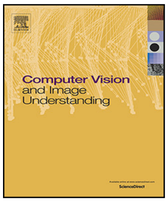
Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Highlights

Video Deblurring via Spatiotemporal Pyramid Network and Adversarial Gradient Prior

Computer Vision and Image Understanding xxx (xxxx) xxx

Tao Wang, Xiaoqin Zhang*, Runhua Jiang, Li Zhao, Huiling Chen, Wenhan Luo

- We propose a spatiotemporal pyramid module as a new tool to model spatiotemporal dynamics within the video for the specific video deblurring task.
- We introduce the gradient space of the image into the discriminator in GAN. With the goal of fooling the discriminator in the differential space, it is easier for the deblurring method to generate sharp videos.
- The proposed methods achieve the state-of-the-art results by comparing with the existing methods on benchmark datasets.

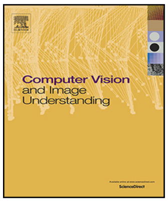
Graphical abstract and Research highlights will be displayed in online search result lists, the online contents list and the online article, but will not appear in the article PDF file or print unless it is mentioned in the journal specific style requirement. They are displayed in the proof pdf for review purpose only.

Please cite this article as: T. Wang, X. Zhang, R. Jiang et al., Video Deblurring via Spatiotemporal Pyramid Network and Adversarial Gradient Prior. Computer Vision and Image Understanding (2020) 103135, https://doi.org/10.1016/j.cviu.2020.103135.

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Video Deblurring via Spatiotemporal Pyramid Network and Adversarial Gradient Prior

Tao Wang [a], Xiaoqin Zhang [a,*], Runhua Jiang [a], Li Zhao [a], Huiling Chen [a], Wenhan Luo [b]

[a] *College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, 325035, China*
[b] *Tencent AI Lab, Shenzhen, 518000, China*

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Video deblurring
Spatiotemporal pyramid
Adversarial learning
Image prior

ABSTRACT

Video deblurring is to restore sharp frames from a blurry sequence. It is a challenging low-level vision task because the blur caused by camera shake, object motions and depth variations is heterogeneous in both spatial and temporal dimensions. Traditional methods usually work on a fixed spatiotemporal scale. However, the spatiotemporal scale of blurs in the video can vastly vary in the real-world situation. To address this challenge, we propose a Spatiotemporal Pyramid Network (SPN) to dynamically learn different spatiotemporal cues for video deblurring. Specifically, inside SPN, a spatiotemporal pyramid module is employed to effectively capture both spatial information and temporal information from the blurry sequence in a pyramid mode. An image reconstruction module constructs the sharp center frame through the obtained spatiotemporal information. Additionally, inspired by the statistical image prior and adversarial learning, we extend SPN and propose a Spatiotemporal Pyramid Generative Adversarial Network (SPGAN), which conducts adversarial discrimination in the gradient space. It helps the network produce more realistic sharp video frames. Experiments conducted on benchmarks demonstrate that the proposed methods achieve state-of-the-art results in terms of PSNR, SSIM and visual quality.

1. Introduction

Deblurring is a typical problem for low-level vision tasks because of its importance in both the academic community and industrial applications. The low-quality image or video causes visually poor quality, which hampers some high-level vision tasks (Zhang et al., 2020c; Lee et al., 2011). In general, great progress has been achieved in the problem of image deblurring, while the problem of video deblurring has not been well explored because of its relatively complicated settings. Compared with image deblurring, the modeling of temporal information should be handled appropriately.

Naively, one can apply modern methods of single image deblurring to each single frame one by one to obtain the results as the deblurred video frames. However, the independence among the processing of multiple frames would inevitably introduce artifacts into the resulting video. To solve this problem, some traditional methods model the temporal information by explicitly forcing the continuous frames to be consistent after warping. However, there exist several drawbacks to these methods. First, forcing continuous frames to be consistent with warping involves many heuristics. For example, it requires the estimation of motion between continuous frames, which is time-consuming. And there are occlusion areas even if the motion is estimated correctly. Second, this kind of solution is typically very complicated and not end to end. It is difficult to tune and diagnose.

With the popularity of deep learning, several video deblurring approaches that using deep neural networks are proposed. For example, Su et al. (2017) propose a neural network that stacks five successive frames as inputs to produce the center sharp frame. Hyun Kim et al. (2017) concatenate the multi-frame features to recover the current frame by a deep recurrent network. These recent deblurring approaches often work on a fixed spatiotemporal scale and do not fully use spatial information from the center blurry frame. However, the spatiotemporal scale of blurs can vastly vary. This is because that most of the camera shake blur is short, spatially uniform and temporally uncorrelated (Su et al., 2017; Xu et al., 2014), while object motion causes long, spatially localized and temporally smooth blurs (Pan et al., 2016). Therefore, we propose a spatiotemporal pyramid module to process the input frames both in spatial and temporal dimensions via a pyramid mode that taking the advantage of the fact that the spatiotemporal scale of blurs in a video can vary vastly. The spatiotemporal pyramid module works on different temporal scales. It first takes the middle frame as the center, and divides five successive input frames into three different lengths in pyramid mode. Then it uses 2D convolution to process the center blurry frame, and captures temporal information from the

* Corresponding author.
E-mail addresses: taowangzj@gmail.com (T. Wang), zhangxiaoqinnan@gmail.com (X. Zhang), ddghjikle1@gmail.com (R. Jiang), lizhao@wzu.edu.cn (L. Zhao), chenhuiling.jlu@gmail.com (H. Chen), whluo.china@gmail.com (W. Luo).

https://doi.org/10.1016/j.cviu.2020.103135
Received 13 May 2020; Received in revised form 17 August 2020; Accepted 3 November 2020
Available online xxxx

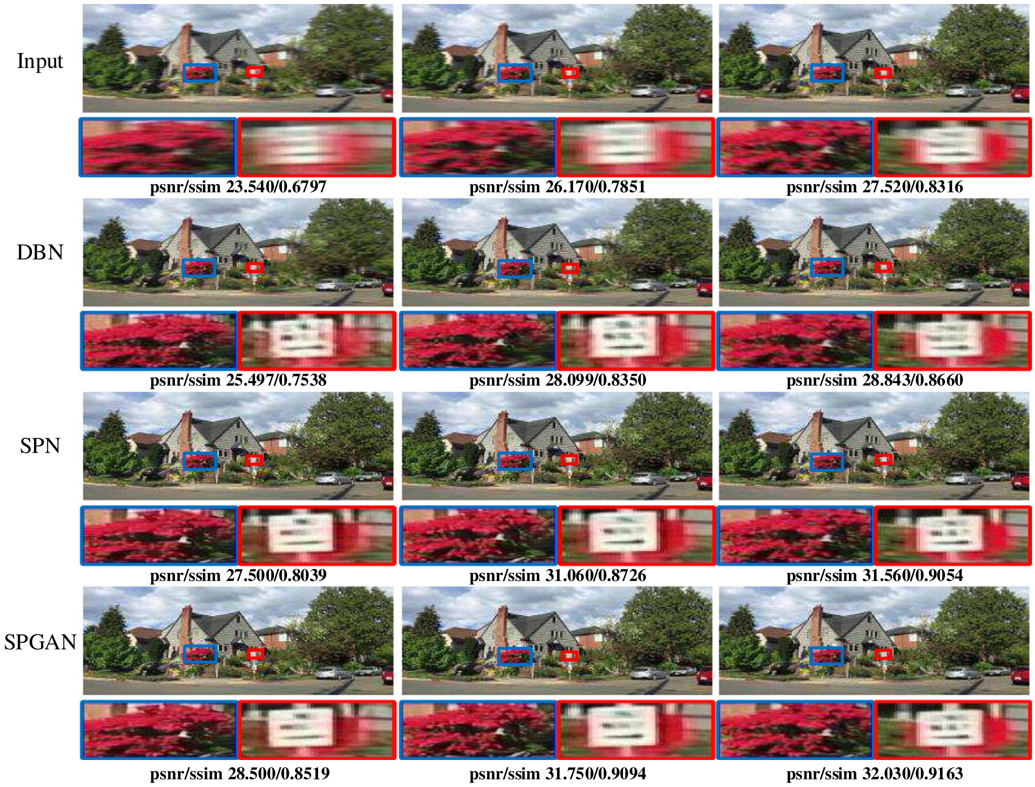


Fig. 1. Exemplar video frames on the DVD dataset (Su et al., 2017) processed by DBN (Su et al., 2017), SPN and SPGAN. The input blurry frames are in the first row. The second row illustrates the deblurring results of DBN. The third row displays the deblurring results of SPN. The fourth row shows the deblurring results of SPGAN. The values of PSNR and SSIM are shown at the bottom of each frame. The higher values demonstrate better performance.

successive subframes via 3D convolution. Finally, it dynamically fuses the spatiotemporal information to sharpen blurry ones. We put this module in front of the network to help the network (SPN) not only focus on the spatial information of the center blurry frame, but also learn the different scales of temporal information from nearby frames. With the help of this module, the deblurring performance of SPN is improved, as Fig. 1 shows.

Moreover, with the emergence of generative adversarial networks, the problem of video deblurring has been addressed with the help of generative adversarial networks by existing methods to boost the realness of the deblurred video. This is conducted by discriminating the generated videos against the real-world sharp videos via an adversarial loss. This is effective in some cases, but with only limited improvement. We suspect that this is partly due to the following reason. The space of natural images can be hardly covered by the limited number of training examples for video deblurring. Thus the volume within which the discrimination between deblurred videos and real-world videos is conducted in fact is limited in the whole visual image space. To this end, we propose to alternatively discriminate the generated videos against real-world videos in the differential space of images. Existing researches (Chen et al., 2019; Pan et al., 2014) have shown that gradient images of sharp images and blur images are completely different. Due to the blurring process, the values of neighboring pixels tend to be closer to each other. Thus, for sharp images, values in their gradient images are usually greater than values in the gradient images of blurred images. In other words, the distribution of gradient images of sharp images is different from those corresponding to blurred images, and the blurring effect is well-observed in the gradient of a blurred image. In this paper, we discriminate the combination of gradient images against those of the real sharp videos. The insight behind is, the freedom degree of differential space of image is much smaller than that of the original visual image space, thus it is easier for the deblurring algorithm to generate videos with the gradient images sharing the same space with those from sharp videos. We call it as the adversarial gradient prior. Fig. 1 shows exemplar results.

Our main contributions are threefold. First, we propose a spatiotemporal pyramid module as a new tool to model spatiotemporal dynamics within the video for the specific video deblurring task. Second, we introduce the gradient space of the image into the discriminator of the generative adversarial network. With the goal of fooling the discriminator in the differential space of image, it is easier for the deblurring method to generate sharp videos. Finally, the proposed methods achieve state-of-the-art results by comparing with the existing methods on benchmark datasets.

The rest of the paper is organized as follows. Section 2 briefly introduces the related work of deblurring. Section 3 represents the proposed method. Experimental results are reported in Section 4. Conclusions are drawn in Section 5.

2. Related work

As the video deblurring problem is closely related to the problem of image deblurring, we introduce related work of both image deblurring and video deblurring as follows.

2.1. Image deblurring

Image deblurring aims at recovering a sharp image X given a blurred image Y. The blur process is usually modeled by convolution operation with a kernel K plus additive noise E, *i.e.,* $Y = X * K + E$, where $*$ means the convolution operation. The image deblurring can be classified into blind deblurring and non-blind deblurring depending on whether the kernel is known (non-blind) (Yuan et al., 2008; Joshi et al., 2009; Cho et al., 2011; Schuler et al., 2013; Schmidt et al., 2013; Javaran et al., 2017) or not (blind) (Shan et al., 2008; Krishnan et al., 2011; Levin et al., 2011; Babacan et al., 2012; Wang et al., 2018). The

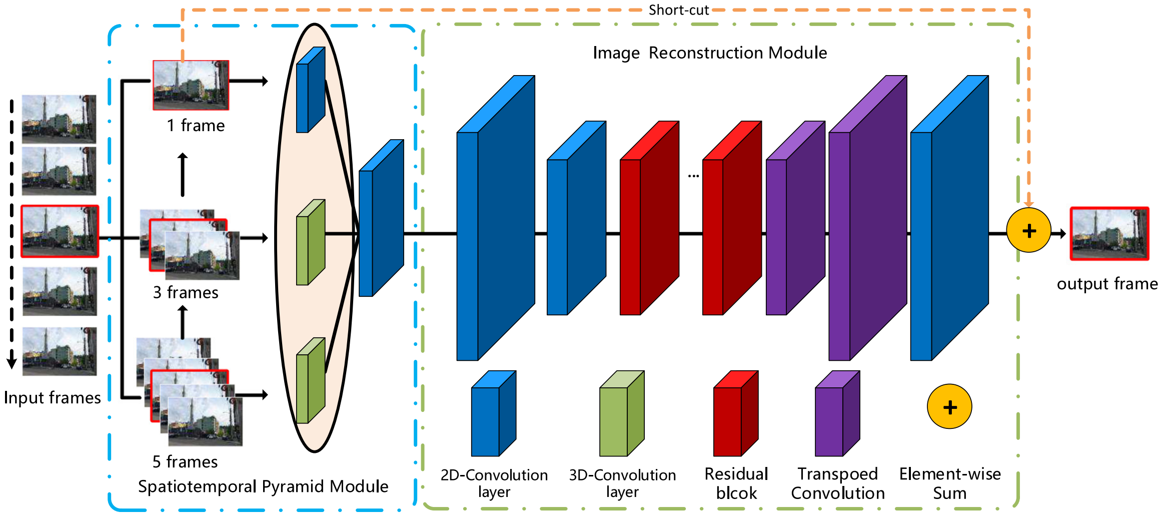


Fig. 2. The architecture of the proposed SPN. The input is five consecutive blurry frames. The output is the central deblurred image frame. The orange arrow indicates the short-cut from the input to the output, so that the network focuses on residual learning. Zoom in for better visibility.

non-blind case is relatively easier compared with the blind case. The blind case is an ill-posed problem as we have only Y in hand while neither the kernel K nor X is known. In most cases, the deblurring problem is blind as the blur process is unknown in advance.

Different methods are proposed to handle the deblurring problem. For example, the prior of sparsity is employed for the task of deblurring in Krishnan et al. (2011), Babacan et al. (2012) and Zhang et al. (2013). Yan et al. (2017) propose an extreme channel prior which is the joint of a dark channel prior and a bright channel prior for deblurring. Data fitting term is also considered for image deblurring. Pan et al. (2017b) learn a data fitting function driven by data to estimate blur kernels of generic scenes for blind image deblurring. An algorithm is proposed in Dong et al. (2017) to handle outliers in the modeling of data fidelity, eliminating the side effect of outliers in the estimation of blur kernels. A self-paced kernel estimation approach is proposed by Gong et al. (2017). In the proposed approach, inlier pixels are gradually detected and incorporated into the process of kernel estimation. This self-paced approach improves the robustness of the estimation. Later, multiple images are considered in image deblurring. Because they provide more information. For instance, a pair of blurred and noisy images are employed in Yuan et al. (2007). Petschnigg et al. (2004) use flash and no-flash image pairs to restore a sharp image. An approach composed of two main components is presented by Bahat et al. (2017). The above methods assume blurred images are noise-free and perform unsatisfactorily on blurry images with noise. To handle noise in image deblurring, Anger et al. (2019) propose an adaptation of the L0-based kernel estimation method, which uses an improved blur kernel estimation method to deal with low noise, and a non-blind deconvolution method to deal with medium and high noise.

In recent years, the deep convolutional neural network (CNN) has witnessed advances in various kinds of vision problems. It has been applied to image deblurring (Zhang et al., 2020b; Sun et al., 2015). A pioneer work is Sun et al. (2015), which estimates the probability of blur kernel using CNN at the patch level. Nah et al. (2017b) propose a multi-scale convolutional neural network with a multi-scale loss function to progressively restore sharp images in multiple scales in an end-to-end manner. Kupyn et al. (2018) present a conditional GAN which produces high-quality deblurred images via the Wasserstein loss. In the literature Ren et al. (2018), a low-rank property is used to compute a set of blur kernels, which are further used to initialize a network. The information contained in the blur kernels enables the trained network to handle various kinds of blur artifacts. Tao et al. (2018) explore what the proper network structure is for using the coarse-to-fine scheme in image deblurring, and propose a scale-recurrent network for image deblurring. Zhang et al. (2018) use an RNN, within which the pixel-wise weights of the RNN are learned from a CNN, to model the spatially varying blur. Zhang et al. (2020a) propose two GAN (BGAN and DBGAN) models for image deblurring. BGAN learns how to blur sharp images with unpaired sharp and blurry image sets, and DBGAN learns to correctly deblur such images. Inspired by spatial pyramid matching, Zhang et al. (2019a) present a Deep Multi-Path Hierarchical Network (DMPHN) to deblur blurry images by a fine-to-coarse hierarchical representation. Zhang et al. (2019a) propose two neural networks (*i.e.*, DeblurRNN and DeblurMerger), to restore clear images, which use a pair of noise/blurred images captured in a burst for end-to-end restoration in a sequential or parallel manner.

2.2. Video deblurring

For video deblurring, temporal information provides additional clues for deblurring. Early video deblurring methods recover sharp details of the current frame by nearby frames via several techniques, such as path matching, motion flow and frame alignment. However, they cannot deal with large movements. The algorithm by Delbracio and Sapiro (2015b) first uses optical flow to warp nearby frames, then fuses them in the Fourier domain to remove blurs in a video. Pan et al. (2017a) propose a single framework to jointly deblur scene videos and estimate the scene flow. The deblurring performance can benefit from the scene flow and blur information. Ren et al. (2017) tackle the video deblurring problem by exploiting the semantic segmentation in each blurry frame and use different models for optical flow estimation in image regions. Hyun Kim and Mu Lee (2015) and Kim et al. (2017) propose a segmentation-free dynamic video deblurring method, which approximates locally varying blur kernels via bidirectional optical flows.

With the emergence of the realistic blur datasets (Nah et al., 2017b; Su et al., 2017), several deep-learning based methods have been proposed for video deblurring. For example, Su et al. (2017) propose a network called DBN. It takes five consecutive blurry frames as an input and deblurs a central frame among them. A spatio-temporal recurrent network is proposed in Hyun Kim et al. (2017). It adaptively enforces temporal consistency among successive frames. Chen et al. (2018) propose an optical flow based reblurring step to reconstruct the blurry input, which is employed to fine-tune deblurring Network via self-supervised learning. To overcome the limitation of optical flow estimation, Zhou et al. (2019) propose a Spatio-Temporal Filter Adaptive Network (STFAN) that dynamically generates element-wise alignment and deblurring filters for video deblurring.

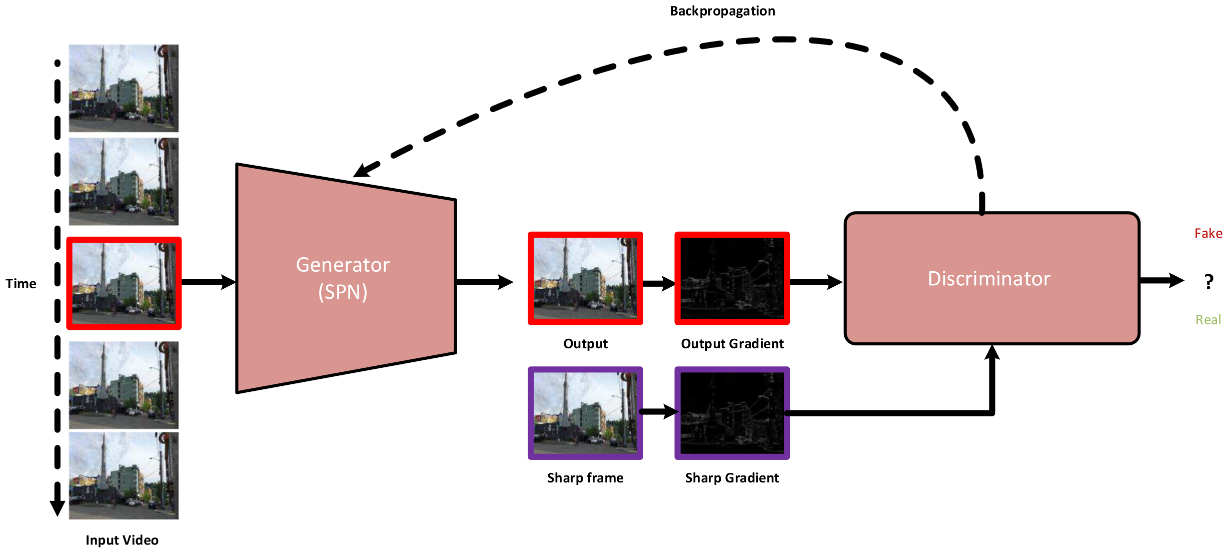


Fig. 3. The SPGAN framework for video deblurring. The generator is SPN, while the discriminator is a VGG-like CNN network with the gradient prior.

3. The proposed models

In this section, we first introduce the proposed network (SPN). Then we introduce SPGAN with the adversarial gradient prior, which is based on both SPN and adversarial gradient prior. Finally, the loss functions are presented.

3.1. Spatiotemporal pyramid network

To provide a clear view, the overall architecture of SPN is presented in Fig. 2. Taking five consecutive blurry frames, SPN aims to recover the center sharp frame, *i.e.,* the third frame. SPN consists of a spatiotemporal pyramid module and an image reconstruction module. To simplify the training, all input frames are transformed into $YCbCr$ space. Only the Y channel of each frame is used like the work in Zhang et al. (2019b). Then these inputs are fed into the spatiotemporal pyramid module. The spatiotemporal pyramid module first processes these inputs to get features in different spatiotemporal scales by using three different paths, and then fuses these features with abundant information in different scales. Finally, the output features of the spatiotemporal pyramid module are processed and reconstructed by the image reconstruction module to produce the output. Besides, a short-cut is employed to maintain the color information of the original center frame. The final center sharp frame is obtained by transforming the output back to the colored image with the original Cb and Cr channels. Next, we detail the architecture of SPN.

Spatiotemporal Pyramid Module. Generally, recent video deblurring methods work on a fixed spatiotemporal scale. For example, a fixed number of blurry frames are stacked as inputs to the model in Su et al. (2017) and Hyun Kim et al. (2017). However, different types of motions lead to different spatiotemporal scales of blur in a video. Therefore, we propose a spatiotemporal pyramid module to simultaneously model spatial and temporal dynamics among successive frames in real videos, as shown in Fig. 2. The spatiotemporal pyramid module consists of a pyramid block and a fusion block. The pyramid block is constructed by a 2D convolutional layer and two 3D convolutional layers. The 3D convolution is related to spatiotemporal representation learning (Akilan et al., 2018). It conducts convolution in both spatial and temporal domains, whereas 2D convolution carries out convolution in the spatial domain. The 3D convolution can be formulated as:

$$V_{ij}^{xyz} = \sigma(\sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} V_{(i-1)m}^{(x+p)(y+q)(z+r)} \cdot g_{ijm}^{pqr} + b_{ij}), \qquad (1)$$

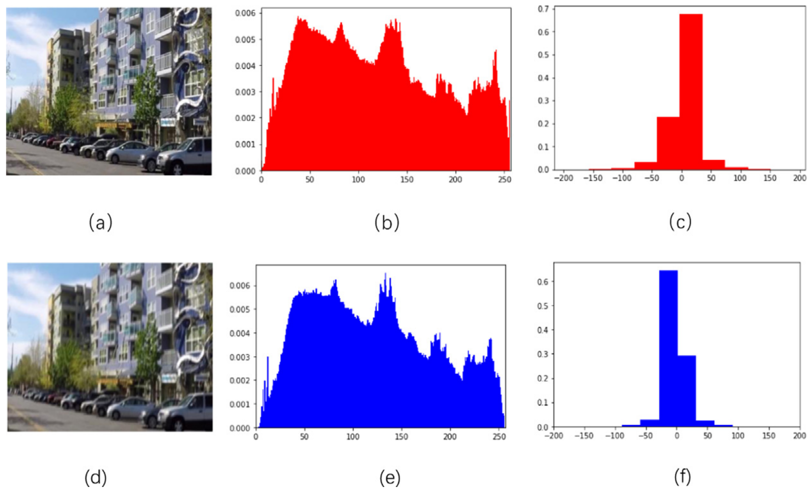
where V_{ij}^{xyz} refers to the value in the jth feature map of the ith layer, g_{ijm}^{pqr} represents the value at (p, q, r) of the 3D kernel connected to the mth feature map from the $(i-1)th$ layer, (P_i, Q_i, R_i) is the size of 3D convolution kernel, and the σ represents an activation function such as $ReLU$, $Tanh$ or $Sigmoid$. The fusion block consists of a concatenation operation and a convolutional layer.

The input of the spatiotemporal pyramid module is five consecutive blurry video frames. These five frames are first divided into three parts by the pyramid operation. The three parts are a center frame, three consecutive frames, and five consecutive frames, which contain spatiotemporal information with different scales. The pyramid block captures spatiotemporal information of different scales by three paths using 2D and 3D convolutions together. Each path in this block increases the width (the number of channels) of the input feature to 16, respectively. Due to its designation, this block can also capture features locally (a center frame) and globally (multiple frames). The output features of these three convolution operations are concatenated together and processed by a convolutional layer in the fusion block (*i.e.,* a 2D convolution) to further enhance the feature representations. The width of the output feature from the fusion block is 64. The spatiotemporal pyramid module helps the model simultaneously consider sharp temporal dependency and different levels of spatially blur in a center frame at the same time. In addition, the global and local information of all input frames is utilized. More analysis of this module is presented in Section 4.

Image Reconstruction Module. As illustrated in Fig. 2, this module consists of three blocks. The first block contains two consecutive convolutional blocks and each of them has one convolution layer, a batch normalization layer and a $ReLU$ activation layer. Each convolutional layer downsamples the feature maps by half and increases the width of feature maps by 2 times (*i.e.,* 64 → 128 → 256). The output feature maps of the first block are fed into the second block, which contains six residual blocks with $ReLU$ as the activation functions. Finally, the output features of the second block are reconstructed by the third block. It consists of two transposed layers and one convolutional layer with $Tanh$ as the activation function. The feature maps after each transposed layer are upsampled by a factor of two and the width of feature maps is reduced by half (*i.e.,* 256 → 128 → 64). The convolutional layer reconstructs the features into an image frame.

3.2. Spatiotemporal pyramid generative adversarial network

GAN is composed of a generator network G and a discriminator network D. It represents a class of generative models based on a game



(a) (b) (c)

(d) (e) (f)

Fig. 4. Statistics of images. (a) sharp image. (b) pixel intensities of (a). (c) horizontal gradient histogram of (a). (d) blurred image. (e) pixel intensities of (d). (f) horizontal gradient histogram of (d).

theory in which a generator network competes against an adversary. For the deblurring task, G is trained to restore a deblurred image to try and fool the discriminator, and D is to distinguish between sharp and blurry images.

Typically, the generated results are discriminated from the real-world samples by the discriminator for tasks of image/video restoration (Zhang et al., 2019b; Kupyn et al., 2018). The discrimination is usually conducted in the image space. However, this space cannot be fully covered by the limited number of training examples. We argue that the discrimination is better achieved using the image gradients rather than the image intensity values. It is based on the observation that the blurred image and the sharp image have different uniform intensity values (*i.e.* the gradient prior). Figs. 4(b) and 4(e) illustrate the pixel intensities of a clear image and a blurred image. It can be seen that the values of neighbor pixels in the blurred image are closer than those in a sharp image. Figs. 4(c) and 4(f) show the horizontal gradient histograms of the sharp image and the blurred one. The values of blurred image gradients change faster than those of sharp image gradients. The distribution of gradient images of sharp images is different from that corresponding to blurry images. Therefore, we propose to use the prior knowledge and improve discriminator with the gradient image. Specifically, we combine the gradient of an image I, which can be formulated as follows:

$$\nabla I = \sqrt{\nabla_v I (\nabla_v I)^T} = \sqrt{(\frac{\partial I}{\partial x})^2 + (\frac{\partial I}{\partial y})^2}, \tag{2}$$

where $\nabla_v I$ is the gradient of an image I, $\frac{\partial I}{\partial x}$ is the derivative with respect to x (gradient in the x direction) and $\frac{\partial I}{\partial y}$ is the derivative with respect to y (gradient in the y direction). Considering images in the gradient space, the discriminator will be more effective in discriminating between real-world sharp frames and the generated sharp frames.

For building our GAN module, the trained SPN is used as a generator, and the network which has the same structure as the VGG network (Simonyan and Zisserman, 2014) is applied as a discriminator. The discriminator contains 14 convolutional layers, and the number of channels changes from 64 to 512. At last, it uses a soft-max function at the last layer of the network to obtain the probability that the generated image is real. Table 1 shows the whole architecture of the discriminator. The whole framework of the proposed SPGAN is shown in Fig. 3.

3.3. Loss function

To optimize the proposed SPN and SPGAN, we utilize two different loss functions, namely content loss and adversarial loss.

Table 1
The structure of the discriminator in SPGAN. BN means batch normalization and ReLU refers to the activation function.

Layers	1–2	3–5	6–9	10–14	15–16	17
kernel	3 × 3	3 × 3	3 × 3	3 × 3	FC	FC
channels	64	128	256	512	4096	2
BN	BN	BN	BN	BN	–	–
ReLU	ReLU	ReLU	ReLU	ReLU	–	–

Content Loss. In the deblurring task, the mean squared error (MSE) is the commonly used content loss function. On the basis of MSE, the content loss is:

$$\mathcal{L}_{cont} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{x,y}^{sharp} - G(I^{blurry})_{x,y})^2, \tag{3}$$

where W and H are the width and height of an image frame, $I_{x,y}^{sharp}$ is the value of sharp image frame at location (x, y), and $G(I^{blurry})_{x,y}$ corresponds to the value of the deblurred image frame which is generated from the proposed network.

Adversarial Loss. We apply the adversarial loss of GAN. The model G is initialized to recover the video frames, while the model D is to distinguish samples in gradient images. The model D drives model G to recover more realistic sharp image frames. The gradient adversarial loss is formulated as:

$$\mathcal{L}_{adv} = \log(1 - D(\nabla G(I^{blurry}))), \tag{4}$$

where $G(I^{blurry})$ is a deblurred image frame by G, $\nabla G(I^{blurry})$ is the gradient of $G(I^{blurry})$, and $D(\nabla G(I^{blurry}))$ is the probability that the gradient image is real.

Total Loss. In order to optimize SPGAN, we combine two loss functions above to form the total loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{cont} + \lambda \cdot \mathcal{L}_{adv}\,. \tag{5}$$

The content loss $\mathcal{L}_{cont}$, and the gradient adversarial loss $\mathcal{L}_{adv}$ are fused in a weighted fashion. The adversarial loss together with the content loss is used to make the GAN module learn the spatial and temporal information and recover the original frame details. To balance these two different losses, the weight λ is set as a value between 0 and 1. In the experiments, we train different models by controlling the value of λ. When $\lambda = 0$, the total loss degenerates into content loss, and the model refers to SPN.

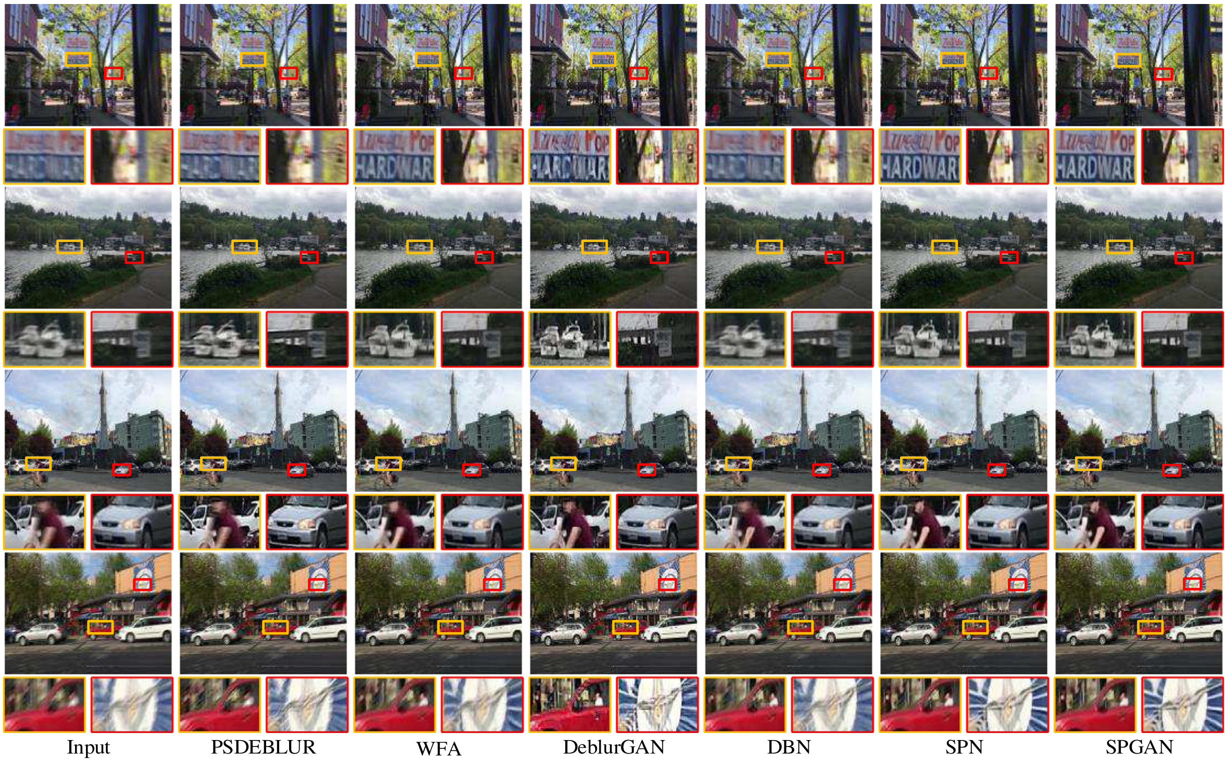


Fig. 5. Comparison with the state-of-the-art deblurring methods on the Deep Video Deblurring dataset (Su et al., 2017) (the quantitative subset). From left to right: Input, PSDEBLUR, WFA (Delbracio and Sapiro, 2015a), DeblurGAN (Kupyn et al., 2018), DBN (Su et al., 2017), SPN, and SPGAN. Zoom in for better visibility.

4. Experimental results

We evaluate the proposed method on standard benchmark datasets to demonstrate the effectiveness of the proposed methods for video deblurring. The dataset introduction and evaluation metrics are given in Section 4.1. Experiment details are reported in Section 4.2. In order to prove the validity of the proposed model, we carry out an ablation study to demonstrate the effectiveness of the spatiotemporal pyramid module and adversarial gradient prior, which is presented in Section 4.3. Furthermore, we compare the proposed model with several state-of-the-art methods, including MSCNN (Nah et al., 2017a), PSDEBLUR, WFA (Delbracio and Sapiro, 2015a), DBN (Su et al., 2017), DeblurGAN (Kupyn et al., 2018), STFAN (Zhou et al., 2019), and DMPHN (Zhang et al., 2019a) in Section 4.4.

4.1. Datasets and metrics

(1) Datasets: Deep Video Deblurring (DVD) dataset (Su et al., 2017) is the most popular benchmark for video deblurring algorithms. DVD dataset consists of two subsets: the qualitative subset and the quantitative subset. Both the qualitative and the quantitative subsets are captured by various devices such as Canon 7D, GoPro Hero 4 Black and iPhone 6s. Each video in the two subsets includes approximately 100 frames of size 1280 × 720 (Su et al., 2017). The qualitative subset has 22 different scenes. However, it does not provide the corresponding ground truth data. For the quantitative subset, there are 6708 synthetic blurry frames with corresponding ground truth from 71 different videos. So we only use the quantitative subset for training our models. We divide the dataset into a training set and a testing set following the previous method (Su et al., 2017), and compare the proposed model with the state-of-the-art methods on the testing set in terms of both the objective measurement and the visual quality. Besides, we conduct tests on the qualitative subset to further illustrate the effects of the proposed models.

(2) Metrics: To evaluate the performance of the proposed methods, two metrics are used in this paper: the Peak Signal to Noise Ratio (RSNR) and the Structural Similarity index (SSIM). At the same time, we also use visual quality to further evaluate the video deblurring performance.

4.2. Implementation details

When training the proposed network, the weights of the network are initialized from a Gaussian distribution $\mathcal{N}(0, 0.01)$. We adopt an ADAM optimizer with a batch of 4, and the exponential decay rate is set as $(\beta_1, \beta_2) = (0.5, 0.999)$. The input size is 128 × 128, which is cropped randomly from any location of the 1280 × 720 image. In this way, there are at least 712,193 possible samples per frame on DVD dataset (Su et al., 2017), greatly increasing the number of training samples in the training stage. We update the weights of SPN in each mini-batch. The learning rate is initialized to 0.0001 for all layers. When the training loss does not change significantly, we reduce the learning rate of the network to its one-tenth and continue the training to further improve the performance.

In the training of SPGAN, the hyper-parameter λ in the loss function is empirically set as 0.00001 to achieve the best performance. There are several benefits by doing so. First, SPGAN is not trained from scratch, we use the trained SPN directly as a generator and fine-tune SPGAN. So when training SPGAN, at first it produces PSNR as high as the SPN. Second, the adversarial loss with the content loss is used to make SPGAN learn the global information and recover the original frame more efficiently. The hyper-parameter λ is set as 0.00001, which is relatively small. In the end, we keep the learning rate at 0.00001, so the result of PSNR does not change dramatically. When the PNSR starts to decrease, we finish training for SPGAN.

4.3. Ablation study

To better validate the effectiveness of different components of the proposed models, we perform an ablation study by individually considering different factors. Specifically, three factors are considered,

Table 2
Performance comparisons in terms of PSNR with MSCNN (Nah et al., 2017a), PSDEBLUR, WFA (Delbracio and Sapiro, 2015a), DBN (SINGLE), DBN (NOALIGN), DBN (FLOW) (Su et al., 2017), DeblurGAN (Kupyn et al., 2018), STFAN (Zhou et al., 2019), DMPHN (Zhang et al., 2019a), SPN and SPGAN on the DVD dataset (Su et al., 2017). All the results are quoted from the corresponding publications. '–' means not reported by the original publications.

Method	1	2	3	4	5	6	7	8	9	10	Average (PSNR)
INPUT	24.14	30.52	28.38	27.31	22.60	29.31	27.74	23.86	30.59	26.98	27.14
MSCNN	26.84	31.56	29.29	29.46	24.19	29.94	28.50	25.18	32.07	27.89	28.49
PSDEBLUR	24.42	28.77	25.15	27.77	22.02	25.74	26.11	19.71	26.48	24.62	25.08
WFA	25.89	32.33	28.97	28.36	23.99	31.09	28.58	24.78	31.30	28.20	28.35
DBN (single)	25.75	31.15	29.30	28.38	23.63	30.70	29.23	25.62	31.92	28.06	28.37
DBN (noalign)	27.83	33.11	31.29	29.73	25.12	32.52	30.80	27.28	33.32	29.51	30.05
DBN (flow)	28.31	33.14	30.92	29.99	25.58	32.39	30.56	27.15	32.95	29.53	30.05
DeblurGAN	–	–	–	–	–	–	–	–	–	–	30.16
STFAN	28.83	34.46	32.56	31.60	26.13	33.32	31.00	27.86	35.36	30.30	31.24
DMPHN	30.48	34.41	32.25	32.10	26.74	33.12	30.86	27.55	35.25	30.60	31.43
SPN	29.53	36.14	32.51	32.58	26.48	34.18	32.24	29.46	37.54	31.27	32.19
SPGAN	29.66	36.21	32.62	32.68	26.53	34.32	32.37	29.58	37.69	31.37	32.30

Table 3
Performance comparison between different variants with different components on the DVD dataset (Su et al., 2017).

Models	PSNR	SSIM
SPN(1)	29.98	0.897
SPN(5)	31.43	0.920
SPN(1-3)	31.72	0.920
SPN(3-5)	31.64	0.924
SPN(1-3-5-7)	31.97	0.929
SPN	32.19	0.930
SPGAN-plain	32.13	0.929
SPGAN	32.30	0.940

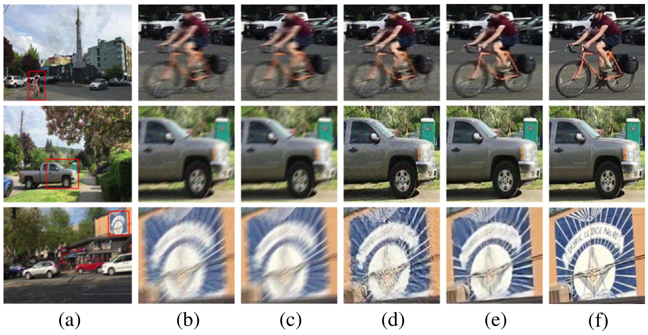


Fig. 6. Deblurring results of frames from the DVD dataset (Su et al., 2017). (a) Blurring frames with full resolution. (b) Input noisy frames. (c) Input blurry frames. (d) Results of the proposed SPGAN with noisy inputs. (e) Results of the proposed SPGAN with blurry inputs. (f) Ground truth frames.

including different spatiotemporal scales, GAN, and adversarial gradient prior. We use different combinations of components to construct the following variant models: (1) SPN(1): a model constructed by the spatiotemporal pyramid module with only one path of one frame; (2) SPN(5): a model constructed by the spatiotemporal pyramid module with only one path of five frames; (3) SPN(1-3): the spatiotemporal pyramid module with two paths of one frame and three frames is put in front of the feature extraction and image reconstruction module (the base model); (4) SPN (3-5): Only spatiotemporal pyramid module with two paths of three frames and five frames is constructed in front of the base model; (5) SPN: a model constructed with spatiotemporal pyramid module with three paths of one frame, three frames and five frames; (6) SPN(1-3-5-7): a variant that constructed with spatiotemporal pyramid module with four paths of one frame, three frames, five frames, and seven frames; (7) SPGAN-plain: SPN with a discriminator that considers original image visual space; (8) SPGAN: SPN with a discriminator that considers image gradient space.

All these variants are trained in the same way as before and tested on the same testing dataset from the DVD dataset. The comparison results are shown in Table 3. It demonstrates that the proposed SPN and SPGAN achieve better performance of video deblurring in terms of both PSNR and SSIM. By changing the input path of the spatiotemporal pyramid module, it is clear that considering both spatial information of center frame and spatiotemporal information from successive frames leads to higher average PSNR and SSIM values (SPN(1), SPN(5), SPN(1-3), SPN(3-5)). Meanwhile, the results (SPN(1), SPN(1-3), SPN, SPN(1-3-5-7)) show that increasing spatiotemporal scales of input frames will not result in better performance, which suggests that choosing the right spatiotemporal scales of input frames is important for information gathering to conduct deblur. Moreover, SPGAN outperforms SPN in terms of both PSNR and SSIM, while the SPGAN-plain does not. It demonstrates that the employed gradient prior helps the discriminator to learn more difference between blurry frames and sharp frames and improve the deblurring performance of SPN.

The ablation study shows that the spatiotemporal pyramid module, GAN and adversarial gradient prior have their own contributions to the performance of the full model, which justifies the overall design.

In addition, we illustrate the sensitivity to noise of the proposed SPGAN. Fig. 6 shows exemplar images of the deblurred results on the DVD dataset. The visual results demonstrate that our method is able to remove blur in the case of blurry and noisy images, and can produce visually pleasing results. The results show that the proposed SPGAN is robust to noise on blurry frames.

4.4. Comparison with state of the art

To further demonstrate the effectiveness of our approach, we compare our proposed models with various state-of-the-art methods for video deblurring. The quantitative comparison results are shown in Table 2, in which the reported values are the average of the result values on the 10 testing videos from the DVD dataset in terms of PSNR. In Table 2, PSDEBLUR refers to the results of deblurring by Photoshop, and WFA acquires the deblurred frames by using multiple frames as input. DeblurGAN, DBN, DMPHN, and STFAN are recent video deblurring methods which achieve state-of-the-art performance. As shown in Table 2, our SPN already outperforms the state-of-the-art methods. It surpasses DMPHM by 0.76 dB, and more than 3.84 dB over WFA. Further equipped with the adversarial gradient prior, our SPGAN yields PSNR results with a 0.11 dB gain over the SPN model.

We also conduct a qualitative comparison between the proposed method and several state-of-the-art methods. Figs. 5 and 7 show exemplar result frames of different approaches in both the quantitative

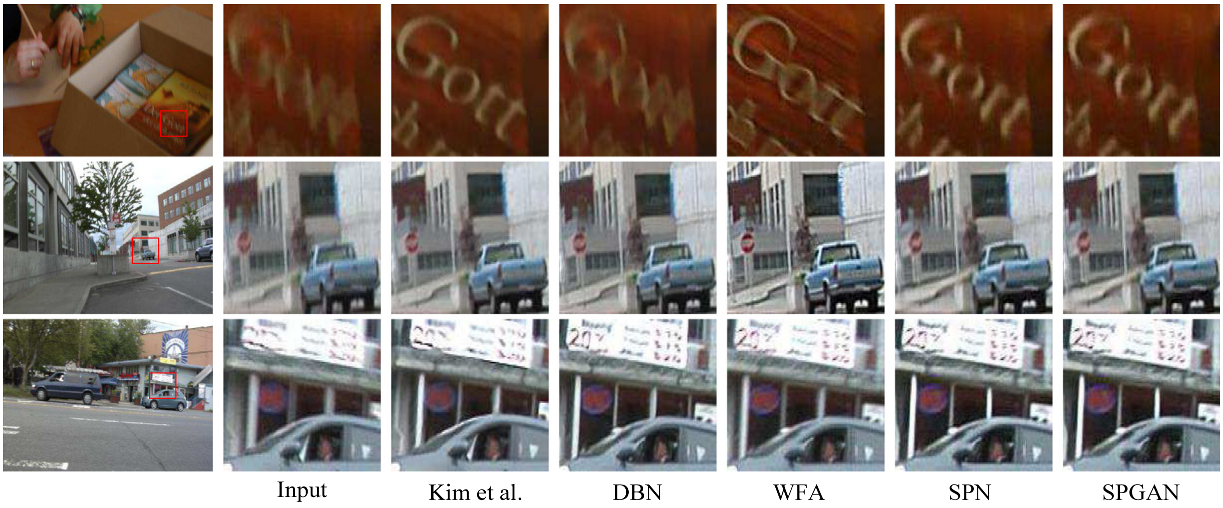


Fig. 7. Comparison with the state-of-the-art deblurring methods on the Deep Video Deblurring dataset (Su et al., 2017) (the qualitative subset). From left to right: Input, Hyun Kim and Mu Lee (2015), DBN (Su et al., 2017), WFA (Delbracio and Sapiro, 2015a), SPN, and SPGAN. Zoom in for better visibility.

and qualitative subsets from the DVD dataset, respectively. In Fig. 5, we illustrate the deblurring performance of different models. When zooming in the main object for clarity, it is observed that results of PSDEBLUR, DEBLURGAN and DBN remain blurry to some extent in part of the generated frames. In contrast, SPN and SPGAN show far fewer artifacts while preserving the sharp structural information. In Fig. 7, we select the images from three different scenes to demonstrate the advantages of SPN and SPGAN. As can be seen, our SPGAN restores the sharpest details in all cases. The comparison results show that the proposed SPN and SPGAN can robustly handle complex blur in the real-world situation, which further demonstrates the superiority of the proposed methods.

5. Conclusion

In this work, we propose SPN and SPGAN for video deblurring. The spatiotemporal pyramid module is used to learn the different scales of spatiotemporal information to conduct deblurring for a video, while the image reconstruction module reconstructs the features to produce the sharp image frame. Different from the previous GAN that the discriminator works on the visual image space for video deblurring, we propose to use the adversarial gradient prior in the GAN model, which is helpful to the discrimination of discriminator. Extensive experimental results on the benchmark datasets show that the proposed methods achieve state-of-the-art performance.

CRediT authorship contribution statement

Tao Wang: Methodology, Investigation, Software, Writing - original draft. **Xiaoqin Zhang:** Funding acquisition, Project administration, Supervision, Conceptualization. **Runhua Jiang:** Writing - review & editing, Software. **Li Zhao:** Software, Resources, Visualization, Validation. **Huiling Chen:** Software, Resources, Visualization, Validation, Formal analysis. **Wenhan Luo:** Conceptualization, Data curation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China [grant no. 61922064], in part by the Zhejiang Provincial Natural Science Foundation [grant nos. LR17F030001, LQ19F020005], in part by the Project of science and technology plans of Wenzhou City [grant nos. C20170008, G20150017, ZG2017016].

References

Akilan, T., Wu, Q.J., Jiang, W., Safaei, A., Huo, J., 2018. New trend in video foreground detection using deep learning. In: Proceedings of Midwest Symposium on Circuits and Systems. pp. 889–892.

Anger, J., Delbracio, M., Facciolo, G., 2019. Efficient blind deblurring under high noise levels. In: Proceedings of International Symposium on Image and Signal Processing and Analysis. IEEE, pp. 123–128.

Babacan, S.D., Molina, R., Do, M.N., Katsaggelos, A.K., 2012. Bayesian blind deconvolution with general sparse image priors. In: Proceedings of the European Conference on Computer Vision. pp. 341–355.

Bahat, Y., Efrat, N., Irani, M., 2017. Non-uniform blind deblurring by reblurring. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3286–3294.

Chen, L., Fang, F., Wang, T., Zhang, G., 2019. Blind image deblurring with local maximum gradient prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1742–1750.

Chen, H., Gu, J., Gallo, O., Liu, M.-Y., Veeraraghavan, A., Kautz, J., 2018. Reblur2deblur: Deblurring videos via self-supervised learning. In: Proceedings of International Conference on Computational Photography. IEEE, pp. 1–9.

Cho, S., Wang, J., Lee, S., 2011. Handling outliers in non-blind image deconvolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 495–502.

Delbracio, M., Sapiro, G., 2015a. Burst deblurring: Removing camera shake through fourier burst accumulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2385–2393.

Delbracio, M., Sapiro, G., 2015b. Hand-held video deblurring via efficient fourier aggregation. IEEE Trans. Comput. Imaging 1, 270–283.

Dong, J., Pan, J., Su, Z., Yang, M., 2017. Blind image deblurring with outlier handling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2478–2486.

Gong, D., Tan, M., Zhang, Y., van den Hengel, A., Shi, Q., 2017. Self-paced kernel estimation for robust blind image deblurring. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1661–1670.

Hyun Kim, T., Mu Lee, K., 2015. Generalized video deblurring for dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5426–5434.

Hyun Kim, T., Mu Lee, K., Scholkopf, B., Hirsch, M., 2017. Online video deblurring via dynamic temporal blending network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4038–4047.

Javaran, T.A., Hassanpour, H., Abolghasemi, V., 2017. Non-blind image deconvolution using a regularization based on re-blurring process. Comput. Vis. Image Underst. 154, 16–34.

Joshi, N., Zitnick, C.L., Szeliski, R., Kriegman, D.J., 2009. Image deblurring and denoising using color priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1550–1557.

Kim, T.H., Nah, S., Lee, K.M., 2017. Dynamic video deblurring using a locally adaptive blur model. IEEE Trans. Pattern Anal. Mach. Intell. 40, 2374–2387.

Krishnan, D., Tay, T., Fergus, R., 2011. Blind deconvolution using a normalized sparsity measure. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 233–240.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J., 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8183–8192.

Lee, H.S., Kwon, J., Lee, K.M., 2011. Simultaneous localization, mapping and deblurring. In: 2011 International Conference on Computer Vision. IEEE, pp. 1203–1210.

Levin, A., Weiss, Y., Durand, F., Freeman, W.T., 2011. Efficient marginal likelihood optimization in blind deconvolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2657–2664.

Nah, S., Kim, T.H., Lee, K.M., 2017a. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 257–265.

Nah, S., Kim, T.H., Lee, M.K., 2017b. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3883–3891.

Pan, L., Dai, Y., Liu, M., Porikli, F., 2017a. Simultaneous stereo video deblurring and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4382–4391.

Pan, J., Dong, J., Tai, Y., Su, Z., Yang, M., 2017b. Learning discriminative data fitting functions for blind image deblurring. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1068–1076.

Pan, J., Hu, Z., Su, Z., Lee, H.-Y., Yang, M.-H., 2016. Soft-segmentation guided object motion deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 459–468.

Pan, J., Hu, Z., Su, Z., Yang, M.-H., 2014. Deblurring text images via L0-regularized intensity and gradient prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2908.

Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K., 2004. Digital photography with flash and no-flash image pairs. ACM Trans. Graph. 23, 664–672.

Ren, W., Pan, J., Cao, X., Yang, M., 2017. Video deblurring via semantic segmentation and pixel-wise non-linear kernel. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1077–1085.

Ren, W., Zhang, J., Ma, L., Pan, J., Cao, X., Zuo, W., Liu, W., Yang, M., 2018. Deep non-blind deconvolution via generalized low-rank approximation. In: Advances in Neural Information Processing Systems. pp. 297–307.

Schmidt, U., Rother, C., Nowozin, S., Jancsary, J., Roth, S., 2013. Discriminative non-blind deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 604–611.

Schuler, C.J., Christopher Burger, H., Harmeling, S., Scholkopf, B., 2013. A machine learning approach for non-blind image deconvolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1067–1074.

Shan, Q., Jia, J., Agarwala, A., 2008. High-quality motion deblurring from a single image. ACM Trans. Graph. 27, 73.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O., 2017. Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1279–1288.

Sun, J., Cao, W., Xu, Z., Ponce, J., 2015. Learning a convolutional neural network for non-uniform motion blur removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 769–777.

Tao, X., Gao, H., Shen, X., Wang, J., Jia, J., 2018. Scale-recurrent network for deep image deblurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 8174–8182.

Wang, H., Pan, J., Su, Z., Liang, S., 2018. Blind image deblurring using elastic-net based rank prior. Comput. Vis. Image Underst. 168, 157–171.

Xu, L., Ren, J.S., Liu, C., Jia, J., 2014. Deep convolutional neural network for image deconvolution. In: Advances in Neural Information Processing Systems. pp. 1790–1798.

Yan, Y., Ren, W., Guo, Y., Wang, R., Cao, X., 2017. Image deblurring via extreme channels prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4003–4011.

Yuan, L., Sun, J., Quan, L., Shum, H., 2007. Image deblurring with blurred/noisy image pairs. ACM Trans. Graph. 26, 1.

Yuan, L., Sun, J., Quan, L., Shum, H., 2008. Progressive inter-scale and intra-scale non-blind image deconvolution. ACM Trans. Graph. 27, 74.

Zhang, H., Dai, Y., Li, H., Koniusz, P., 2019a. Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 5978–5986.

Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., Li, H., 2019b. Adversarial spatio-temporal learning for video deblurring. IEEE Trans. Image Process. 28, 291–301.

Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., Li, H., 2020a. Deblurring by realistic blurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

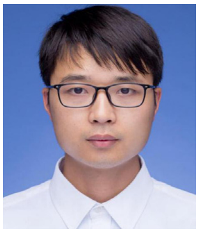
Zhang, J., Pan, J., Ren, J., Song, Y., Bao, L., Lau, R.W., Yang, M.-H., 2018. Dynamic scene deblurring using spatially variant recurrent neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 2521–2529.

Zhang, X., Wang, T., Wang, J., Tang, G., Zhao, L., 2020b. Pyramid channel-based feature attention network for image dehazing. Comput. Vis. Image Underst. 197–198, 103003.

Zhang, X., Wang, D., Zhou, Z., Ma, Y., 2020c. Robust low-rank tensor recovery with rectification and alignment. IEEE Trans. Pattern Anal. Mach. Intell..

Zhang, H., Wipf, D., Zhang, Y., 2013. Multi-image blind deblurring using a coupled adaptive sparse prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1051–1058.

Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J., 2019. Spatio-temporal filter adaptive network for video deblurring. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2482–2491.



Tao Wang is currently a graduate student at College of Computer Science and Artificial Intelligence, Wenzhou University, China. He received the B.Sc. degree in information and computing science from Hainan Normal University, China, in 2018. His research interests include several topics in computer vision and machine learning, such as object tracking, image/video quality restoration, adversarial learning, image-to-image translation and reinforcement learning.

Xiaoqin Zhang received the B.Sc. degree in electronic information science and technology from Central South University, China, in 2005 and Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China, in 2010. He is currently a professor in Wenzhou University, China. His research interests are in pattern recognition, computer vision and machine learning. He has published more than 100 papers in international and national journals, and international conferences, including IEEE T-PAMI, IJCV, IEEE T-IP, IEEE T-IE, IEEE T-C, ICCV, CVPR, NIPS, IJCAI, AAAI, and among others.

Runhua Jiang is currently a graduate student majoring in computer software and theory at College of Computer Science and Artificial Intelligence, Wenzhou University, China. He received his B.Sc. degree in department of information science at Tianjin University of Finance and Economy, China. His research interests include image and video processing, pattern recognition and machine learning.

Li Zhao received the B.Sc. degree in automation in 2005 and M.Eng degree in control theory and control engineering in 2008 from Central South University, China. She is currently an assistant researcher in Wenzhou University. Her research interests are in pattern recognition, computer vision, and machine learning.

Huiling Chen is currently an associate professor in the college of computer science and artificial intelligence at Wenzhou University, China. He received his Ph.D. degree in department of computer science and technology at Jilin University, China. His present research interests center on evolutionary computation, machine learning and data mining, as well as their applications to medical diagnosis, bankruptcy prediction and parameter extraction of solar cell. He is currently serving as an associate editor of IEEE ACCESS and the editorial board member of Com-

putational and Mathematical Methods in Medicine. He is also a reviewer for many journals such as Applied Soft Computing, Artificial Intelligence in Medicine, Knowledge-based Systems, Future Generation Computer System. He has published more than 100 papers in international journals and conference proceedings, including Information Sciences, Pattern Recognition, Future Generation Computer System, Expert Systems with Applications, Knowledge-based Systems, Neurocomputing, PAKDD, and among others. He has more than 10 ESI highly cited papers and 2 hot cited papers.

1

Wenhan Luo received the Ph.D. degree from Imperial College London, UK, 2016, M.E. degree from Institute of Automation, Chinese Academy of Sciences, China, 2012 and B.E. degree from Huazhong University of Science and Technology, China, 2009. His research interests include several topics in computer vision and machine learning, such as motion analysis (especially object tracking), image/video quality restoration, object detection and recognition, reinforcement learning. 2