table 1 Performance comparison with the-state-of-the-art methods by two quantitative evaluation metrics

Method	CelebA		Helen	
	PSNR	SSIM	PSNR	SSIM
Jin et al. [15]	25.9943	0.8053	24.1320	0.7854
Kupyn et al. [21]	21.6635	0.7073	20.3665	0.7465
Pan et al. [35]	19.4853	0.6964	20.1337	0.7353
Pan et al. [36]	19.9021	0.7121	20.6253	0.7693
Zhong et al. [54]	17.3053	0.6221	18.4043	0.6865
Xu et al. [51]	20.1932	0.7174	20.4282	0.7493
Nah et al. [33]	21.3335	0.7764	22.0790	0.7862
Wen et al. [47]	19.2765	0.6024	20.1252	0.6453
Ours	26.1273	0.8176	24.3253	0.7952

We compute the average PSNR(dB) and SSIM values on the test dataset of CelebA and Helen

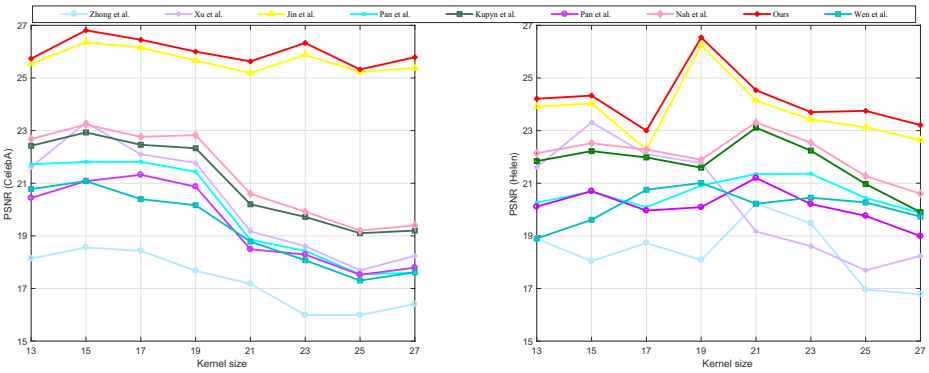


Fig. 4 Quantitative evaluation on different blur kernels. 200 images are randomly selected from the test datasets of CelebA and Helen, and 8 sizes of kernels are generated by [4]. Each of 25 images is convolved with 4 different kernels, and total of 1600 images are prepared for evaluation

is tailored for the face image deblurring task, however the external exemplar references developed by the authors are impossible to cover all facial contours and expressions. The results of [35] generate overly smooth results rather than having photo-realistic quality. [47] attempts to deblur face images with the priors that developed from natural images, however it stills a challenging task. Due to the lack of consideration for face image structure, the visual results of [21][33] still need improvement in image detail recovery. Furthermore, we bring in the image deblurring method which specially tailored for face images. Their results [15] are considerable to ours in qualitative and quantitative evaluation. The proposed model is tested on two different face datasets, and it proves that the proposed method has a good ability of generalization for different face images and blur kernels size. Specifically,

(a) GT (b) Input (c) [21] (d) [15] (e) [35] (f) [36] (g) [54] (h) [51] (i) [47] (j) [33] (k) Ours

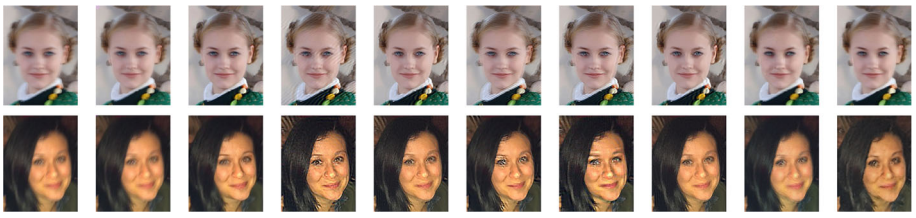
Fig. 5 Visual comparison of synthesized blurry face images with the-state-of-the-art methods on test dataset CelebA. The results of the proposed method leads to more clearer and visually pleasant results with fewer artifacts. (**zoom in for best view**)

(a) GT (b) In-put (c) [21] (d) [21] (e) [35] (f) [36] (g) [54] (h) [51] (i) [47] (j) [33] (k) Ours

Fig. 6 Visual comparison of synthesized blurry face images with the-state-of-the-art methods on test dataset Helen. The results of the proposed method leads to more clearer and visually pleasant results with fewer artifacts. (**zoom in for best view**)

in Fig. 5 the degraded images contain both the face region and the complicated background, our method can generate fewer artifacts and provide better details in the key parts such as mouths, eyes, and wrinkles.

Real-world images In order to further verify the effectiveness and generalization of the proposed method, we also test our method on real-world blurred images. As shown in Fig. 7, the methods [51, 54] rely on the hand-crafted priors and developed on the MAP-framework. Different priors need to be developed according to specific image deblurring task requirements. Due to the limitations of the applicability of the methods, they are incapable of deblurring face images, both the background and the face region have severe ringing artifacts. Although [35] develop external exemplar references, they merely extract face contour rather than learn the structural and detailed features. The results of [35] are still found to generate an overly smooth result on the "girl" image, it looks like an oil painting-like image rather than a photo-realistic image. Compared to other methods [51, 54] developed from the hand-crafted prior, [47] can achieve image deblurring to a certain extent. However



(a) In-put (b) [21] (c) [15] (d) [35] (e) [36] (f) [54] (g) [51] (h) [47] (i) [33] (j) Ours

Fig. 7 Visual comparison on real-world blurry face images. (**zoom in for best view**)

wen [47] fails to avoid the effect bring by the kernel estimation, and the ringing artifacts emerge at the regions of the collar in the "girl" image. Due to the lack of features enhancement process, The results of "girl" image obtained by Kupyn et al. [21] is close to ours, however the other deburred image is worse than ours. Furthermore, the results are overly smoothed which as the same as [35]. The method [36] employs dark-channel prior, however sometimes converges to a local minimum that causes the incorrect kernel estimation and final ringing artifacts image. Although the method [33] has large receptive fields, distortion appears in the eyes of the deblurred "girl" image. From the results of qualitative and quantitative evaluation, the proposed method is comparable with the method of Jin [15]. Due to the limited real-world images, our model is trained on synthetic blurry face images dataset. Our method generates images with fewer artifacts compared with the methods [33, 36]. It further illustrates that the synthetic dataset is helpful to restore real-world images.

5 Analysis and discussion

In this part, we study the loss functions and the network architecture of the proposed method by performing ablation studies.

5.1 Ablation study of each building blocks in the proposed network

To demonstrate the effectiveness obtained by the building blocks employed in the proposed method, we perform an ablation study which involves the following four experiments: the proposed network without enhancement module (**Net w/o EM**); the proposed network without ResBlock, and replace it with two convolutional layers (**Net w/o ResBlock**); the proposed network without DRDBU (**Net w/o DRDBU**); the proposed network without RRRB (**Net w/o RRRB**).

In these experiments, we use identical parameter settings with the proposed method. Quantitative evaluation is conducted on **CelebA dataset**. As shown in Table 2 and Fig 8, (1) removing the enhancement module, the performance of the proposed network considerably decreases, which implies the enhancement module makes contributions to reinforcing feature representations and constructing complicated high-order maps; and (2) removing the building blocks of **ResBlock**, **DRDBU**, and **RRRB**, the performance of the network slightly decreases, which manifest these three building blocks also make contributions to our network.

5.2 Ablation study of loss functions in the proposed network

To demonstrate the effectiveness obtained by loss functions in the proposed method, we conduct an ablation study which involves the following two experiments:

Table 2 An ablation study of removing/changing building blocks in the proposed method

We compute the average PSNR(dB) and SSIM values on the dataset of CelebA

Methods	PSNR	SSIM
Net w/o EM	25.3254	0.7794
Net w/o DRDBU	25.6475	0.7832
Net w/o RRRB	25.8319	0.7905
Net w/o ResBlock	25.9876	0.7944
Ours	26.1273	0.8176

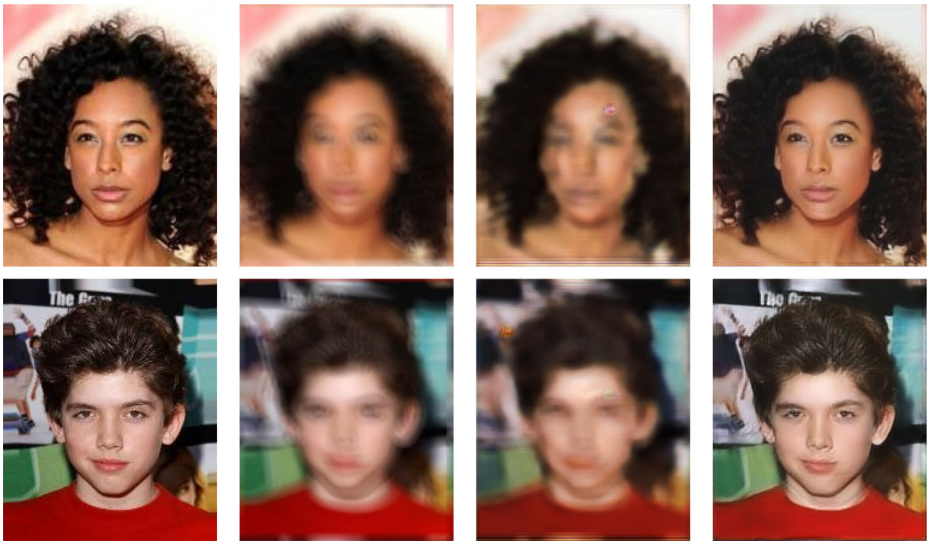
(a) Blurry image (b) Net w/o ResBlock (c) Net w/o RRB (d) Net w/o DRDBU (e) Net w/o EM (f) Ours

Fig. 8 Effect of different building blocks. The visualization blurry face images are processed by different building blocks. The proposed network leads to clearer and more visually realistic results. **a** Blurry inputs, **b** The result of Net w/o ResBlock, **c** The result of Net w/o RRRB, **d** The result of Net w/o DRDBU, **e** The result of Net w/o EM, **f** Ours

Table 3 An ablation study of removing/changing loss functions in the proposed method

Methods	CelebA PSNR	CelebA SSIM
Net w GANl + conl	22.9319	0.6634
GANl + conl + perl(proposed)	26.1273	0.8176

We compute the average PSNR(dB) and SSIM values on the dataset of CelebA



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

Fig. 9 Effect of different loss functions. The visualization blurry face images are processed by different loss functions. The total loss leads to clearer and more visually realistic results. **a** Ground truth **b** Blurry inputs **c** GANs loss + Content loss **d** Total loss

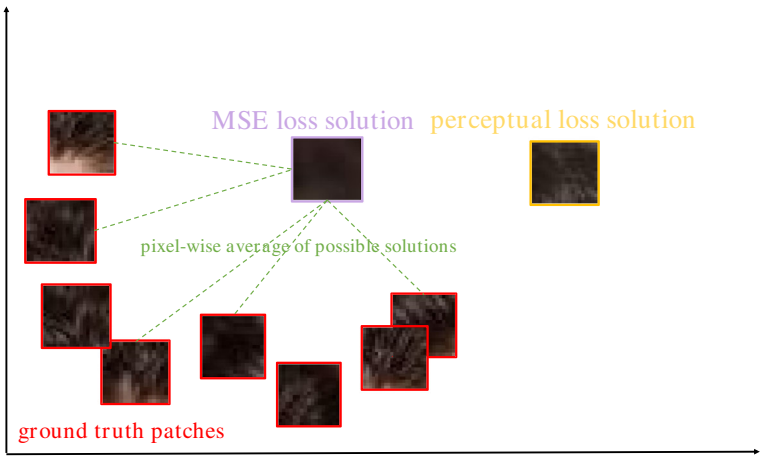


Fig. 10 Illustration of patches from the ground truth images (red) and deblurred face images patches obtained with MSE (purple) and visual loss functions (orange). The MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space, while visual loss functions force the recovery results towards the realistic face images producing perceptually pleasant results. (The images patches are hairs, **zoom in for best view**)

The proposed network with GAN loss function and content loss function (**Net w GANl + conl**); the proposed network with perceptual loss function (**Net w GANl + conl + perl**);

In these experiments, we use identical parameter settings and network architecture with the proposed method. Quantitative evaluation is conducted on **CelebA dataset**. We calculate the PSNR and SSIM values on deblurred results obtained from models trained according to different loss functions in Table 3. The generated face images based on the content loss and GANs loss are shown in Fig. 9c. The overall visual effect of Fig. 9c is of no difference from blurry input Fig. 9b. Apparently, the results obtained by this loss function doesn't deblur at all. Fig. 10 shows that the MSE-based solution appears overly smooth due to the pixel-wise average of possible solutions in the pixel space. The structural and detailed features are beyond the learning range of the generator. As shown in Fig. 9d, the quantitative values of PSNR and SSIM significantly increase when introducing the perceptual loss, which demonstrates that the perceptual loss determines the performance of our network.

Limitations Our method may fail when the blurry images are in the condition of low-light, as shown in Fig. 11. Future work includes increasing the performance of face images deblurring in low-light or highly blurry conditions.

Fig. 11 Failure case. The proposed face deblurring method may not robust to the images in the low-light condition

(a) Blurry input

(b) Deblurred image

6 Conclusions

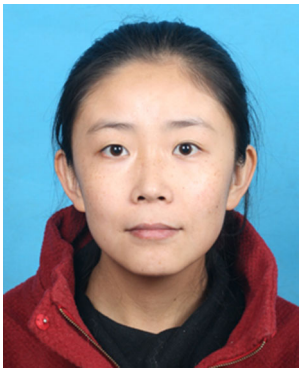
In this work, we presented a solution for blind face images deblurring problem. To solve this problem, we proposed a data-driven method that semantically generates sharp face images via GANs. The proposed method takes advantage of adversarial learning to generate images with similar statistics to those of sharp face images. In addition, according to the properties of the face images, we adopt a perceptual loss function to recover the face images. We further develop a generator that includes image encoding and decoding process as well as an enhancement process step to reinforce feature representation and encourage feature propagation. Finally, the experiments demonstrate that the performance of the proposed method is comparable with state-of-the-art face images deblurring methods.

Acknowledgements This document is the results of the research project funded by the National Science Foundation of China Grant No.61771334, the ChunHui project, Ministry of education, China No.Z2016105.

References

1. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK (2018) Improved inception-residual convolutional neural network for object recognition
2. Anwar S, Phuoc Huynh C, Porikli F (2015) Class-specific image deblurring
3. Ayan C (2016) a neural approach to blind motion deblurring
4. Boracchi G, Foi A (2012) Modeling the performance of image restoration from motion blur. IEEE Trans Image Proc 21:3502–3517
5. Cho S, Lee S (2009) Fast motion deblurring
6. Chrysos GG, Zafeiriou S (2017) Deep face deblurring
7. Cronje J (2015) Deep convolutional neural networks for dense non-uniform motion deblurring
8. Fergus R, Singh B, Hertzmann A, Roweis ST, Freeman WT (2006) Removing camera shake from a single photograph. ACM transactions on graphics 25:787–794
9. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y et al (2014) Generative adversarial nets, Proc
10. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans
11. Hacohen Y, Shechtman E, Lischinski D (2013) Deblurring by example using dense correspondence
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition
13. Hradiš M., Kotera J, Zemcík P., Šroubek F. (2015) Convolutional neural networks for direct text deblurring. Proc British machine vis Conf 10:2
14. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks
15. Jin M, Hirsch M, Favaro P, fast Learningfacedeblurring, wide Proc. IEEEConf. Comput. Vis. (2018) Patt Recog. workshops
16. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution
17. Joshi N, Szeliski R, Kriegman DJ (2008) Psf estimation using sharp edge prediction
18. Kingma DP, Ba J (2014). arXiv:1412.6980
19. Krishnan D, Tay T, Fergus R (2011) Blind deconvolution using a normalized sparsity measure
20. Kupyn O, Budzan V, Mykhailych M, Mishkin D, Matas J (2018) Deblurgan: Blind motion deblurring using conditional adversarial networks
21. Kupyn O, Martyniuk T, Wu J, Wang Z (2019) Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better
22. Le V, Brandt J, Lin Z, Bourdev L, Huang TS (2012) Interactive facial feature localization
23. Levin A, Weiss Y, Durand F, Freeman WT (2009) Understanding and evaluating blind deconvolution algorithms
24. Li C, Anwar S, Porikli F (2019) Underwater scene prior inspired deep underwater image and video enhancement Pattern Recognition
25. Li C, Cong R, Hou J, Zhang S, Qian Y, Kwong S (2019). arXiv:1901.05495
26. Li C, Cong R, Hou J, Zhang S, Qian Y, Kwong S (2019). arXiv:1906.08462

27. Li C, Guo J, Guo C (2018) Emerging from water: Underwater image color correction based on weakly supervised color transfer. IEEE Signal processing letters 25(3):323–327
28. Li C, Guo C, Guo J, Han P, Fu H, Cong R (2019) PDR-Net Perception-Inspired Single Image Dehazing Network with Refinement
29. Li X, Liu M, Ye Y, Zuo W, Lin L, Yang R (2018) Learning warped guidance for blind face restoration
30. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild
31. Maas AL, Hannun AY, Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. Proc Int Conf Machine Learn 30:3
32. Mao X, Shen C, Yang YB (2016) Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections
33. Nah S, Kim TH, Lee KM (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring
34. Nishiyama M, Hadid A, Takeshima H, Shotton J, Kozakaya T, Yamaguchi O (2011) Facial deblur inference using subspace analysis for recognition of blurred faces. IEEE Trans patt analy machine intel 33:838–845
35. Pan J, Hu Z, Su Z, Yang MH (2014) Deblurring face images with exemplars
36. Pan J, Sun D, Pfister H, Yang MH (2016) Blind image deblurring using dark channel prior
37. Ren W, Cao X, Pan J, Guo X, Zuo W, Yang MH (2016) Image deblurring via enhanced low-rank prior. IEEE Trans. Image Proc. 25:3426–3437
38. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation
39. Schuler CJ, Hirsch M, Harmeling S, Schölkopf B. (2016) Learning to deblur. IEEE Trans. patt. analy. machine intel. 38:1439–1451
40. Shi W, Caballero J, Huszr F, Totz J, Aitken AP, Bishop R et al (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network
41. Simonyan K (2014). arXiv:1409.1556
42. Su S, Delbracio M, Wang J, Sapiro G, Heidrich W, Wang O (2017) Deep video deblurring for hand-held cameras. In: Proc IEEE Int Conf Comput Vis, p 6
43. Sun L, Cho S, Wang J, Hays J (2013) Edge-based blur kernel estimation using patch priors
44. Ulyanov D, Vedaldi A, Lempitsky V (2016). arXiv:1607.08022
45. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
46. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C et al (2018) Esrgan: Enhanced super-resolution generative adversarial networks
47. Wen F, Ying R, Liu P, Truong TK (2019) Blind Image Deblurring Using Patch-Wise Minimal Pixels Regularization. arXiv:1906.06642
48. Wieschollek P, Hirsch M, Schölkopf B., Lensch HP (2017) Learning blind motion deblurring
49. Xu L, Jia J (2010) Two-phase kernel estimation for robust motion deblurring
50. Xu L, Ren JS, Liu C, Jia J (2014) Deep convolutional neural network for image deconvolution
51. Xu L, Zheng S, Jia J (2013) Unnatural l_0 sparse representation for natural image deblurring
52. Yan R, Shao L (2016) Blind image blur estimation via deep learning. IEEE Trans. Image Proc. 25:1910–1921
53. Zhang H, Yang J, Zhang Y, Huang TS (2011) Sparse representation based blind image deblurring
54. Zhong L, Cho S, Metaxas D, Paris S, Wang J (2013) Handling noise in single image deblurring using directional filters



Qing Qi

Jichang Guo

Chongyi Li

Lijun Xiao