

Blind Motion Deblurring Super-Resolution: When Dynamic Spatio-Temporal Learning Meets Static Image Understanding

Wenjia Niu, Kaihao Zhang, Wenhan Luo, and Yiran Zhong

Abstract—Single-image super-resolution (SR) and multi-frame SR are two ways to super resolve low-resolution images. Single-Image SR generally handles each image independently, but ignores the temporal information implied in continuing frames. Multi-frame SR is able to model the temporal dependency via capturing motion information. However, it relies on neighbouring frames which are not always available in the real world. Meanwhile, slight camera shake easily causes heavy motion blur on long-distance-shot low-resolution images. To address these problems, a **Blind Motion Deblurring Super-Resolution Networks**, **BMDSRNet**, is proposed to learn dynamic spatio-temporal information from single static motion-blurred images. Motion-blurred images are the accumulation over time during the exposure of cameras, while the proposed BMDSRNet learns the reverse process and uses three-streams to learn Bidirectional spatio-temporal information based on well designed reconstruction loss functions to recover clean high-resolution images. Extensive experiments demonstrate that the proposed BMDSRNet outperforms recent state-of-the-art methods, and has the ability to simultaneously deal with image deblurring and SR.

Index Terms—Blind motion deblurring, single image super-resolution, multi-frame super-resolution, dynamic spatio-temporal learning.

1 INTRODUCTION

Super Resolution (SR) [1] has been an active topic for decades due to its utility in various applications. Its aims to improve the resolution of images given an input low-resolution image and output an image of high resolution. In most cases, low-resolution images also exhibit the artifact of blur. For example, capturing a fast-moving vehicle from a far distance produces an image of both low resolution and blur artifact. This paper focuses on super-resolving a low-resolution image with motion blur artifact.

Existing super resolution solutions approach the SR problem in both single-image [2] and multi-frame ways [3]. Solutions in the case of single image extract features in spatial domain only. These solutions can hardly work satisfactorily as they ignore the temporal information caused by the motion. Multi-frame super resolution is capable of using both spatial and temporal information contained in the multiple given frames, thus performs better than single-image super resolution. However, in our task, multiple neighbouring frames are not available.

As a fact, though we have only a single image, it does include rich temporal dynamics. The blurred single image

can be considered as an overlay of a sequence of images shot in multiple time steps during the exposure time window (assuming we have a camera of higher shutter speed than the original camera capturing the given image) [4]. Once we are able to obtain the assumed sequence of multiple frames, multi-frame super resolution can spontaneously be carried out for better performance.

Inspired by this, this paper tries to learn the dynamic spatio-temporal information from a static motion-blurred low-resolution image. We decouple the problem into two sub problems, motion deblurring and (multi-frame) super resolution. For the first one, we aim to extract multiple clear frames from the given single motion-blurred image, thus extract the spatio-temporal information and solve the motion deblurring sub problem. For the second one, with the produced multiple frames, we conduct multi-frame super resolution by utilizing both temporal and spatial cues contained in the multiple low-resolution frames and produce a high resolution image free of blur defect. By doing so, we borrow the temporal dynamics enclosed in the static motion-blurred image and solve both the problem of blind motion blur and super resolution.

Specifically, for solving the motion deblurring problem, we propose a Blind Motion Deblurred Net (BMDNet) which is composed of convolutional layers and residual blocks. The BMDNet is successful in recovering a sequence of clear images from a given single motion-blurred image, disentangling the fused multiple images corresponding to finer-scale moments. As such, it solves the static-to-dynamic problem. Following the BMDNet there are parallel three streams within the BMDSRNet. To better utilize the bidirectional temporal cues contained in the sequence of frames. The first stream, ForNet, and the second stream, BackNet, process the recovered sequence

• Wenjia Niu is with the School of electronic and information engineering, Hebei University of Technology, Tianjin, China. E-mail: {niuwenjia9064@hotmail.com}

• Kaihao Zhang and Yiran Zhong are with the College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia. E-mail: {kaihao.zhang@anu.edu.au; yiran.zhong@anu.edu.au}

• Wenhan Luo is with Tencent, Shenzhen, China. E-mail: {whluo.china@gmail.com}

from BMDNet in the forward and backward directions, respectively, with LSTM structure. The third stream is a CoreNet dedicated for the central frame. The three streams are learned with a well-designed reconstruction loss. The outputs of the three streams are fused by a FuNet to produce a high-resolution counterpart corresponding to the given low-resolution motion-blurred image. This thus tackles the LR-to-HR problem.

Our contributions are three-fold: 1) For the first time, we tackle the super resolution of single motion-blurred image as two sub-problems of motion deblurring and multiple frame super resolution. With this strategy, temporal cues can be utilized for the task of single image super resolution. 2) We propose an BMDSRNet to solve these two sub-problems. By learning bidirectional spatio-temporal dynamics with additional reconstruction loss, both these sub-problems are addressed. 3) Experimental results on public datasets demonstrate the superiority of the proposed method.

2 RELATED WORK

2.1 Image Super-resolution

Early solutions [5], [6] to image super resolution address this problem with sampling based interpolation techniques. Natural image prior [7], neighbour embedding and sparse coding are employed to better predict finer textures by the subsequent works [8], [9]. Recently, deep learning achieves significant success in low-level vision tasks [10], [11], [12], [13], which also include image super-resolution [2], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. In the deep learning, Dong *et al.* [14] for the first time propose to solve the image SR problem using deep convolutional neural networks and the method surpasses the traditional methods. The success of residual structure in the recognition task inspires Kim *et al.* to introduce the residual structure and thus train much deeper neural network for the image super-resolution task in [19]. A deeply-recursive convolutional network (DRCN) is proposed in [16]. A deep recursive layer is included in DRCN to improve the performance without new parameters. Huang *et al.* employ bi-directional recurrent convolutional network to tackle the problem of video super-resolution in [30]. Generative Adversarial Networks (GAN) is introduced for image super resolution in [31]. A GAN based network is firstly trained to learn how to downgrade image resolution with unpaired data. The paired output of this network is used to train the desired image SR network. This method verifies its effectiveness in real-world images. The popular attention mechanism is introduced in a very deep residual channel attention networks (RCAN) [32] to improve the representation ability of CNNs and ease the difficulty of training deep networks for image SR task.

2.2 Image Deblurring

Image deblurring has been addressed by early methods based on priors or constraints [33], [34], [35]. These kind of solutions operates on multiple scales which are also time consuming. Recently, many deep deblurring methods are proposed to address the problem of image deblurring [4], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47],

[48], [49], [50], [51], [52], [53], [54], [55] and video deblurring [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66]. Schuler *et al.* [67] propose a pioneer work of using deep learning for image deblurring. A two-stage architecture is developed and trained in an end-to-end manner for the blind image deblurring. Sun *et al.* [36] use CNN to estimate blur kernel and deblur images based on the estimated kernel. Targeted on non-blind deblurring, Xu *et al.* [68] develop connection between deep neural networks and traditional optimization based approaches, propose a structure of two sub-modules, and achieve better performance. Similar to the traditional methods before deep learning era [69], [70], [71], the multi-scale strategy is also employed by Nah *et al.* [38]. The sharp image is directly generated by a network without estimating the unknown kernel in [37]. CNN is also utilized along with RNN for image deblurring in [41]. LSTM and CNNs are combined in a proposed SRN-DeblurNet to tackle image deblurring in a multi-scale manner by Tao *et al.* [42]. A nested skip connection structure is developed in [47]. In addition, there also exist some methods focusing on recover sharp high-resolution images from blurry low-resolution images. Zhang *et al.* [72] propose a deep plug-and-play super-resolution network to handle LR image with arbitrary blur kernels. This method is a non-blind deblurring SR method, which relies on known blur kernels. For blind deblurring SR, the most related work is GFN [73]. This method directly extracts the spatial information from a motion-blurred image to recover its sharp SR version, but ignores the temporal information implied in the motion-blurred image.

In this paper, we address the problem of blind motion deblurring super-resolution, which is a more difficult task than the individual problems of image deblurring and super-resolution. One reason is that the motion-blurred image includes spatio-temporal information, which is difficult to extract. In this paper, we propose a new method to address this problem via employing the “divide and conquer” scheme. Specially, to extract the spatio-temporal information, we first use a “from static to dynamic” network to generate a sequence of LR sharp images from an input motion blurred image. During the training stage, the input and output of the “from static to dynamic” network are blurry image and its corresponding sharp frames, respectively. Therefore, it has better ability to extract the information implied in the motion-blurred image. Then, we can use the restored sharp images to help the following network to finish the video super-resolution process and obtain the final results.

3 PROPOSED METHOD

3.1 The Overall Architecture of BMDSRNet

Fig. 1 shows the overview architecture of BMDSRNet. Primarily, it is composed of two parts, aiming to solve the “static-to-dynamic” and “LR-to-HR” problems individually. The first part is a designated Blind Motion Deblurring Net (BMDNet), which takes a given low resolution image with motion blur as input and outputs a sequence of motion deblurred images, corresponding to the multiple clear images during the exposure time period. This part does not address the resolution issue. The second part specifically focuses

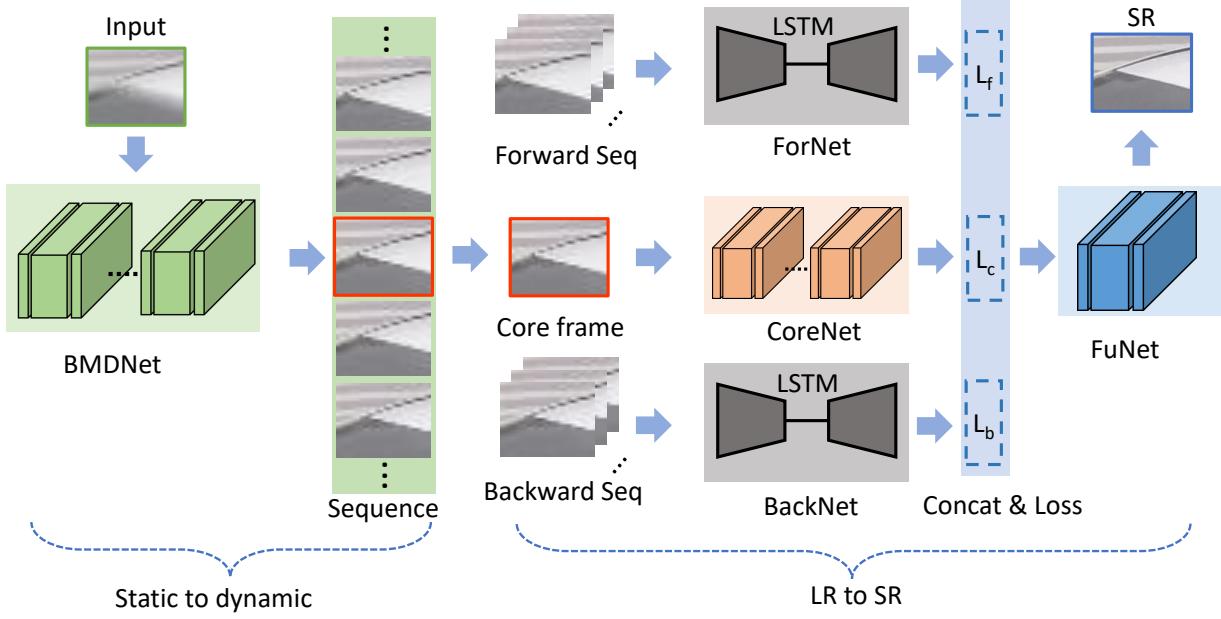


Fig. 1. The architecture of the motion deblurring super-resolution networks. Static to dynamic: One Static low-resolution motion blurred image is put into our model to remove unwanted blur and extract a video sequence. LR to SR: The generated video sequence is fed into three-stream networks, which includes ForNet, CoreNet and BackNet, to generate a high-resolution image.

on the super-resolution aspect. This part consists of three parallel streams, a ForNet, a CoreNet, and a BackNet. The ForNet processes the sequence of images from BMDNet in a forward direction. On the contrary, the BackNet learns the backward temporal information by processing the reverse order of the sequence. Both of ForNet and BackNet are of the LSTM structure and share weights during the training stage. The CoreNet operates on the central frame as it is the most important one among the sequence. The outputs of these three streams are concatenated and fused by a FuNet aiming at recovering a high resolution image corresponding to the given low resolution image, free of motion blur.

3.2 From Static to Dynamic

The first part in the proposed BMDSRNet aims to extract a set of continuous frames from a given motion-blurred image. The intuition behind is that, a motion blurred image can be considered as an accumulation of multiple instant frames. By decomposing the blurred image into multiple clear images with BMDNet (as Fig. 2 shows), we can transform a *static* motion-blurred image into a sequence of *dynamic* frames. This strategy benefits our ultimate goal in two aspects. Firstly, this kind of decomposition solves the deblurring problem. Secondly, by transforming a single image into multiple continuous frames, the difficulty of super-resolution is eased as we can subsequently utilize the spatio-temporal cues contained in the multiple resulted frames, as illustrated in [30].

To accomplish this “*static-to-dynamic*” task, we develop a neural network called BMDNet. Fig. 3 presents the structure of BMDNet. The input is a single motion-blurred image and the output is seven continuous clear frames. There are sequentially two convolutional layers of kernel size 3×3 , nine residual blocks [74] and three convolutional layers of kernel size 3×3 . Through the last convolutional

layer, the output feature channels turn 21, corresponding to seven frames. Table 1 illustrates the detailed configuration of BMDNet.

Let the input of BMDNet be denoted as X_{blurry} , and the output as $\{Y_{out}^i, i = 1, 2, \dots, 7\}$. In our practice, we use seven consecutive clear frames $\{Y_{sharp}^i, i = 1, 2, \dots, 7\}$ to synthesize the motion-blurred image. Thus, the used seven clear frames serve as the ground truth frames. The cost function of training the BMDNet is composed of a term of reconstruction loss \mathcal{L}_{S2D}^c regarding the central frame and terms of pair-wise content loss \mathcal{L}_{S2D}^p regarding the other six frames. The “S2D” here indicates “static-to-dynamic”.

The reconstruction content loss of the central frame is

$$\mathcal{L}_{S2D}^c = \|Y_{sharp}^4 - Y_{out}^4\| + \|\Phi(Y_{sharp}^4) - \Phi(Y_{out}^4)\|, \quad (1)$$

where the Φ extracts features using the last convolutional layer of VGG19 network [75]. This loss term measures both the pixel-wise and perceptual level error between the recovered frame and the ground truth.

Given a set of continuous images, changing the order of images would not change the accumulation of multiple frames. This means there could be multiple possible solutions to the reverse process, *i.e.* extracting multiple frames from a given image. Fig. 2(a) and 2(b) respectively show the concept and an example of decomposing a single image into possible two sets of sequential images. This means it is not appropriate to use image-wise content loss like the central frame for the other six images. Thus, following [76], we also employ a pair-wise content loss to ensure the output sequence frames are reasonable, formulated as,

$$\begin{aligned} \mathcal{L}_{S2D}^p = & \sum_{i=1}^3 \left([Y_{sharp}^i, Y_{sharp}^{8-i}]_+ - [Y_{out}^i, Y_{out}^{8-i}]_+ \right) \\ & + \left([Y_{sharp}^i, Y_{sharp}^{8-i}]_- - [Y_{out}^i, Y_{out}^{8-i}]_- \right), \end{aligned} \quad (2)$$

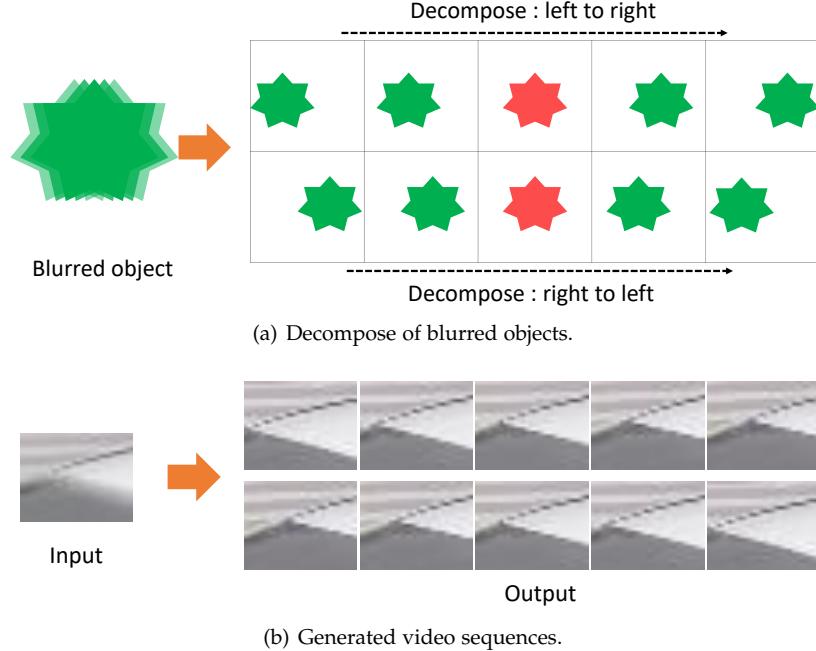


Fig. 2. The analyses of motion blurred images. (a): Instant frames are accumulated over time to create a motion blurred images, which thus can be decomposed to different sequence. (b): Two different video sequences from a same motion blurred image.

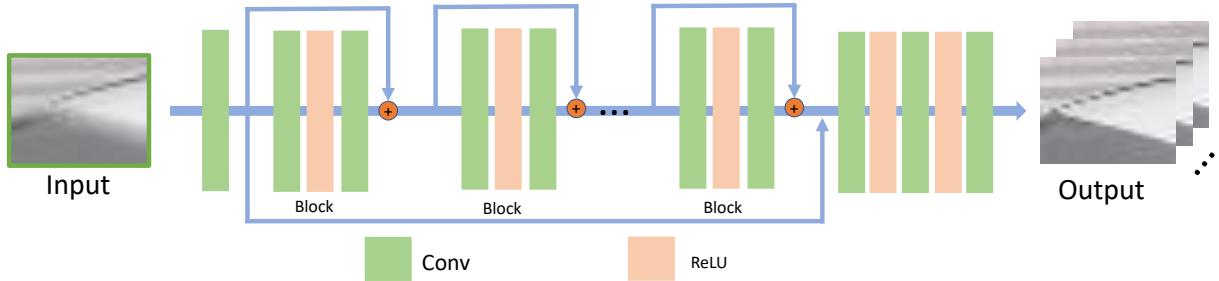


Fig. 3. The architecture of the BMDNet. One blurred image is put into BMDNet to generate a video sequence.

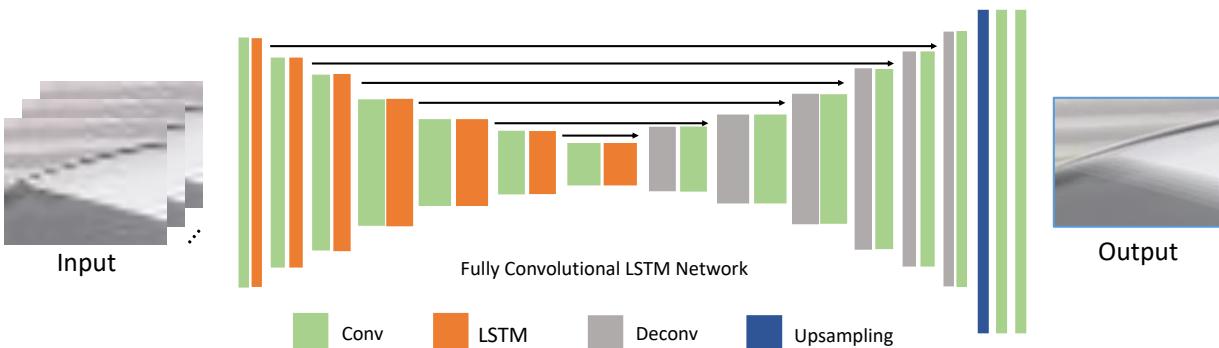


Fig. 4. The architecture of the ForNet and BackNet. One video sequence is put into to generate a SR clean image. The input are the multi-frames generated by the BMDNet.

where $[\cdot, \cdot]_+$ and $[\cdot, \cdot]_-$ denote the pixel-wise summation and subtraction between a pair of images, respectively.

Notably, it is not the first time in the community that a neural network is proposed to extract multiple frames from a single static image. [76] is a pioneer work. However, our BMDNet is distinctly different from the method in [76].

Our method uses only a single model to carry out the task of extracting multiple frames, but [76] requires multiple models for the task (as many as the number of output frames), which is not efficient in practice.

In the real world, it is difficult to determine if the input image is "LR" or "HR". This is because the LR and

TABLE 1

Configurations of the proposed BMDNet. It is composed of two convolutional layers (L1 and L2), nine residual blocks and three additional convolutional layers (L3, L4 and L5). Each residual block contains two convolutional layers. The output size is the same to input.

layers	Kernel size	output channels	operations
L1	3×3	64	-
L2	3×3	64	ReLU
B1-B8	3×3	64	ReLU
B9	3×3	64	ReLU
L3	3×3	64	ReLU
L4	3×3	64	ReLU
L5	3×3	21	-

TABLE 2

Configurations of the proposed ForNet and BackNet. It is composed of convolutional, ConvLSTM and deconvolutional layers. The output size is N times larger than input.

layers	Kernel size	output channels
Conv+ConvLSTM	3×3	32
Conv+ConvLSTM	3×3	64
Conv+ConvLSTM	3×3	128
Conv+ConvLSTM	3×3	256
Conv+ConvLSTM	3×3	256
Conv+ConvLSTM	3×3	256
Conv+ConvLSTM	3×3	512
DeConv+Concat+Conv	3×3	256
DeConv+Concat+Conv	3×3	128
DeConv+Concat+Conv	3×3	128
DeConv+Concat+Conv	3×3	128
DeConv+Concat+Conv	3×3	64
DeConv+Concat+Conv	3×3	32
DeConv	3×3	32
Upsampling+Conv	3×3	32
Conv	3×3	3

HR are relative concepts. For example, the 4K images are high-resolution compared to the 2K images, but are low-resolution compared to the 8K versions. The “From static to dynamic” network is proposed to extract a sequence of sharp images from a motion-blurred image. Therefore, we mainly ensure that the input and ground truth images are the motion-blurred images and their seven neighboring sharp images during the training stage.

3.3 From Low-resolution to High-resolution

The first part of BMDSRNet, *i.e.* BMDNet, has decomposed a single low resolution into seven low resolution frames free of blur artifact. These multiple continuous frames benefit the multi-frame super resolution task, accomplished by the second part.

There are three streams in the second part of BMDSRNet, *i.e.*, ForNet, CoreNet and BackNet. As mentioned before, there could be multiple reasonable solutions to the “static-to-dynamic” sub-problem. Fig. 2 shows an example of two possible solutions of different directions, which both can result in the blurred image. This inspires us to learn the temporal dynamics in both the forward and backward directions, to cover the variety. To this end, the ForNet and

TABLE 3

Configurations of the proposed CoreNet. It is composed of two convolutional layers (L1 and L2), 9 residual blocks and three additional convolutional layers (L3, L4 and L5). Each residual block contains two convolutional layers. The output size is N times larger than input.

layers	Kernel size	output channels	operations
L1	3×3	64	-
L2	3×3	64	ReLU
B1-B8	3×3	64	ReLU
B9	3×3	64	ReLU+Upsampling
L3	3×3	64	ReLU
L4	3×3	64	ReLU
L5	3×3	3	-

BackNet are designed for the purpose of learning both the forward and backward motion.

ForNet differs from BackNet in terms of the input. We reverse the order of input frames of ForNet as the input to the BackNet. Both of them are networks of fully convolutional layers with LSTM structure. As Fig. 4 shows, taking ForNet as an example, it consists of several layers of convolution and ConvLSTM, followed by several layers of deconvolution, convolution and upsampling. The size of all the kernels is 3×3 . Table 2 represents the configurations of ForNet and BackNet.

The CoreNet is specifically for the central frame, as the center frame is the most important one in the sequence. Similarly, it is composed of a sequence of convolutional layers, residual blocks and convolutional layers, as shown in Fig. 5. The detailed configurations of CoreNet is given in Table 3.

In order to push the intermediate output become the sharp high-resolution images, we train out ForNet, BackNet and CoreNet based on MSE criterion. The loss function can be formulated as:

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

where W and H are the width and height of a frame, $I_{x,y}^{sharp}$ is the value of high-resolution images at location (x, y) , and $G(I_{x,y}^{blurry})$ corresponds to the value of super-resolution images which are recovered from ForNet and BackNet.

Finally, we develop a FuNet to fuse three different intermediate results and generate the final finer high-resolution images. FuNet has a similar structure as CoreNet, but with two primary differences. Firstly, it is a smaller structure, which has only three blocks. The reason is that the input to FuNet is already high-quality super-resolution images generated by the preceding sub-networks. The second difference is that FuNet does not have the upsampling layer. Fig. 6 shows the architecture of FuNet.

4 EXPERIMENTS

4.1 Motion-Blurred LR Dataset

One of the most popular motion-blurred datasets is the GOPRO dataset [38]. Using a high-speed camera, Nah *et*

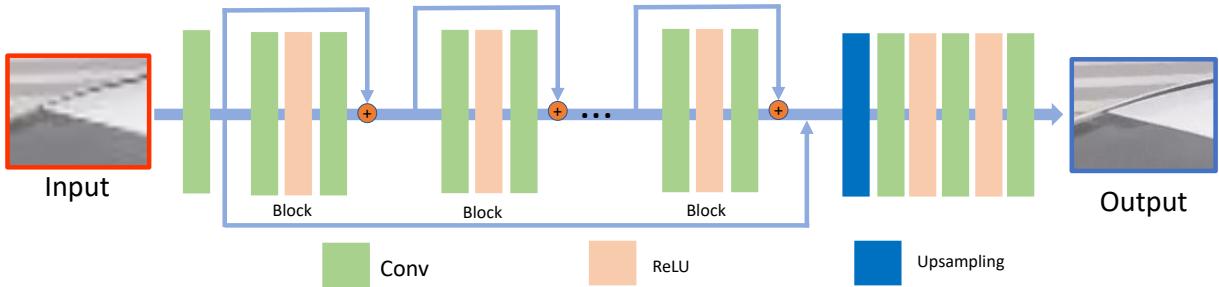


Fig. 5. The architecture of the CoreNet. One low-level image is put into to generate a super-resolution image. Then input is the center frame generated by BMDNet.

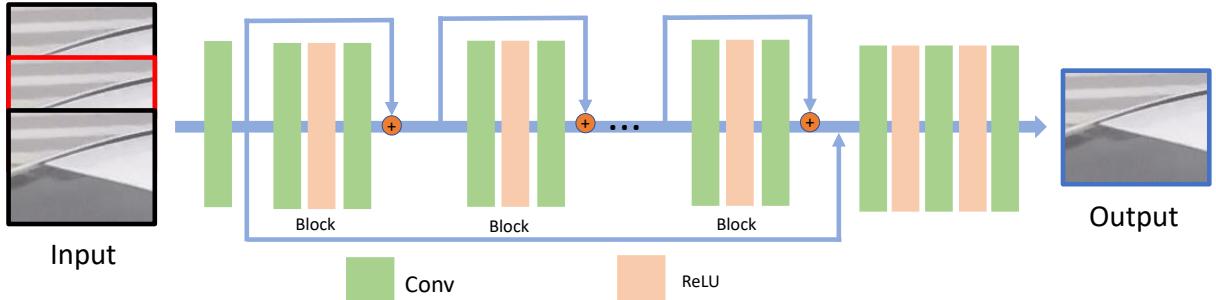


Fig. 6. The architecture of the FuNet. Three intermediate results generated by preceding networks are utilized to recover the final finer super-resolution images. Then input are three images, which are the output of ForNet, CoreNet and BackNet, respectively.

al. capture 33 videos (22 for training and 11 for testing respectively), and then synthesize 3,214 pairs of blurry image and sharp image for training and testing. This synthesized dataset cannot be utilized to train our model directly because we need a blurry image and its corresponding seven sharp images. to learn the process of “*static-to-dynamic*”. In order to address these problem, we re-synthesize a Motion-Blurred LR dataset based on the GOPRO dataset. We firstly extract all the frames from the 33 provided videos. Then we synthesize blurry images via averaging the seven neighbouring sharp images. By doing this, a motion-blurred image corresponds to seven sharp images and thus we can train our BMDNet to learn the process of “*static-to-dynamic*”. In order to model the stage of “*from low-resolution to high-resolution*”, we further down-sample the images to generate low-resolution counterparts. The training and testing sets are generated based on 22 and 11 videos, respectively, which is same as the original GOPRO dataset.

4.2 Ablation Study

In order to evaluate the effects of different components in the proposed model, we develop four networks, SRNet, BMDSRNet(C), BMDSRNet(F+C) and BMDSRNet(F+C+B). SRNet is a simplified version of BMDSRNet without the BMDNet module. BMDSRNet(C), BMDSRNet(F+C) and BMDSRNet(F+C+B) are three versions, whose inputs to the stage of *LR to SR* are different. All these networks are illustrated in the following. During the training stage, we update all weights with a mini-batch of size 4 in each iteration. 128 × 128 patches are cropped at random locations. The learning rate and training epoch is 0.0001 and 400, respectively. To

TABLE 4
Ablation study for scale factor 2, 3 and 4 on the Motion-Blurred LR dataset in terms of PSNR and SSIM.

Scale	Methods	PSNR	SSIM
× 2	SRNet	30.88	0.9437
× 2	BMDSRNet(C)	31.28	0.9457
× 2	BMDSRNet(F+C)	31.45	0.9479
× 2	BMDSRNet(F+C+B)	31.62	0.9483
× 3	SRNet	29.49	0.9256
× 3	BMDSRNet(C)	29.95	0.9302
× 3	BMDSRNet(F+C)	30.21	0.9337
× 3	BMDSRNet(F+C+B)	30.44	0.9356
× 4	SRNet	28.06	0.8997
× 4	BMDSRNet(C)	28.43	0.9068
× 4	BMDSRNet(F+C)	28.62	0.9113
× 4	BMDSRNet(F+C+B)	28.78	0.9132

train the final BMDSRNet, the weight of loss functions in “*static to dynamic*” and “*LR to SR*” are the same.

SRNet. The architecture of SRNet is same to CoreNet. The main difference is that the input to SRNet is the blind motion-blurred images, rather than the deblurred results from BMDNet.

BMDSRNet(C). It consists of BMDNet and CoreNet. One motion-blurred image is put into BMDNet to recover seven deblurred low-resolution images. Then we input the central frame of them to CoreNet to generate the sharp high-resolution image.

BMDSRNet(F+C). It consists of BMDNet, CoreNet, ForNet and FuNet. The main difference from BMDSRNet(C) is that it uses an additional ForNet module to extract temporal

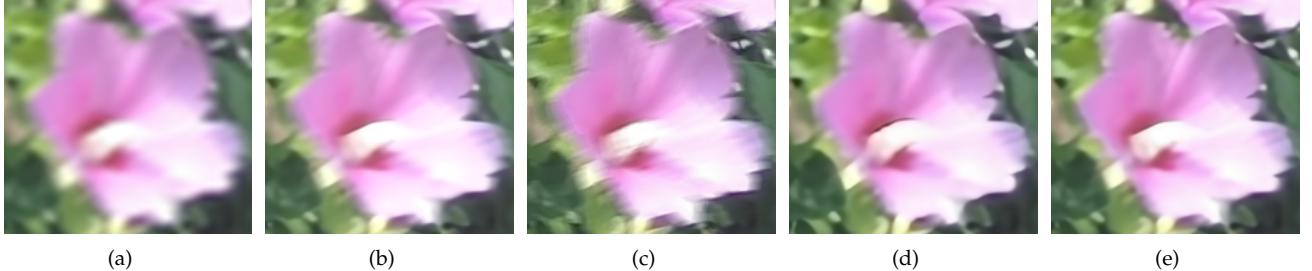


Fig. 7. Examples of blind motion deblurring super-resolution. From left to right are the input image, deblurring SR results from SRNet, BMDSRNet(C), BMDSRNet(F+C) and BMDSRNet(F+C+B), respectively. Zoom in the figure for better visibility.



Fig. 8. Comparison with state-of-the-art deblurring and super-resolution methods. From left to right are input blurry image, RCAN+SRN, SRFBN+SRN, GFN and BMDSRNet, respectively. Zoom in the figure for better visibility.



Fig. 9. Comparison with state-of-the-art deblurring and super-resolution methods on the Motion-blur LR GOPRO dataset [38]. From left to right are the input blurry image, RCAN+SRN, SRFBN+SRN, GFN and BMDSRNet, respectively. Zoom in the figure for better visibility.



Fig. 10. Comparison with state-of-the-art deblurring and super-resolution methods on the Motion-blur LR GOPRO dataset [38]. From left to right are the input blurry image, RCAN+SRN, SRFBN+SRN, GFN and BMDSRNet, respectively. Zoom in the figure for better visibility.

information from the deblurred neighboring frames. Then the FuNet generates the final deblurred high-resolution images based on the results from CoreNet and ForNet.

BMDSRNet(F+C+B). This architecture is our whole blind motion-deblurred super-resolution networks. It consists of BMDNet, CoreNet, ForNet, BackNet and FuNet. ForNet and BackNet extract temporal information from bidirection to help FuNet generate final deblurred high-resolution images.

Results on different scales. We report the PSNR and SSIM of the aforementioned four models on three down-sampling scales. The quantitative and visual results are shown in Tab. 4 and Fig. 7. The performance difference between BMDSRNet(C) and SRNet illustrates the effect of our “divide-and-conquer” strategy. BMDSRNet(F+C) achieves better performance than BMDSRNet(C), which corresponds to the well-known knowledge that multi-frame

super-resolution methods have advantage over the single image super-resolution methods. This demonstrates the effect of the “static-to-dynamic” stage. The BMDSRNet(F+C+B) is better than BMDSRNet(F+C), suggesting extracting temporal features from bidirection is helpful to generate high-resolution images.

4.3 Comparison with State-of-the-art Methods

We have verified the effects of different parts of the proposed BMDSRNet. In this section, we compare our method with other state-of-the-art algorithms, including two SR methods (RCAN [25], SRFBN [27]), one motion-deblurred method (SRN [42]), and methods which jointly address SR and image deblurring tasks. RCAN is one of the best methods for bicubic degradation. SRFBN consists of several feedback blocks and achieves state-of-the-art performance



Fig. 11. Qualitative comparison on the real motion-blurred dataset [56]. The input image is shown in the first row. The second and third rows show results of the methods by GFN [73] and BMDSRNet, respectively. Zoom in the figure for better visibility.

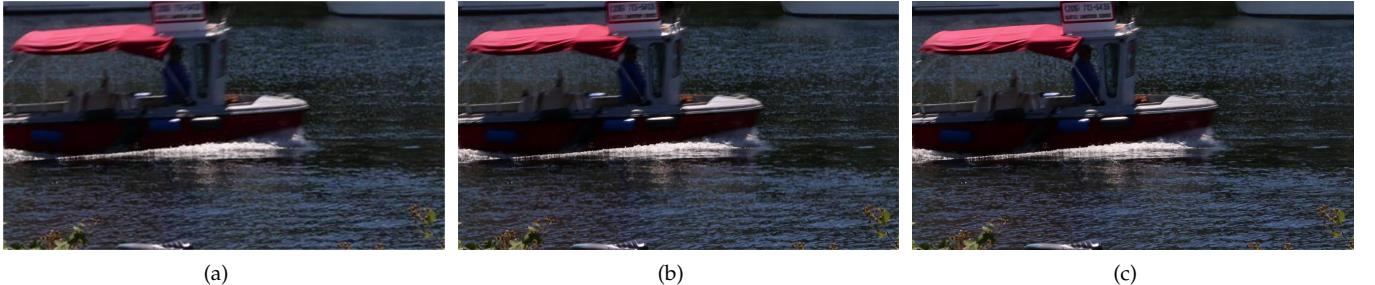


Fig. 12. Qualitative comparison on the real motion-blurred dataset [56]. The input image is shown in the first row. The second and third rows show results of the methods by GFN [73] and BMDSRNet, respectively. Zoom in the figure for better visibility.



Fig. 13. Qualitative comparison on the real motion-blurred dataset [56]. The input image is shown in the first row. The second and third rows show results of the methods by GFN [73] and BMDSRNet, respectively. Zoom in the figure for better visibility.

TABLE 5

Performance of different model structures on the Motion-Blurred LR dataset in terms of PSNR and SSIM for scale factors 2, 3 and 4, respectively.

Method	$\times 2$		$\times 3$		$\times 4$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN	27.57	0.8908	27.18	0.8846	26.65	0.8737
RCAN + SRN	30.94	0.9443	29.54	0.9264	28.18	0.9029
SRN + RCAN	28.88	0.9180	27.08	0.9059	27.56	0.8938
SRFBN	27.57	0.8907	27.16	0.8843	26.63	0.8730
SRFBN + SRN	30.92	0.9196	29.50	0.8901	28.09	0.8504
SRN + SRFBN	30.40	0.9119	29.03	0.8856	27.63	0.8400
GFN	-	-	-	-	27.47	0.8926
BMDSRNet	31.62	0.9483	30.44	0.9356	28.78	0.9132

for SR. SRN is one popular image deblurring method. Specially, it is trained without adversarial loss and one of the state-of-the-art methods on removing motion blur [77]. We directly use the official deblurred models which are trained on the GOPRO dataset. For fair comparisons, we also re-train the two SR methods on the GOPRO dataset. The values of PSNR and SSIM of different methods on

three different motion-blurry LR sets are shown in Tab. 5 and Fig. 8. Results show that the proposed BMDSRNet outperforms two popular SR methods, and the combination of SR and image deblurring methods, and the previous blind deblurring super-resolution network, i.e., GFN.

TABLE 6

Performance of different model structures on the LR GORPO dataset in terms of PSNR and SSIM for scale factor 4.

Method	PSNR	SSIM	Params	Times
SCGAN	22.74	0.783	1.1M	0.66
SRResNet	24.40	0.827	1.5M	0.07
EDSR	24.52	0.836	43M	2.10
DeepDeblur + SRResNet	24.99	0.827	13M	0.66
SRResNet + DeepDeblur	25.93	0.850	13M	6.06
DeepDeblur + ED-DSRN	21.53	0.682	54M	2.18
ED-DSRN + DeepDeblur	24.66	0.827	54M	2.95
GfN	27.74	0.896	12M	0.07
BMDSRNet	28.15	0.904	2.9M	0.11

4.4 Test on the LR GOPRO Dataset

In order to further evaluate the proposed method, we also test it on the LR GOPRO dataset. This dataset is synthesized based on GOPRO dataset via down-sampling to downscale blurry images by $4\times$ to generate blurry LR images. The ground-truth sharp high-resolution images are the original sharp images in the GOPRO dataset.

We compare our method with two SOTA SR methods (SRResNet [21], EDSR [17]), two joint image deblurring and SR approaches (SCGAN [78], GfN [73]), and the combinations of blind image deblurring algorithms (DeepDeblur [38]) and SR algorithms (SRResNet [21], ED-DSRN [79]). The values of PSNR and SSIM of different methods on the LR GOPRO set are shown in Tab. 6. Fig. 9 and 10 show the qualitative results. The results show that the proposed BMDSRNet does not only outperform the traditional SR methods, and the joint image deblurring and SR approaches, but also achieves better performance than the previous state-of-the-art method, GfN, which is also designed for blind motion deblurring super-resolution.

As we all know that blind motion deblurring super-resolution is a more ill-posed problem. The proposed BMDSRNet achieves better performance because we employ the “divide-and-conquer” scheme, rather than in-one-go SR network. Results are reported in Tab. 6. Firstly, the RCAN and SRFBN are two state-of-the-art SR methods, which can directly super-resolve an LR blurry image to an HR sharp image. However, these methods achieve worse performance than RCAN + SRN and SRFBN + SRN. SRN is one of the state-of-the-art deblurring methods. It shows that “divide-and-conquer” scheme is a better option than an in-one-go network for the blind motion deblurring super-resolution. Secondly, the SRFBN + SRN and SRFBN + SRN achieve worse performance than the proposed BMDSRNet, showing that extracting a sequence from a blurry image is better than using only the intermediate sharp image. Thirdly, the previous work [4] shows that the intermediate sharp image in the extracted sharp sequence is of high-quality with satisfied PSNR values, guaranteeing the quality of intermediate sharp images. Finally, SRN + RCAN and SRN + SRFBN, which firstly predict one sharp LR image and then super-resolve to corresponding HR version, achieve worse performance than firstly extracting a sharp sequence of LR images. It verifies that “only predict one sharp LR image” is worse than predicting a sharp sequence.

4.5 Test on the real motion-blurred Dataset

Then we compare the performance of our approach with the previous state-of-the-art blind motion-blurred SR method, i.e., GFN [73], on real blurry images [56]. The results can refer to Fig. 11, 12 and 13.

5 CONCLUSION

In this paper, we propose a blind motion deblurring super-resolution networks to recover sharp high-resolution images from motion-blurred low-resolution input. Our main contribution is the novel use of spatio-temporal information implied in motion-blurred images for single image super-resolution. We first design a motion deblurring network which can model the reverse process of generation of motion-blurred images. This network can extract several neighbouring frames and thus transfer the static super-resolution to dynamic super-resolution. Then a well-designed three-streams network is developed to learn bidirectional spatio-temporal information to recover the final sharp high-resolution images. The experimental results on two datasets suggest that the proposed model outperforms existing methods for super-resolving blind motion-blurred images.

ACKNOWLEDGMENT

This work is funded in part by the ARC Centre of Excellence for Robotics Vision (CE140100016), ARC-Discovery (DP 190102261) and ARC-LIEF (190100080) grants, as well as a research grant from Baidu on autonomous driving. The authors gratefully acknowledge the GPUs donated by NVIDIA Corporation. We thank all anonymous reviewers and editors for their constructive comments.

REFERENCES

- [1] Z. Wang, J. Chen, and S. C. Hoi, “Deep learning for image super-resolution: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [2] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, “Single image super-resolution via a holistic attention network,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 191–207.
- [3] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, “Video super-resolution with convolutional neural networks,” *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.
- [4] M. Jin, M. Hirsch, and P. Favaro, “Learning face deblurring fast and wide,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2018.
- [5] X. Li and M. T. Orchard, “New edge-directed interpolation,” *IEEE TIP*, 2001.
- [6] L. Zhang and X. Wu, “An edge-guided image interpolation algorithm via directional filtering and data fusion,” *IEEE TIP*, 2006.
- [7] J. Sun, Z. Xu, and H.-Y. Shum, “Image super-resolution using gradient profile prior,” in *CVPR*, 2008.
- [8] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, “Coupled dictionary training for image super-resolution,” *IEEE TIP*, 2012.
- [9] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” *IEEE TIP*, 2010.
- [10] L. Zheng, Y. Li, K. Zhang, and W. Luo, “T-net: Deep stacked scale-iteration network for image dehazing,” *arXiv preprint arXiv:2106.02809*, 2021.
- [11] K. Zhang, R. Li, Y. Yu, W. Luo, C. Li, and H. Li, “Deep dense multi-scale network for snow removal using semantic and geometric priors,” *arXiv preprint arXiv:2103.11298*, 2021.

- [12] K. Zhang, W. Luo, Y. Yu, W. Ren, F. Zhao, C. Li, L. Ma, W. Liu, and H. Li, "Beyond monocular deraining: Parallel stereo deraining network via semantic prior," *arXiv preprint arXiv:2105.03830*, 2021.
- [13] K. Zhang, W. Luo, W. Ren, J. Wang, F. Zhao, L. Ma, and H. Li, "Beyond monocular deraining: Stereo image deraining via semantic understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 71–89.
- [14] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014, pp. 184–199.
- [15] ———, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
- [16] J. Kim, J. Kwon Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *CVPR*, 2016.
- [17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPR*, 2017.
- [18] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 624–632.
- [19] J. Kim, J. Kwon Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016.
- [20] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 391–407.
- [21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.
- [22] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "EsrGAN: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 0–0.
- [23] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500.
- [24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.
- [25] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.
- [26] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 065–11 074.
- [27] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *CVPR*, 2019.
- [28] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3217–3226.
- [29] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271.
- [30] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Neurips*, 2015.
- [31] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *ECCV*, 2018.
- [32] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.
- [33] T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE TIP*, 1998.
- [34] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *IEEE TOG*, 2008.
- [35] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *NIPS*, 2009.
- [36] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *CVPR*, 2015.
- [37] A. Chakrabarti, "A neural approach to blind motion deblurring," in *ECCV*, 2016.
- [38] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017, pp. 3883–3891.
- [39] T. M. Nimisha, A. Kumar Singh, and A. N. Rajagopalan, "Blur-invariant deep learning for blind-deblurring," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4752–4760.
- [40] X. Xu, J. Pan, Y.-J. Zhang, and M.-H. Yang, "Motion blur kernel estimation via deep learning," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 194–205, 2017.
- [41] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R. W. Lau, and M.-H. Yang, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *CVPR*, 2018.
- [42] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *CVPR*, 2018.
- [43] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *CVPR*, 2018.
- [44] T. Madam Nimisha, K. Sunil, and A. Rajagopalan, "Unsupervised class-specific deblurring," in *European Conference on Computer Vision*, 2018.
- [45] D. Ren, K. Zhang, Q. Wang, Q. Hu, and W. Zuo, "Neural blind deconvolution using deep priors," *arXiv preprint arXiv:1908.02197*, 2019.
- [46] J. Mustaniemi, J. Kannala, S. Särkkä, J. Matas, and J. Heikkilä, "Gyroscope-aided motion deblurring with deep networks," in *IEEE Winter Conference on Applications of Computer Vision*, 2019.
- [47] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *CVPR*, 2019.
- [48] B. Lu, J.-C. Chen, and R. Chellappa, "Unsupervised domain-specific deblurring via disentangled representations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [49] R. Aljadaany, D. K. Pal, and M. Savvides, "Douglas-rachford networks: Learning both the image prior and data fidelity terms for blind image deconvolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [50] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *IEEE International Conference on Computer Vision*, 2019.
- [51] K. Purohit and A. Rajagopalan, "Region-adaptive dense network for efficient motion deblurring," *arXiv preprint arXiv:1903.11394*, 2019.
- [52] K. Zhang, W. Luo, Y. Zhong, B. Stenger, L. Ma, W. Liu, and H. Li, "Deblurring by realistic blurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [53] M. Suin, K. Purohit, and A. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," *arXiv preprint arXiv:2004.05343*, 2020.
- [54] A. Kaufman and R. Fattal, "Deblurring using analysis-synthesis networks pair," *arXiv preprint arXiv:2004.02956*, 2020.
- [55] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, "Learning event-based motion deblurring," *arXiv preprint arXiv:2004.05794*, 2020.
- [56] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *CVPR*, 2017.
- [57] T. H. Kim, K. M. Lee, B. Schölkopf, and M. Hirsch, "Online video deblurring via dynamic temporal blending network," in *IEEE International Conference on Computer Vision*, 2017.
- [58] M. Aittala and F. Durand, "Burst image deblurring using permutation invariant convolutional neural networks," in *European Conference on Computer Vision*, 2018.
- [59] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 291–301, 2018.
- [60] H. Chen, J. Gu, O. Gallo, M.-Y. Liu, A. Veeraraghavan, and J. Kautz, "Reblur2deblur: Deblurring videos via self-supervised learning," in *ICCP*, 2018.
- [61] S. Nah, S. Son, and K. M. Lee, "Recurrent neural networks with intra-frame iterations for video deblurring," in *CVPR*, 2019.
- [62] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "EDVR: Video restoration with enhanced deformable convolu-

- tional networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2019.
- [63] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *IEEE International Conference on Computer Vision*, 2019.
- [64] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [65] D. Li, C. Xu, K. Zhang, X. Yu, Y. Zhong, W. Ren, H. Suominen, and H. Li, "Arvo: Learning all-range volumetric correspondence for video deblurring," *arXiv preprint arXiv:2103.04260*, 2021.
- [66] K. Zhang, W. Luo, B. Stenger, W. Ren, L. Ma, and H. Li, "Every moment matters: Detail-aware networks to bring a blurry image alive," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 384–392.
- [67] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1439–1451, 2015.
- [68] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *NIPS*, 2014.
- [69] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *CVPR*, 2018.
- [70] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *ICCV*, 2019.
- [71] K. Purohit, A. Shah, and A. Rajagopalan, "Bringing alive blurred moments," in *CVPR*, 2019.
- [72] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [73] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for joint image deblurring and super-resolution," 2018.
- [74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.
- [76] M. Jin, G. Meishvili, and P. Favaro, "Learning to extract a video sequence from a single motion-blurred image," in *CVPR*, 2018.
- [77] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *European Conference on Computer Vision*. Springer, 2020, pp. 184–201.
- [78] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *IEEE International Conference on Computer Vision*, 2017.
- [79] X. Zhang, F. Wang, H. Dong, and Y. Guo, "A deep encoder-decoder networks for joint deblurring and super-resolution," in *ICASSP*, 2018.