Computer Vision and Image Understanding

journal homepage: www.elsevier.com

Attentive Deep Network for Blind Motion Deblurring on Dynamic Scenes

Yong Xu[a,b], Ye Zhu[a], Yuhui Quan[a,**], Hui Ji[c]

[a] *School of Computer Science and Engineering, South China University of Technology,Guangzhou 510006, China*
[b] *Peng Cheng Laboratory, Shenzhen, 518052, China*
[c] *Department of Mathematics, National University of Singapore, Singapore 119076, Singapore*

ABSTRACT

Non-uniform blind motion deblurring is a challenging yet important problem in image processing that receives enduring attention in the last decade. The non-uniformity nature of motion blurring leads to great variations on the blurring effects across image regions and over different images, which makes it very difficult to train an end-to-end deblurring neural network (NN) with good generalization performance. This paper introduces an attention mechanism for the blind deblurring NN, including both spatial and channel attention, so as to effectively handle the significant spatial variations on blurring effects. In the attention mechanism, the spatial attention is introduced in both the encoder for discriminative exploitation of image edges and smooth regions and the decoder for discriminative treatment on different regions with different blurring effects. The channel attention is introduced for better generalization performance of the NN, as it allows adaptive weighting on intermediate features for a particular image. Building such an attention mechanism into a multi-scale encoder-decoder framework, an attentive NN is developed for practical non-uniform blind image deblurring. The experiments on several benchmark datasets show that the proposed NN can effectively restore the images degraded by spatially-varying blurring, with state-of-the-art performance.

1. Introduction

Image blurring is one often-seen type of image degradation, which causes the loss of image details. In addition to yielding poor picture quality unwanted in digital photography, image blurring also have negative impacts on many vision tasks, *e.g.* automatic driving, object tracking, and visual surveillance. Image deblurring is then about recovering a clear image with sharp details from an input blurred image. The effect of image blurring may come from multiple sources, *e.g.* out-of-focus and motion blur. In practice, motion blurring effect often is non-uniform (spatially-varying), *i.e.*, different image regions have different blurring effects, when the camera motion is not the translation along image plane, or there are large variations on scene depth, or there exist independent moving objects. This paper focuses on the study of how to remove spatially-varying motion blurring effect from the images of dynamic scenes.

This paper concerns the motion blurring process that can be formulated as:

$$\boldsymbol{g} = \boldsymbol{K}\boldsymbol{f} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{g}$ denotes the input blurred image, $\boldsymbol{f}$ denotes the latent image with sharp details, $\boldsymbol{n}$ denotes the measurement noise, and $\boldsymbol{K}$ denotes some linear operator that models the blurring process such that $\sum_i \boldsymbol{K}(i,j) = 1, \forall j$ and $\boldsymbol{K}(i,j) \geq 0, \forall i,j$. In other

**Corresponding author
e-mail: csyhquan@scut.edu.cn (Yuhui Quan)

words, the value of each blurred pixel is the weighted average of the values of all its neighboring sharp pixels (Denis et al., 2015). It is noted that the linear model (1) is not applicable to the images where there are occlusions occurring during shutter time. In the case of uniform blurring, as most existing deblurring methods (*e.g.* (Cai et al., 2009; Danielyan et al., 2011; Shan et al., 2008; Xu and Jia, 2010)) assume, all blurred pixels take the same weighting average scheme. Thus, the operator $\boldsymbol{K}$ can be expressed as a convolution operator with a smoothing kernel. When the blurring is not uniform over the image, different blurred pixels will take different weighting schemes. Clearly, as both the operator $\boldsymbol{K}$ and image $\boldsymbol{f}$ are unknown, the blind deblurring is a very challenging ill-posed problem to solve.

For images with complex spatially-varying blurring effects ((Nah et al., 2017; Caglioti and Giusti, 2009; Seibold et al., 2017)), the operator $\boldsymbol{K}$ has a very high degree of freedom, which makes it very difficult to resolve the ambiguity of the solution in blind image deblurring. In the past, there have been several approaches that impose certain structural prior on the blurring operator $\boldsymbol{K}$, *e.g.* two-layer-based model for defocus blurring (Chan and Nguyen, 2011), patch-based model for non-uniform motion blurring (Ji and Wang, 2012), and non-uniform motion blurring model parameterized by 3D camera intrinsic motion (Whyte et al., 2012). Nevertheless, the applicability of these models is limited. For instance, they are not applicable to the blurring effects in dynamic scenes with moving objects of different speeds or for the blurring effects caused by very complex scene depths.

In order to have a blind deblurring method that covers a wide range of spatially-varying blurring effects, an alternative approach is to directly recover each blurred image pixel without explicitly modeling its associated blurring operator. Deep learning provides a powerful tool to learn such a direct recovery process. In recent years, many deep-learning-based approaches (*e.g.* (Nah et al., 2017; Nimisha et al., 2017; Sharma et al., 2018; Su et al., 2017; Xin et al., 2018)) have been proposed for blind image deblurring. Most of these methods train a convolutional neural network (CNN) that models the mapping between a blurred image to its clear version, using many pairs of blurred images and their clear versions. The NN models trained by these approaches have shown promising performance on removing spatially-varying blur from input blurred images.

1.1. Motivations

To train an NN that models the mapping between the pair of blur/clear images with good generalization performance, a great amount of training data is needed to provide a comprehensive coverage of the instances of different image contents and different blurring effects. In comparison to uniform blur, the variations of blurring effects in non-uniform blur are much more significant, as the spatial configurations of blurring effects can be very different across image regions and over different images. Thus, it is overwhelming to build a training data set that is sufficiently comprehensive to avoid possible overfitting when training the NN. As a result, the performance gain of existing deep-learning-based approaches over the traditional ones is limited, and increasing their model size does not help much for further performance improvement; see *e.g.* the studies in (Xin et al., 2018; Zhang et al., 2019).

It is well known in human visual perception that blur directly participates in visual experience especially for space perception. It is shown in (Khan et al., 2011) that blurring has an important influence on visual attention, and there is deep connection between blur and extraction of salient regions. Indeed, human visual system can directly estimate local blur effects from many salient structures (*e.g.* edges and corner points) and generalize them to more global salient regions. This motivated us to investigate the introduction of the spatial attention mechanism to the NN so that the NN can be learned to effectively exploit salient image features to deal with spatially-varying blurring effects. Also, how to restore image regions with different blurring effects in one NN is another concern that needs to be handled. As different blur effects require different processes for restoration, the NN for processing non-uniformly blurred images needs to be spatially-varying as well. Clearly, spatial attention is one solution to introduce such a spatially-varying nature in the NN, specially for the CNN.

Channel attention is another widely-used attention mechanism in deep NNs for image classification and processing (Hu et al., 2018). Channel attention allows the intermediate features of the CNN have varying weights over different images, and thus improve the adaptivity of the features of the CNN to different image contents. Such an adaptivity is certainly very appealing when the CNN need to handle a wide range of image contents, as well as blurring effects.

In summary, the potential benefits of spatial attention and channel attention in handling non-uniform blur, inspired us to investigate the attention mechanism for deep-learning-based non-uniform blind deblurring.

1.2. Basic Ideas

In this paper, using a multi-scale encoder-decoder CNN as the backbone, we propose a deep attentive NN with built-in attention mechanism for non-uniform blind image deblurring. The attention mechanism we use for the deblurring NN include both spatial attention and channel attention.

In the proposed approach, the spatial attention is introduced in both the encoder and the decoder, and they have different functions. In an encoder-decoder CNN, the encoder functions as a feature extractor for capturing essential image features that provide essential information for image recovery while are robust to the blurring. It is shown in edge-selection-based uniform blind motion deblurring methods (*e.g.* (Cho and Lee, 2009; Xu et al., 2013; Yang and Ji, 2019)) that focusing on strong image edges with different orientations for kernel estimation can provide very robust estimation of blur kernels, which in turn greatly improves the deblurring performance. For instance, strong horizontal/vertical edges will not be erased by blurring, and they provide all information regarding the blurring effect along the vertical/horizontal direction.

According to the success of edge selection techniques in uniform blind deblurring methods, different stages of an effective encoder should discriminatively treat image edges and smooth regions, or say some stages emphasize edges and some emphasize smooth regions. Therefore, we introduce the spatial attention into the encoder part. Such an attention mechanism is expected to allow the encoder to treat different spatial image features with different weights. As a result, the features extracted from the encoder with spatial attention will focus on those image features encoding more information regarding the blurring effect, *e.g.* strong image edges with isotropic orientations. See Fig. 1 (middle row) for an illustration of the spatial attention in the encoder part, which makes the NN focus more on strong image edges with various orientations.

In our approach, spatial attention is also introduced into the decoder part of the encoder-decoder CNN, but with a different function from its counter-part in the encoder part. Recall that the decoder can be interpreted as an image recovery process that maps the extracted features from the encoder to a clear image. In the case of non-uniform blind deblurring, different image regions have different blurring effects. Thus, different image regions should be treated by different reconstruction processes. For example, a region with more severe blurring should be paid more attention to, as more details need to be recovered. A plain version of the deblurring CNN without spatial attention is not effective on modeling such highly location-dependent mappings. The introduction of spatial attention in the decoder enables efficient modeling on location-dependent operations. See Fig. 1 (bottom row) for an illustration of the spatial attention in the decoder part, where the spatial attention distinguishes well the image regions with different blur degrees, *e.g.* focusing more on fast-moving persons.

In addition to spatial attention, the channel attention is also employed for further improvement on the generalizability of the CNN in non-uniform blind deblurring. As the intermediate features from the CNN are supposed to cover a wide range of images with different contents, many features are not very related to one particular image. Such a redundancy in features will cause severe issues in the case of image blurring, as there will exist certain ambiguities among different images when they are severely blurred. The channel attention allows the CNN to impose different weights on different channels (*i.e.* different feature maps), which makes the CNN more adaptive to the input image.

Similar to other existing works (Xin et al., 2018; Zhang et al., 2019), we also implement a multi-scale version of the encoder-decoder CNN as the backbone, and built the aforementioned attention mechanisms into the NN. The multi-scale architecture of the NN provides better guidance for the encoder to extract the blur-invariant representations as well as the decoder to recover image details.

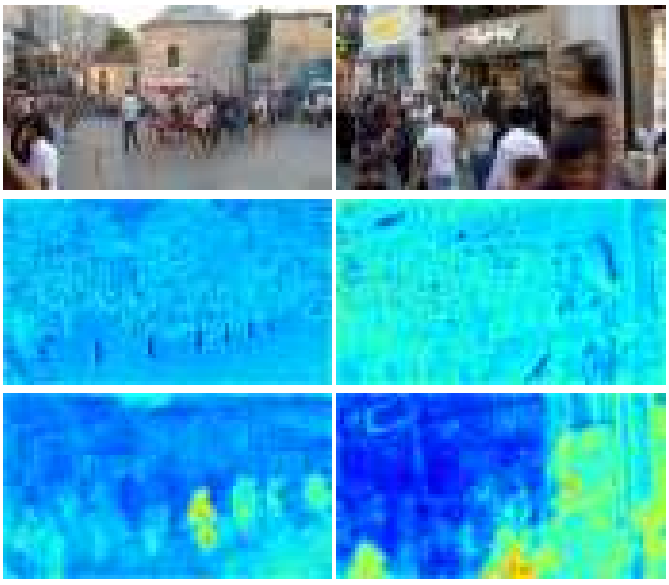

Fig. 1. Images with spatially-varying blurring and their spatial attention maps generated by the proposed attentive CNN. Upper row: input images. Middle row: spatial attention maps generated in the 5^{th} block of encoder. Bottom row: spatial attention maps generated in the 5^{th} block of decoder.

1.3. Main Contributions

The main contribution of this paper is the introduction of attention mechanism into the NN for non-uniform blind image deblurring. One main limitation of the existing deep learning methods for non-uniform blind deblurring lies in their unsatisfactory generalization performance, owing to significant variations among spatially-varying blurring effects. Introducing the attention mechanism has several benefits toward better generalization performance, including *(i)* the spatial attention in the encoder makes the NN focus more on those image features that are closely related to the blurring estimation; *(ii)* the spatial attention in the decoder enables spatially-varying treatments on different image regions, and *(iii)* the channel attention allows image-adaptive deblurring procedures.

Based on the spatial and channel attention mechanisms, this paper present an encoder-decoder CNN with light-weight concurrent spatial and channel attention modules. The proposed CNN can effectively restore the images degraded by complex spatially-varying blurring, with relatively-small model size. The experiments on standard benchmark datasets show that the proposed model achieved the state-of-the-art performance, which have justified the value of the attention mechanism in deep-learning-based non-uniform blind image deblurring.

2. Related work

In the last decade, there have been many approaches proposed for single-image-based blind motion deblurring. Depending on the setting of motion blurring effects, most existing approaches can be classified into three categories: non-blind motion deblurring which assumes the parameters of blur processing are known, blind uniform motion deblurring which assumes the blur is generated by convolving with an unknown kernel, and blind non-uniform motion deblurring which considers complex spatially-varying blurring effects. As the paper aims at non-uniform blind motion deblurring, the following literature review will focus more on the last category.

2.1. Non-Blind Image Motion Deblurring

Non-blind motion deblurring often assumes the blurring effect is uniform and models the blurring process by the convolution with a given low-pass filter. The focus of non-blind deblurring is about designing suitable priors to regularize clear images so as to suppress the magnification of the measurement noise when reversing the convolution process. The often-used image priors include *(i)* the sparsity of image gradients which is often implemented by total variation (TV) minimization (*e.g.* (Chan and Wong, 1998)), ℓ_0/ℓ_1-norm regularization under wavelet frames (*e.g.* (Cai et al., 2009; Bao et al., 2016)) or gradient sparsity (Javaran et al., 2017); and *(ii)* the patch recurrence prior implemented by nonlocal operators (*e.g.* (Danielyan et al., 2011; Quan et al., 2014)). Recently, there are some deep-learning-based approaches (*e.g.* (Zhang et al., 2017; Kruse

et al., 2017)), which unroll some optimization process of certain regularization method and replace the pre-defined prior by the learnable prior modeled by a deep NN. The NNs in these methods act as the denoisers. These deep-learning-based approaches showed better performance than the non-learning methods.

In comparison to the development of non-blind uniform motion deblurring methods, the study on non-blind non-uniform methods has been scant in the literature. One main reason is the non-uniform blurring operator is difficult to simulate or capture. One available work is (Tai et al., 2010). Assuming the projective motion is given, this work extended the traditional Richardson-Lucy algorithm for uniform deblurring to handling non-uniform motion deblurring. In (Whyte et al., 2012), based on the rotational camera motion during exposure, a parameterized geometric model is built up to capture the non-uniform motion blur caused by camera shake. With the parameterized model, the non-blind deblurring is conducted.

2.2. Blind Uniform Motion Deblurring

Blind uniform motion deblurring methods also assume the blurring is uniform and model the blurring by the convolution with an unknown blur kernel; see *e.g.* (Shan et al., 2008; Levin et al., 2009; Sun et al., 2013; Perrone and Favaro, 2014; Ren et al., 2016; Yang and Ji, 2019). In comparison to the non-blind ones, these blind uniform deblurring methods aim at estimating the blur kernel. Once the kernel is determined, the clear image can be restored by calling some non-blind deblurring method. Many existing methods estimate the blur kernel based on selected strong edges. Cho and Lee (Cho and Lee, 2009) proposed to use simple image processing techniques to predict strong edges from an estimated latent image, which are then solely used for kernel estimation. Sun *et al.* (Sun et al., 2013) proposed to estimate the blur kernel and latent image by imposing a patch prior specifically tailored towards modeling the appearances of image edges and corner primitives. Wang *et al.* (Wang et al., 2018) developed an elastic-net regularization of singular values computed from similar patches of an image to guide kernel estimation. Schuler *et al.* (Schuler et al., 2015) proposed to use a CNN to extract local image features for blur kernel estimation. Ren *et al.* (Ren et al., 2016) proposed to retain the dominant edges and eliminate fine texture and slight edges in the intermediate images using a thresholding strategy, allowing robust kernel estimation. Pan *et al.* (Pan et al., 2016) proposed a maximum a posterior framework for moving object deblurring, which jointly estimates object segmentation and camera motion. Yang *et al.* (Yang and Ji, 2019) proposed an adaptive edge selection scheme for blur kernel estimation, which is implemented by a variational probabilistic framework.

2.3. Non-uniform Blind Motion Deblurring

Uniform deblurring methods have their limitations in real applications. There are many approaches proposed for handling spatially-varying motion blur. Levin (Levin, 2007) proposed to segment a motion-blurred image into layers that contain different blur generated from an one-dimensional box filter. The sharp image is recovered by the uniform deblurring at each layer. Ji and Wang (Ji and Wang, 2012) proposed to approximate non-uniform motion blurring by a piece-wise uniform blurring model. Then the image is deblurred by using a robust version of ℓ_1-norm relating regularization method. The above two approaches are based on two-stage frameworks which may be sub-optimal. Kim *et al.* (Kim et al., 2013) proposed a unified framework that jointly conducts blur region segmentation, local blur kernel estimation and sharp image recovery for deblurring images of dynamic scenes.

Recent approaches for blind non-uniform motion deblurring are based on deep learning. Sun *et al.* (Sun et al., 2015) proposed to estimate the heterogeneous motion blur in the form of motion field by a CNN and then deconvolve the blurred image with the estimated motion field. The image is first divided into overlapping patches on which the local blur kernels are learned with a CNN. Then the learned blur kernels are merged into the motion field based on a Markov random process. Gong *et al.* (Gong et al., 2017) proposed to learn a CNN to predict the motion flow directly without post-processing.

The aforementioned deep approaches explicitly model the blurring process with spatially-varying blur kernels. There are also kernel-free deep approaches that learn to end-to-end de-

blurring, *i.e.* learning the mapping from blurry images to the clear ones directly. Nah *et al.* (Nah et al., 2017) proposed a multi-scale CNN with coarse-to-fine structure. A multi-scale loss is used to train the CNN. To enlarge the receptive field of the CNN, they used a large number of convolutional layers with residual connections in each scale, which however increases/decreases the difficulty/efficiency of training. For reducing the number of parameters so as to make the CNN training easier, Tao *et al.* (Xin et al., 2018) imposed the recurrent structure onto the multi-scale encoder-decoder CNN. They also introduced the long short-term memory (LSTM) units to model the dependencies of the intermediate features across scale. Gao *et al.* (Gao et al., 2019) proposed a selective weights sharing mechanism onto the multi-scale CNN. They further extended the skip connections to nested skip connections which encode second-order information for better capturing image features. Zhang *et al.* (Zhang et al., 2018) proposed a spatially-variant recurrent NN for dynamic scene deblurring, where the recurrent structures are inspired by the recursive filtering. Instead of using downsampling to generate multi-scale representations as input, Zhang *et al.* (Zhang et al., 2019) proposed to crop image patches to generate the multi-scale input. At each scale, the cropped patches are processed and then merged as the patch of the upper scale. In (Zhang et al., 2019), an efficient scheme is also proposed for stacking deblurring CNN models for better performance.

There are some approaches using generative adversarial networks (GANs). Kupyn *et al.* (Kupyn et al., 2018) modeled image deblurring as a style transfer problem and use a GAN from image translation for deblurring. Liu *et al.* (Liu et al., 2018) combined a conditional GAN with the NN of (Xin et al., 2018) to enhance the visual quality of the deblurring results. It is worth mentioning that there are some deep approaches specifically designed for text images (*e.g.* Quan et al. (2020)) instead of natural dynamic scenes, which is not considered in our work.

2.4. Attention Mechanism for Image Deblurring

There are some deep learning approaches using attention mechanisms for deblurring. Purohit *et al.* (Purohit and Rajagopalan, 2019) proposed a feature transformation with a non-local spatial attention to deal with motion blurs. The spatial attention is computed based on pair-wise inner products of pixels, which has higher computational cost than ours. Wu *et al.* (Wu et al., 2020) used a dual attention for video deblurring, where an internal attention module is used to select the optimal temporal scales for restoring the sharp center frame, and an external attention module is used to aggregate and refine multiple sharp frame estimates. This method does not employ spatial or channel attention as ours to handle spatially-varying blurs. Shen *et al.* (Shen et al., 2019) proposed an NN that uses a mask to separate a blurry image into background and front objects (*i.e.* humans), with two decoders to recover the sharp results of the background and objects separately. This method requires manual masks to train the NN and essentially assume all the moving objects as the same type, which limits its applications. In contrast, our method has no such requirements and assumptions.

3. Proposed method

3.1. Network Architecture

The proposed CNN for image deblurring is outlined in Fig. 2, whose backbone is an encoder-decoder NN repeated with a multi-scale fashion. Such a backbone is inspired by the work of (Xin et al., 2018; Zhang et al., 2019). Concretely, there are T modules, denoted by $\mathcal{M}_1(\cdot;\Theta), \cdots, \mathcal{M}_T(\cdot;\Theta)$, in the proposed CNN, each of which module is an encoder-decoder network whose weights are shared with other modules. Given a blurry image $\boldsymbol{f}$ as input, we first generate its multi-scale representations, denoted by $\boldsymbol{f}_{\downarrow s_1}, \ldots, \boldsymbol{f}_{\downarrow s_T}$ by downsampling $\boldsymbol{f}$ with scale factors $s_1, \cdots, s_T$ and the bi-linear interpolation, where $s_t = c^{t-1}$ with $c > 1$ for all t. We set $T = 3$ and $c = 2$ in practice. The proposed CNN generates a sequence of deblurred images $\boldsymbol{g}_1, \cdots, \boldsymbol{g}_T$ at different scales by the modules:

$$\mathcal{M}_T(\cdot;\Theta) : [\boldsymbol{f}_{\downarrow s_T}, \boldsymbol{f}_{\downarrow s_T}] \rightarrow \boldsymbol{g}_T, \tag{2}$$

$$\mathcal{M}_t(\cdot;\Theta) : [\boldsymbol{f}_{\downarrow s_t}, \boldsymbol{g}_{t+1}^{\uparrow s_t}] \rightarrow \boldsymbol{g}_t, \tag{3}$$

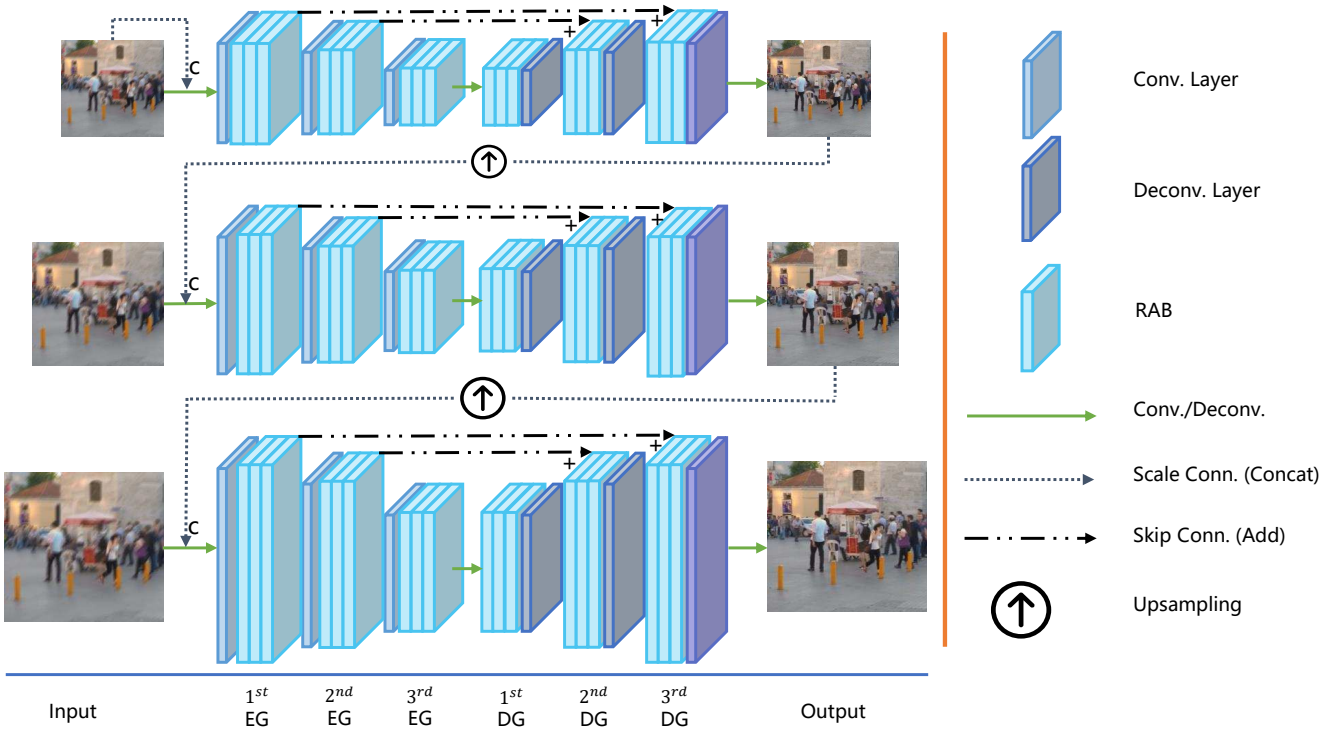

Fig. 2. Diagram of framework of proposed NN. The symbol 'c' indicates concatenation in 'Scale Connection'. The symbol '+' indicates element-wise addition in 'Skip Connection'.

where $\uparrow s_t$ denotes the operation of upsampling to scale s_t with bi-linear interpolation, and $[\cdot,\cdot]$ denotes the concatenation operation. In other words, at each scale, we concatenate the downsampled image of the current scale and the deblurred image from the previous scale as input, and then generate the deblurred image at the current scale as output. The output of $\mathcal{M}_1$ is defined as the final output of the CNN.

All the modules $\mathcal{M}_1, \cdots, \mathcal{M}_T$ share the same structure as well as the same weights. Each module is an encoder-decoder network which consists of an encoder and a decoder. For the convenience of presentation, the encoder/decoder is divided into several groups, denoted by EGs/DGs, as shown in Fig. 2. The EGs and DGs have symmetric structures. The feature maps passing through the $1^{\text{st}}/2^{\text{nd}}/3^{\text{rd}}$ EG are with the same size as those of the $3^{\text{rd}}/2^{\text{nd}}/1^{\text{st}}$ DG. Each EG sequentially connects a convolutional/downsampling layer and three residual attention blocks (RABs), while each DG sequentially connects three RABs and a upsampling/deconvolutional layer. Skip connections are added from the $1^{\text{st}}/2^{\text{nd}}$ EG to the $3^{\text{rd}}/2^{\text{nd}}$ DG.

The RABs in the EGs and DGs have the same structure, which is shown in Fig. 3. An RAB sequentially connects a convolutional (Conv) layer, a rectified linear unit (ReLU), a Conv layer and an attention module, with a skip connection that connects the RAB's input and output by a summation operation. The number of convolution kernels is 32/64/128 on the $1^{\text{st}}/2^{\text{nd}}/3^{\text{rd}}$ RAB as well as on the $3^{\text{rd}}/2^{\text{nd}}/1^{\text{st}}$ RAB. The sizes of the convolution kernels in the two Conv layers in all RABs are set to 5×5. The residual links in the RABs, as well as the skip connections between EGs and DGs, bring two benefits. For the forward pass, it enables the network to reuse the features output by its previous layers for gaining higher visual quality. For the backward pass, it help avoids the gradient vanishing problems and thus allows the NN to be deeper while trainable. The attention mechanisms in RABs are another critical parts in our model, which will be detailed in the next.

3.2. Attention Modules

Ideally, for image deblurring, the encoder acts as a robust image feature extractor that preserves essential image components while eliminating the blurring effects, while the decoder progressively recovers the image details on the output of the en-

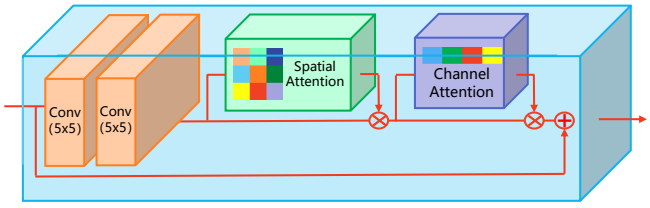

Fig. 3. Diagram of residual attention block.

coder. When the blur is spatially-varying, it is very challenging for the encoder to extract the essential yet robust features, as the regions with very similar contents can have totally different blur degrees and different regions can degenerate to similar ones due to the varying blur. It is also challenging for the decoder to the recover the image details as the degradation is varying on different regions owing to the spatially-varying blur. To better handle such problems, we introduce spatial attention (SA) modules to both encoder part and decoder part. Recall that the role of SA is different in the encoder part and the decoder part. The former is to discriminatively exploit image edges and smooth regions while the latter one is to exploit the spatially-varying blur degrees. However, considering simplicity, we use the same scheme to define them.

Borrowing the idea of squeeze-and-excitation (SE) (Hu et al., 2018) for constructing channel attention, we define the SA module, which is shown in Fig. 4, as follows. Let $\boldsymbol{X} \in \mathbb{R}^{W\times H\times C}$ denote the feature maps of C channels with spatial size $W \times H$. Let $\boldsymbol{p}_{i,j} = \boldsymbol{X}(i, j, :) \in \mathbb{R}^C$ denote vector that collects all features across different channels at the spatial location (i, j). We pass $\boldsymbol{p}_{i,j}$ to a multi-layer perceptron (MLP) to form the spatial attention $\boldsymbol{A}(i, j)$ at the spatial location (i, j), for all (i, j). The MLP sequentially contains a fully-connected (FC) layer, a rectified linear unit (ReLU), a FC layer and a sigmoid function. The sizes of two FC layers are $C \times \frac{C}{16}$ and $\frac{C}{16} \times 1$ respectively. Formally, $\boldsymbol{A}(i, j)$ is generated by

$$\boldsymbol{A}(i, j) = \sigma(\boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{p})), \tag{4}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{C}{16}\times C}, \boldsymbol{W}_2 \in \mathbb{R}^{1\times\frac{C}{16}}$ denote two FC layers and σ denotes the sigmoid function. Due to the use of sigmoid function, all elements in $\boldsymbol{A}(i, j)$ are in $[0, 1]$. The proposed spatial attention can lead to good results in the experiments with few parameters involved.

The channel attention (CA), shown in Fig. 5, is for improving the generality of the network across different images. When handling the images with different types/degrees of blur or different image patterns, it is better to give different contributions to different feature channels. The CA determines the contribution of each feature channel by exploiting the interdependencies among different channels. Let $\boldsymbol{q} = [q_1, \cdots, q_C]$ where q_c is the result of global average pooling on the feature map of the cth channel, *i.e.* calculating the mean value of all elements on $\boldsymbol{X}(:, :, c)$. The vector $\boldsymbol{q}$ encodes the information of each feature map, and it is input to an MLP to predict the importance of each feature map. The MLP has a similar structure with that used by the SA module, which sequentially contains an FC layer, an RelU, an FC layer and a sigmoid function. The sizes of two FC layers are $\frac{C}{16} \times C$ and $C \times \frac{C}{16}$ respectively. The output of the MLP is denoted by $\bar{\boldsymbol{a}} = [\bar{a}_1, \cdots, \bar{a}_C]$ and used as the channel attention map. Formally, $\bar{\boldsymbol{a}}$ is calculated by

$$\bar{\boldsymbol{a}} = \sigma(\boldsymbol{V}_2\mathrm{ReLU}(\boldsymbol{V}_1\boldsymbol{q})), \tag{5}$$

where $\boldsymbol{V}_1 \in \mathbb{R}^{\frac{C}{16}\times C}, \boldsymbol{V}_2 \in \mathbb{R}^{C\times\frac{C}{16}}$ denote the two FC layers.

The combined attention is done by re-calibrating the feature maps using the combination of SA and CA as follows:

$$\hat{\boldsymbol{X}} = \boldsymbol{X} \odot (\boldsymbol{A} \otimes \bar{\boldsymbol{a}}), \tag{6}$$

where $\otimes$ denotes the Kronecker product and $\odot$ denotes the element-wise product.

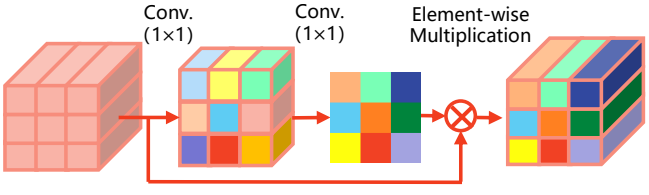

Fig. 4. Diagram of spatial attention module.

3.3. Loss

Let ∇ denote the gradient operator. In training, given the blurry/clear image pair set $\{(\boldsymbol{g}^k, \boldsymbol{f}^k)\}_{k=1}^K$, we optimize the fol-

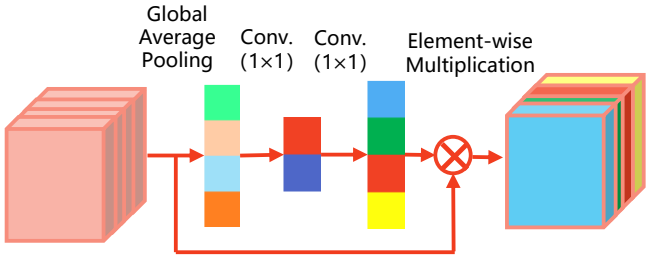

Fig. 5. Diagram of channel attention module.

lowing loss function to train our network:

$$\min_{\Theta} \mathcal{L}(\Theta) := \mathcal{L}_{\text{mse}}(\Theta) + \lambda \mathcal{L}_{\text{grad}}(\Theta), \tag{7}$$

where $\mathcal{L}_{\text{mse}}$, $\mathcal{L}_{\text{grad}}$ the multi-scale mean square error (MSE) and gradient loss respectively:

$$\mathcal{L}_{\text{mse}}(\Theta) := \sum_{k=1,t=1}^{K,T} \|\mathcal{M}_t(\boldsymbol{g}^k_{\downarrow s_t}; \Theta) - \boldsymbol{f}^k_{\downarrow s_t}\|_2^2, \tag{8}$$

$$\mathcal{L}_{\text{grad}}(\Theta) := \sum_{k=1}^{K} \|\nabla\mathcal{M}_1(\boldsymbol{g}^k_{\downarrow s_1}; \Theta) - \nabla\boldsymbol{f}^k_{\downarrow s_1}\|_2^2. \tag{9}$$

Following the previous work (Xin et al., 2018; Zhang et al., 2019), we use the multi-scale MSE to measure the loss of the model at each scale level. Regarding the loss function, we adopted a simple gradient loss on the final output of the network to train the network. The gradient loss makes the model pay more attention on the edges of output images, as image edges are the main focus of deblurring and more sensitive to blurring than the smooth regions. As a result, such a loss encourages the model to recover images with sharper edges. It is noted that more sophisticated loss functions, such as the perceptual loss and adversarial loss used in (Nah et al., 2017; Liu et al., 2018) might also work for the similar purpose.

4. Experiments

4.1. Datasets and Configurations

The proposed approach is evaluated on three public benchmark datasets for blind image deblurring, including the GoPro dataset (Nah et al., 2017), the VideoDeblurring dataset (Su et al., 2017) and the Köhler dataset (Köhler et al., 2012). The details of these three datasets are as follows:

- The GoPro dataset (Nah et al., 2017) contains 3214 blurry/sharp image pairs of resolution 720×1280, which are extracted from 33 videos captured by the GoPro Hero 4 Black Camera. The blurred images are generated by averaging seven to thirteen successive latent frames to simulate complex camera shakes and complex object motions. Same as the protocol of (Xin et al., 2018), 2103 image pairs are used for model training and the remaining 1111 pairs are used for test. The performance is measured by the PSNR (Peak-Signal-to-Noise Ratio) and SSIM (Structural Similarity).

- The Köhler dataset (Köhler et al., 2012) contains 4 clear images as the ground truths, each of which has 12 different blurry versions. The blurry versions are generated by replaying the recorded 6D camera motions with linear CRF (Camera Response Function) assumed. Same as (Xin et al., 2018), we train our model on the GoPro dataset and test the trained model on all images in the Köhler dataset. Following the standard protocol, the PSNR and MS-SSIM (Multi-Scale SSIM) are used as the quantitative metrics.

- The VideoDeblurring dataset (Su et al., 2017) contains videos captured by various devices (*e.g.* iPhone, GoPro and Nexus). The dataset contains 71 videos, each of which consists of 100 frames of resolution 720×1280. Following (Zhang et al., 2019), two schemes are used for the evaluation: *(i)* training on 61 videos and testing on the remaining ten videos; and *(ii)* training on GoPro's training set and testing on the ten videos. During training, each video is used as an image. In test, the videos are processed frame by frame. The PSNR is used as the quantitative metric.

In addition to the three datasets, we also use some real degraded images to evaluate the generalizability of the proposed approach to real scenarios.

4.2. Implementation Details

Our approach is implemented using TensorFlow and run on a PC with Intel Core i7-6700K CPU and an NVIDIA Titan V GPU. To train our model, we used the Adam solver (Kingma

and Ba, 2014) with default parameters (*i.e.* $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$). The learning rate is initialized to $1e^{-4}$ and exponentially decayed with power 0.3. Totally 4000 epochs are used for training. At each epoch, a batch of 16 blurry/sharp image pairs are generated by randomly cropping 256×256-pixel patches from the training set. In addition, the data augmentation is used by left-right flipping and up-down flipping the training images. During training, each image is normalized to the range $[0, 1]$, and all the trainable variables are initialized by using Xavier (Glorot and Bengio, 2010). The hyper-parameter setting described above is consistent throughout all experiments. The parameter λ is empirically set by the following rule, We set $\lambda = 1$ first to train the model. Then the values of the two terms are calculated, and their ratio which is approximately 2.5, is used to set the λ. All our codes will be available at our website.

4.3. Results and Comparisons

Several state-of-the-art image deblurring approaches are selected for performance comparison, including WFA (Delbracio and Sapiro, 2015), Sun *et al.* (Sun et al., 2015), Nah *et al.* (Nah et al., 2017), Zhang *et al.*(Zhang et al., 2018), Liu *et al.* (Liu et al., 2018), SRN (Xin et al., 2018), and DMPHN (Zhang et al., 2019). Note that there are several models in DMPHN (Zhang et al., 2019), we choose the DMPHN(1-2-4-8) which has almost the same parameter numbers with ours for fair comparison. We also list the results of Stack(4)-DMPHN, the largest model of DMPHN, for comparison. There are also two versions of Sun *et al.* (Sun et al., 2015) and we choose its "single frame" version which works on single image. Regarding the experimental results of these methods, whenever possible, we directly quote the results reported in the literature. Otherwise, we use the pretrained models from the authors to generate the results. If only code is available, we made the effort on adjusting the parameters for optimal performance on test data. If none is available, we leave it blank.

4.3.1. Results on GoPro dataset

The quantitative results on the GoPro dataset and are listed in Table 1. It can be seen that the proposed approach outperforms other compared methods in terms of both PSNR and SSIM, and the PSNR improvement over the SRN (Xin et al., 2018) is about 1.1dB. Such noticeable performance improvement has demonstrated the effectiveness of the proposed approach. We also list the model size and running time of all compared methods in Table 1. It can be seen that our proposed CNN performs the best with a relative lighter model size and less running time.

We show some deblurring results in Fig. 6 for visual inspection. It can be seen that the deblurred images output by our model have the best visual quality. Whyte *et al.*'s method (Whyte et al., 2012) yields unsatisfactory results. Sun *et al.* (Sun et al., 2015)'s methods does not work well on most images. The results of Tao *et al.* (Xin et al., 2018) are better than the previous two, but still with some artifacts as well as blurry edges and unclear objects. Note that Tao *et al.* (Xin et al., 2018) uses a multi-scale CNN similar to ours but without the attention mechanisms. In comparison, benefiting the use spatial/channel attention in both the encoder and decoder, our results are of the highest visual quality with sharper edges and clearer objects.

Table 1. Quantitative results on GoPro test set. The best results are boldfaced and the second best results are underlined.

Model	PSNR (dB)	SSIM	Model Size (MB)	Running Time
Sun *et al.*(Sun et al., 2015)	24.64	0.8429	54.10	20min
Nah *et al.*(Nah et al., 2017)	29.23	0.9162	303.60	3.09s
Zhang *et al.*(Zhang et al., 2018)	29.19	0.9306	37.10	1.4s
SRN (Xin et al., 2018)	30.10	0.9323	33.60	1.87s
Liu *et al.*(Liu et al., 2018)	30.28	-	33.60	1.87s
Gao *et al.*(Gao et al., 2019)	30.92	0.9421	**2.84**	2.3s
DMPHN (Zhang et al., 2019)	30.25	0.9351	29.01	**0.032s**
DMPHN(4) (Zhang et al., 2019)	<u>31.20</u>	<u>0.9453</u>	86.8	0.57s
Ours	**31.23**	**0.9455**	<u>26.34</u>	<u>0.28s</u>

4.3.2. Results on Köhler dataset

Table 2 summarizes the quantitative results of different methods on the Köhler dataset. Recall that the tested models are trained on the GoPro dataset. Thus, the test on the Köhler dataset can evaluate whether a model can generalize well across different datasets. It can be seen that our model again outperforms other compared ones by a considerable margin (around 0.9dB over the second best), which double confirms the effec-

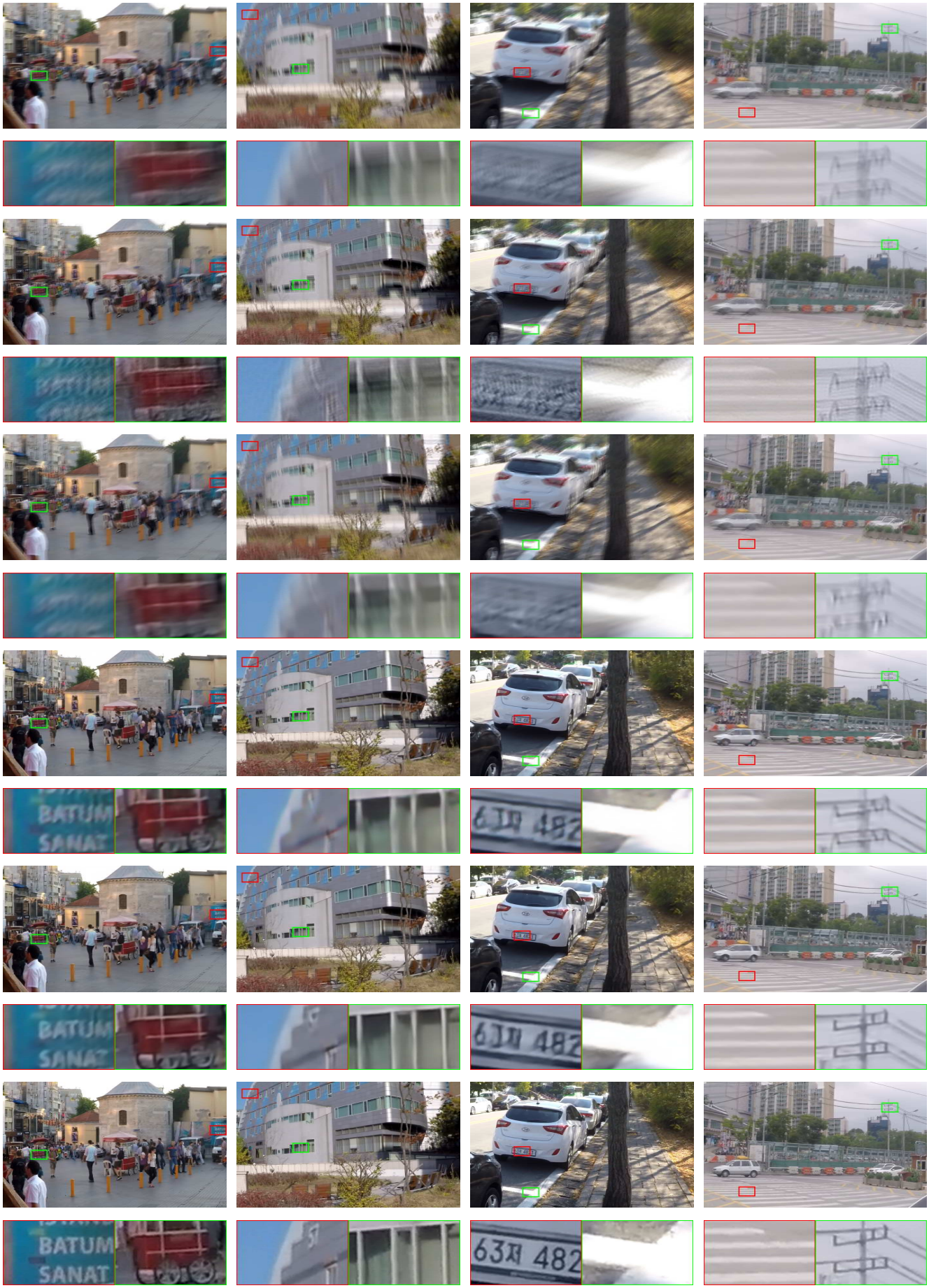

Fig. 6. Visual comparisons of different methods on some blurred images from the test datasets. From top to bottom: Input, Whyte *et al.* (Whyte et al., 2012), Sun *et al.* (Sun et al., 2015), SRN (Xin et al., 2018) , ours and groundtruth.

tiveness and generalizability of the proposed approach. The better generalizability of our model comes from the use of the attention mechanisms which make the model more adaptive to the spatially-varying blur and image structures of a test image.

Table 2. Quantitative results on Köhler test set. The best results are bold-faced and the second best results are underlined.

Model	PSNR(dB)	MS-SSIM
Kim *et al.*(Kim et al., 2013)	24.68	0.7937
Sun *et al.*(Sun et al., 2015)	25.22	0.7735
Nah *et al.*(Nah et al., 2017)	26.48	0.8079
SRN (Xin et al., 2018)	<u>26.75</u>	<u>0.8370</u>
DMPHN (Zhang et al., 2019)	24.66	0.7641
Ours	**27.65**	**0.8596**

4.3.3. Results on VideoDeblurring dataset

The results on the VideoDeblurring dataset are listed in Table 3, where both individual results and the overall results are given. Recall that there are two training setting on this dataset, by which both the generalizability within the dataset and that across different datasets can be evaluated. It can be seen from Table 3 that in both settings, our model exhibit superior performance to the compared methods. Furthermore, our model performs consistently better across all the test images. Such results have demonstrated both the effectiveness and stability of the proposed approach. Note that our model works well on videos even that it processes the video frame by frame without utilizing the temporal cues. Thus, the results also suggest the potential of the extension of the proposed approach to video deblurring.

4.3.4. Results on real images

The blurring effects in the above test data are synthesized based on specific cameras, which may still differ from the real blurry images taken from conventional cameras. Therefore, we also test on some real blurry images obtained from the Internet without ground-truths. The model trained on the GOPRO dataset is used. Please see Fig. 7 for the visual comparison on some real degraded images. It can be seen that our model generalizes well on these real images. The deblurred images of our method have the highest visual quality and contain fewer artifacts than other compared methods. For instance, our model can recover sharp edges of the flowers (the 3^{rd} row) well, while the results of other approaches are more blurry with detail loss. It can be seen that our method as well as others does not perform well on the zoomed-in text region in the last row. The reason is probably that the training set (*i.e.* GoPro dataset) is mainly on outdoor scenes and contains little text content, which limits the generalizability of the trained model in handling blurry text. See more failure cases in Section 4.6.

4.4. More Analysis on Spatial Attention

In order to verify the different roles of SA layers in encoder and decoder, we visualize the SA maps in different blocks in Fig. 8. The first row of Fig. 8 shows the input images with complex blur due to camera motion or object motion. The second row gives the corresponding deblurring results of our method. The 3^{rd} to 5^{th} rows are the spatial attention maps from encoder part. It can be observed that in the encoder part, the SA maps mainly focus on the edges of the input image (3^{st} row and 4^{nd} row) and gradually turn to focus on some smooth areas at the end of the encoder part (5^{rd} row). Such observations indicate that there exists high correlation between the estimated SA map and the edges/smooth areas of input image in the encoder part. Such a characteristic of SA helps the model for better feature extraction in the encoder part. The 6^{th} to 8^{th} rows are the SA maps from the decoder part. It can be seen that the SA maps in decoder firstly focus more on the less-blurry areas such as still buildings and peoples (6^{th} row). At the deeper layers of decoder, the SA tends to pay more attention on the dominant motion blurred areas (7^{th} row), *e.g.* the moving persons. Such observations indicate the high relations between the estimated SA maps and the dominant motion blurred regions. It can be considered that the SA in decoder helps to guide the model to distinguish blurry regions from non-blurred regions during the recovery process in decoder.

4.5. Ablation Study

4.5.1. Ablation study on cross-scale weight sharing

We verified the effectiveness of sharing weights across different scales by examining the performance influence caused

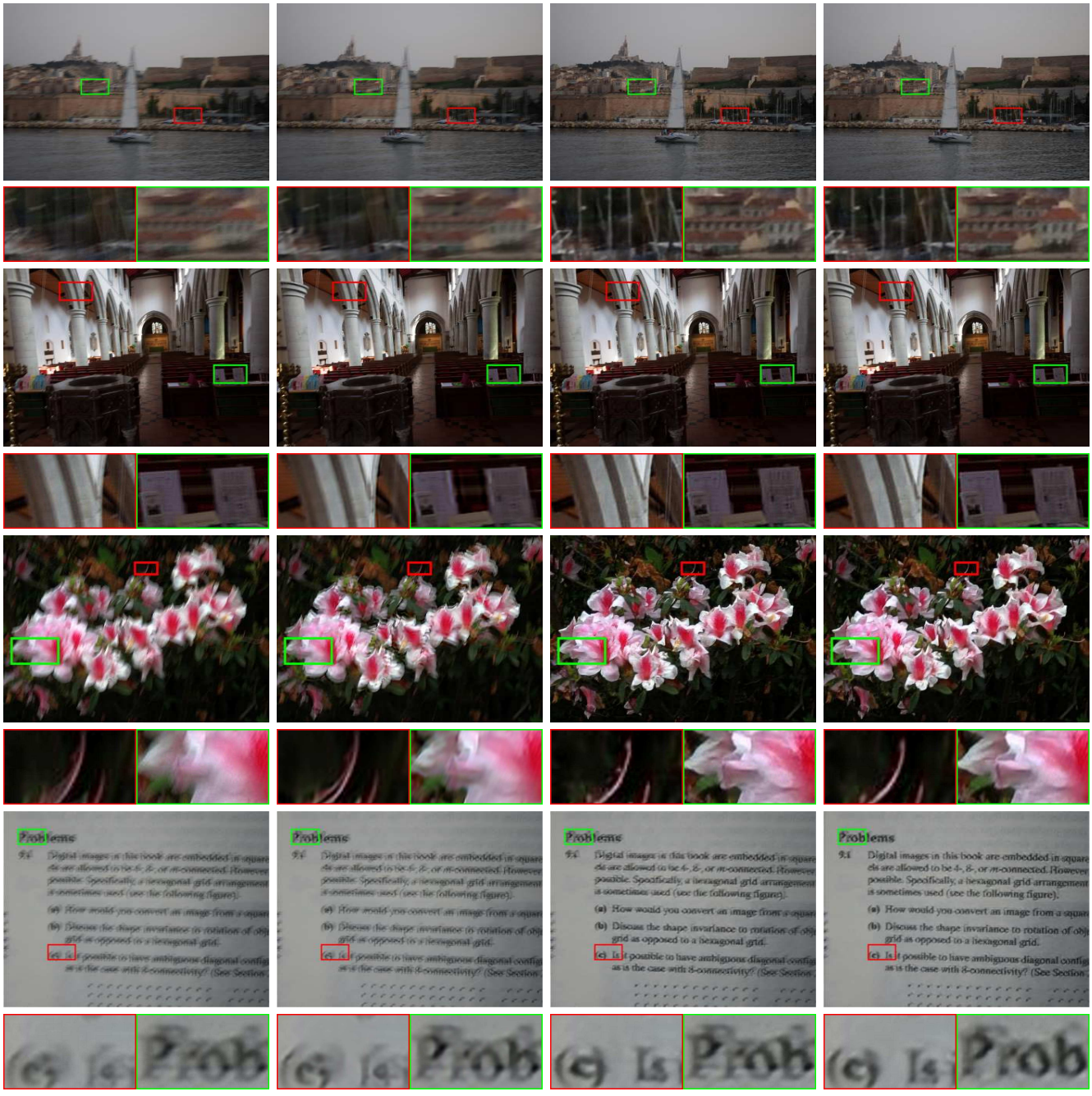

Fig. 7. Visual comparison of different methods on real blurred images, from left to right are the blurry input, results of Sun *et al.* (Sun et al., 2015), results of SRN (Xin et al., 2018) and Ours.

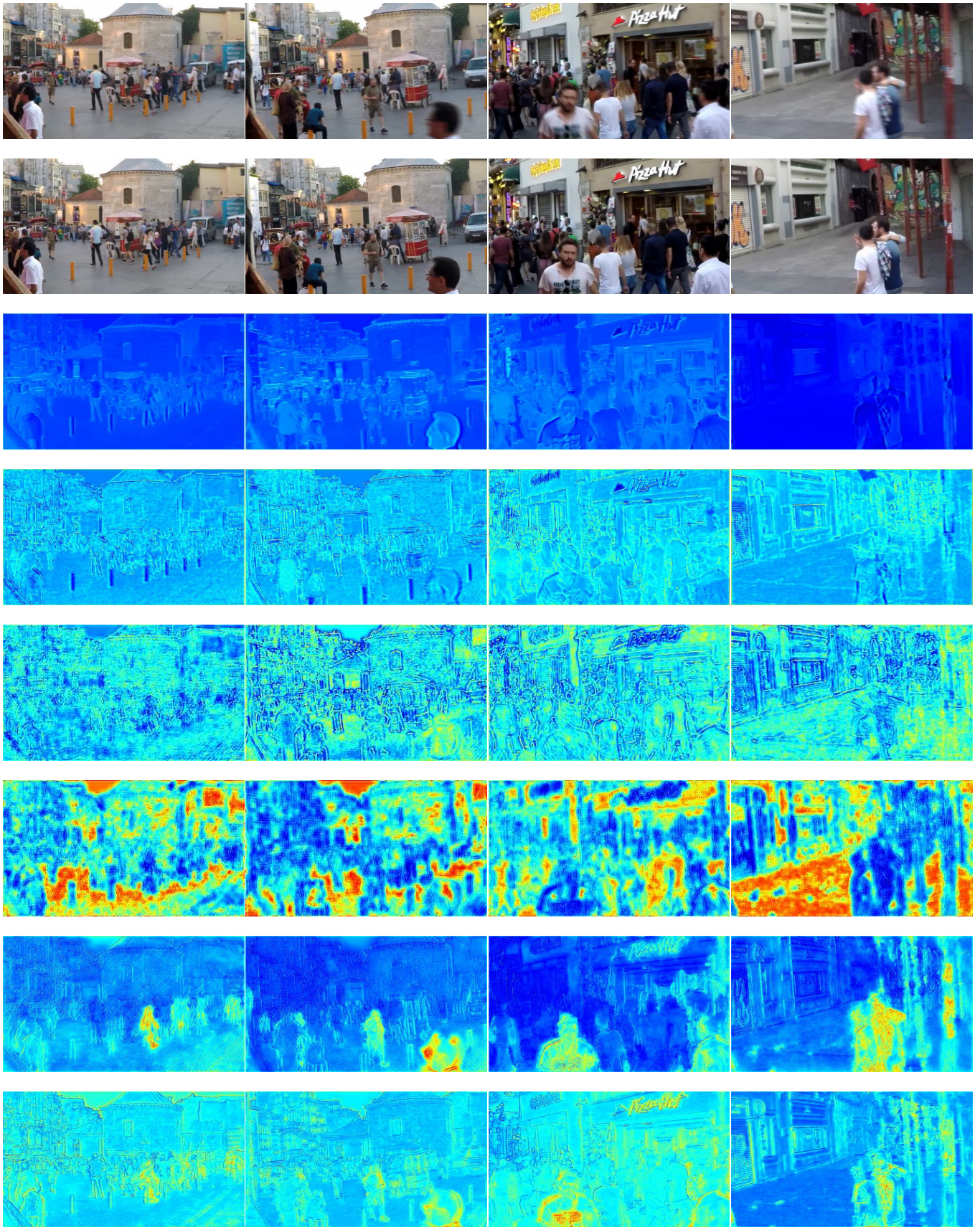

Fig. 8. Visualization of spatial attention maps. Red color indicates higher values in attention maps while the blue color indicates lower values. The 1^{st} and 2^{nd} rows are the input images and the corresponding deblurring results of our method. The 3^{rd} to 5^{th} rows are the spatial attention maps from the 2^{nd}, 5^{th} and 8^{th} encoder block respectively. The 6^{th} to 8^{th} rows are the spatial attention maps from 2^{nd}, 5^{th} and 8^{th} decoder block respectively.

Table 3. PSNR results (dB) on the ten test videos of VideoDeblurring dataset. In the names of methods, no '+' indicates that the results are obtained by the model trained on the GoPro dataset, and '+' indicates that the results are obtained by the model trained on the VideoDeblurring training set. The best results are boldfaced and the second best results are underlined. Note that WFA is a learning-free method and thus the results of WFA and WFA+ are the same.

.

Model	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average
Input	24.14	30.52	28.38	27.31	22.60	29.31	27.74	23.86	30.59	26.98	27.14
WFA (Delbracio and Sapiro, 2015)	25.89	32.33	28.97	28.36	23.99	31.09	28.58	24.78	31.30	28.20	28.35
Su *et al.*(Su et al., 2017)	24.95	30.75	29.05	28.12	22.97	29.73	28.61	25.35	31.14	27.56	27.82
SRN (Xin et al., 2018)	28.37	32.42	30.03	30.78	25.24	31.10	29.22	25.82	33.21	28.57	29.48
DMPHN (Zhang et al., 2019)	25.91	30.71	28.95	28.84	23.28	30.13	28.32	24.70	32.42	27.76	28.10
Ours	**28.80**	**32.92**	**30.90**	**31.06**	**25.47**	**31.55**	**29.61**	**26.23**	**33.58**	**29.39**	**29.95**
WFA+ (Delbracio and Sapiro, 2015)	25.89	32.33	28.97	28.36	23.99	31.09	28.58	24.78	31.30	28.20	28.35
Su *et al.*+ (Su et al., 2017)	25.75	31.15	29.30	28.38	23.63	30.70	29.23	25.62	31.92	28.06	28.37
SRN+ (Xin et al., 2018)	29.07	33.39	30.86	31.07	25.33	32.11	29.86	26.71	34.14	29.76	30.23
DMPHN+ (Zhang et al., 2019)	29.89	33.35	31.82	31.32	26.35	32.49	30.51	27.11	34.77	30.02	30.76
Ours+	**30.47**	**34.16**	**32.21**	**32.05**	**26.43**	**32.71**	**30.53**	**27.42**	**35.28**	**30.63**	**31.19**

by setting the parameter weights non-shared. The results of using weight sharing and non-sharing weights are listed in Table 4. It can be seen that sharing weights across different scales leads to better performance. The reason is probably that sharing weights across different scales can make each scale aim to solve the same problem; otherwise, using different weights may introduce instability and cause the extra problems of un-restrictive solution space. Moreover, without weight sharing, the solution may over-fit to a specific image resolution or motion scale (Xin et al., 2018).

Table 4. Ablation study on weights sharing across different scales. Superior results are boldfaced.

weights sharing	PSNR (dB)	SSIM
×	30.86	0.9413
√	**31.23**	**0.9455**

4.5.2. Ablation study on attention mechanisms

To verify the necessity of the attention mechanisms, we formed different ablated versions of our model by removing some of the attention modules:

- 'w/o all': removing all attention modules;
- 'w/ CA': removing all except the CA modules;
- 'w/ SA(E)': only keeping SA module in encoder part;
- 'w/ SA(D)': only keeping SA module in decoder part;
- 'w/ SA': removing all except the SA modules.
- 'w/ (SA+CA)': the proposed model with both CA and SA modules.

The results in PSNR value of the ablated versions on the GoPro dataset are listed in Table 5, where 'w/ (SA+CA)' refers to the proposed model with all attention modules. It can be seen that spatial attention or channel attention alone brings noticeable performance improvement, as both 'w/ SA' and 'w/ CA' has around 0.9dB advantage over 'w/o all' in PSNR value. While both spatial attention and channel attention can lead to noticeable performance gain alone, their combination leads to further improvement. However, the benefit of the combination of both attentions does not double the performance gain of each individual attention. The further performance gain over each individual attention, spatial attention or channel attention, is around 0.22dB. One plausible cause of such a minor performance improvement when combining both attention mechanism is that there exist redundancy between the SA and CA modules in terms of their functions. As a result, the combination of both modules does not provide significantly further improvement over individual module.

In addition, the ablation studies showed how the SA in encoder and that in decoder contribute to the performance improvement; see the results with respect to 'w /SA(E)' and 'w /SA(D)' in the table. It shows that separate treatment on

different spatial locations and different channels is useful for blind motion deblurring, and spatially-varying treatment plays a more important role than channel-varying treatment in handling spatially-varying blurring. See also Fig. 9 for the deblurring results of 'w/o all' and 'w/ (SA+CA), where the deblurred result obtained by the method with both attention mechanism is sharper than that result of 'w/o all'.

Table 5. Ablation study on attention mechanisms. Best results are boldfaced.

Model	CA	SA (encoder)	SA (decoder)	PSNR (dB)
w/o all	×	×	×	30.09
w/ CA	√	×	×	30.95
w/ SA(E)	×	√	×	30.73
w/ SA(D)	×	×	√	30.82
w/ SA	×	√	√	31.01
w/(SA+CA)	√	√	√	**31.23**

Fig. 9. Comparison of visual results of our model w/ and w/o attention.

4.5.3. Ablation study on combination configurations of attention modules

We formed different combination configurations of attention modules to verify the impact of different combination configurations of attention modules:

- 'parallel-concat': parallel concatenation combination of SA and CA modules;
- 'parallel-addition': parallel addition combination of SA and CA modules;
- 'serial-cascade': serial cascade combination of SA and CA modules.

The PSNR results of different combination configurations on the GoPro dataset are presented at Table 6. It can be seen that the serial cascade combination of SA and CA modules performs best.

Table 6. Ablation study on different combination configurations of attention modules. Best results are boldfaced.

Model	PSNR (dB)	SSIM
parallel-concat	31.14	0.9449
parallel-addition	31.11	0.9446
serial-cascade	**31.23**	**0.9455**

4.5.4. Ablation study on loss function

We verified the effectiveness of each component of the total loss in training our CNN by *(i)* replacing the multi-scale MSE loss with the finest-scale MSE loss (*i.e.* single-scale MSE); *(ii)* removing the gradient loss from the total loss function. The results are listed in Table 7. It can be seen that both multi-scale MSE loss and gradient loss are useful in our CNN training.

Table 7. Ablation study on loss function. Best results are boldfaced.

MSE loss	gradient loss	PSNR (dB)	SSIM
single-scale	×	30.53	0.9386
single-scale	√	30.71	0.9409
multi-scale	×	31.05	0.9432
multi-scale	√	**31.23**	**0.9455**

4.6. Study of the cases that challenge the proposed method

To have a more complete picture of the proposed method, we select several representative unsatisfactory cases from test datasets, and compare the results from our methods to that from other methods. See Fig. 10 for a visual inspection. It can be seen that, our method does not perform well on the image with occlusion of similar objects shown in the first row, and the image with dominant text content shown in the second row.

The unsatisfactory performance on handling occlusion is not supervising as the linear model used for modeling non-uniform

Input (Whyte et al., 2012) (Sun et al., 2015) SRN (Xin et al., 2018) Ours Ground-Truth

Fig. 10. Illustration of some unsatisfactory results from the proposed method and the comparison to other methods.

blurring does not take occlusions into consideration. Indeed, how to effectively handle occlusions in blind motion deblurring remains an open question. Regarding the ineffectiveness on processing texts. One reason is that the training dataset do not contain images with intensive text contents. As a result, the model trained on such training dataset cannot effectively process text. Indeed, the recovery of text images have their specifically designed deep learning solutions (*e.g.* Quan et al. (2020)), which is trained over a dataset with text images.

We would like to point out that while our method did not perform well on these images, other existing methods do not perform well either. Indeed, in comparison, our method still output comparably with or relatively better results than other compared ones. The images presented in Fig. 10 raise great challenges to existing blind non-uniform deblurring methods, and we will investigate how to handle these cases in our future study on blind non-uniform deblurring.

5. Summary

In this paper, we tackle the challenging single-image blind deblurring problem based using a multi-scale residual CNN with spatial attention and channel attention. One big challenge is the large variations from the spatially-varying blurring effects across image regions and over different images, which makes a deep NN hard to generalize well. This paper demonstrated that introducing the spatial and channel attention mechanisms can improve the generalizability and performance a deep neural network in blind image deblurring. Our model exhibited state-of-the-art performance with relatively-small model size. While our method is applied to deblurring dynamic scenes, it can be also applied to other non-uniform blur setting, *e.g.* out-of-focus, which is one of our future work.

We also showed that the different roles played by the spatial attention in the encoder part and that in the decoder part. The former is for discriminative exploitation of image edges and smooth regions, while the latter is for discriminative treatment on different regions with different blurring effects. We believe such results can benefit the better understanding on the attention mechanism for blind image deblurring and motivate the studies on new attention mechanisms for removing spatially-varying blur. One possible improvement on the attention mechanism comes from that the overall attention in our model is formed by the tensor product of spatial attention and channel attention, which assumes independence between them. In the future, we would like to investigate new mechanisms that consider region-varying channel attention.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grants 61872151 and 62072188, in part by Natural Science Foundation of Guangdong Province under Grants 2017A030313376 and 2020A1515011128, in part by Science and Technology Program of Guangdong Province under Grant 2019A050510010, in part by Science and Technology Program of Guangzhou under Grant 201802010055, and in part by Singapore MOE AcRF under Grant MOE2017-T2-2-156.

References

Bao, C., Dong, B., Hou, L., Shen, Z., Zhang, X., Zhang, X., 2016. Image restoration by minimizing zero norm of wavelet frame coefficients. Inverse Problems 32, 115004.

Caglioti, V., Giusti, A., 2009. Recovering ball motion from a single motion-blurred image. Computer Vision and Image Understanding 113, 590–597.

Cai, J.F., Osher, S., Shen, Z., 2009. Split bregman methods and frame based image restoration. Multiscale modeling & simulation 8, 337–369.

Chan, S.H., Nguyen, T.Q., 2011. Single image spatially variant out-of-focus blur removal, in: IEEE Int. Conf. Image Proces., IEEE. pp. 677–680.

Chan, T.F., Wong, C.K., 1998. Total variation blind deconvolution. IEEE Trans. Image Process. 7, 370–375.

Cho, S., Lee, S., 2009. Fast motion deblurring. ACM Trans. Graphics 28, 145.

Danielyan, A., Katkovnik, V., Egiazarian, K., 2011. Bm3d frames and variational image deblurring. IEEE Trans. Image Process. 21, 1715–1728.

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

Denis, L., Thiébaut, r., Soulez, F., Jean-Marie, B., Mourya, R., 2015. Fast approximations of shift-variant blur. Int. J. Comput. Vision 115. doi:10.1007/s11263-015-0817-x.

Gao, H., Tao, X., Shen, X., Jia, J., 2019. Dynamic scene deblurring with parameter selective sharing and nested skip connections, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 3848–3856.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proc. Int. Conf. artificial intelligence and statistics, pp. 249–256.

Gong, D., Yang, J., Liu, L., Zhang, Y., Reid, I., Shen, C., Van Den Hengel, A., Shi, Q., 2017. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 2319–2328.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 7132–7141.

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

Ji, H., Wang, K., 2012. A two-stage approach to blind spatially-varying motion deblurring, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, IEEE. pp. 73–80.

Khan, R.A., Dinet, E., Konik, H., 2011. Visual attention: effects of blur, in: Proc. Int. Conf. Image Proces., IEEE. pp. 3289–3292.

Kim, T.H., Ahn, B., Lee, K.M., 2013. Dynamic scene deblurring, in: Proc. IEEE Int. Conf. Comput. Vision.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Köhler, R., Hirsch, M., Mohler, B., Schölkopf, B., Harmeling, S., 2012. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database, in: Proc. European Conf. Comput. Vision, Springer. pp. 27–40.

Kruse, J., Rother, C., Schmidt, U., 2017. Learning to push the limits of efficient fft-based image deconvolution, in: Proc. IEEE Int. Conf. Comput. Vision, IEEE. pp. 4586–4594.

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J., 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 8183–8192.

Levin, A., 2007. Blind motion deblurring using image statistics, in: Proc. Advances Neural Info. Process. Syst., pp. 841–848.

Levin, A., Weiss, Y., Durand, F., Freeman, W.T., 2009. Understanding and evaluating blind deconvolution algorithms, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, IEEE. pp. 1964–1971.

Liu, J., Sun, W., Li, M., 2018. Recurrent conditional generative adversarial network for image deblurring. IEEE Access 7, 6186–6193.

Nah, S., Hyun Kim, T., Mu Lee, K., 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 3883–3891.

Nimisha, T.M., Kumar Singh, A., Rajagopalan, A.N., 2017. Blur-invariant deep learning for blind-deblurring, in: Proc. IEEE Int. Conf. Comput. Vision, pp. 4752–4760.

Pan, J., Hu, Z., Su, Z., Lee, H.Y., Yang, M.H., 2016. Soft-segmentation guided object motion deblurring, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 459–468.

Perrone, D., Favaro, P., 2014. Total variation blind deconvolution: The devil is in the details, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 2909–2916.

Purohit, K., Rajagopalan, A., 2019. Efficient motion deblurring with feature transformation and spatial attention, IEEE. pp. 4674–4678.

Quan, Y., Ji, H., Shen, Z., 2014. Data-driven multi-scale non-local wavelet frame construction and image recovery. Journal of Scientific Computing 63. doi:10.1007/s10915-014-9893-2.

Quan, Y., Yang, J., Chen, Y., Xu, Y., Ji, H., 2020. Collaborative deep learning for super-resolving blurry text images. IEEE Trans. Comput. Imaging PP, 1–1. doi:10.1109/TCI.2020.2981758.

Ren, W., Cao, X., Pan, J., Guo, X., Zuo, W., Yang, M.H., 2016. Image deblurring via enhanced low-rank prior. IEEE Trans. Image Process. 25, 3426–3437.

Schuler, C.J., Hirsch, M., Harmeling, S., Schölkopf, B., 2015. Learning to deblur. IEEE Trans. Pattern Anal. Mach. Intell. 38, 1439–1451.

Seibold, C., Hilsmann, A., Eisert, P., 2017. Model-based motion blur estimation for the improvement of motion tracking. Computer Vision and Image Understanding 160, 45–56.

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

Sharma, M., Verma, A., Vig, L., 2018. Learning to clean: A gan perspective, in: Asian Conf. Comput. Vision, Springer. pp. 174–185.

Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L., 2019. Human-aware motion deblurring, pp. 5572–5581.

Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O., 2017.

Deep video deblurring for hand-held cameras, in: Proc. IEEE Conf. Comput. Vision 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: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 769–777.

Sun, L., Cho, S., Wang, J., Hays, J., 2013. Edge-based blur kernel estimation using patch priors, in: Proc. IEEE Int. Conf. Comput. Photography, IEEE. pp. 1–8.

Tai, Y.W., Tan, P., Brown, M.S., 2010. Richardson-lucy deblurring for scenes under a projective motion path. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 1603–1618.

Wang, H., Pan, J., Su, Z., Liang, S., 2018. Blind image deblurring using elastic-net based rank prior. Computer Vision and Image Understanding 168, 157–171.

Whyte, O., Sivic, J., Zisserman, A., Ponce, J., 2012. Non-uniform deblurring for shaken images. Int. J. Comput. Vision 98, 168–186.

Wu, J., Yu, X., Liu, D., Chandraker, M., Wang, Z., 2020. David: Dual-attentional video deblurring, pp. 2376–2385.

Xin, T., Gao, H., Yi, W., Shen, X., Wang, J., Jia, J., 2018. Scale-recurrent network for deep image deblurring. Proc. IEEE Conf. Comput. Vision Pattern Recognition .

Xu, L., Jia, J., 2010. Two-phase kernel estimation for robust motion deblurring, in: Proc. European Conf. Comput. Vision, Springer. pp. 157–170.

Xu, L., Zheng, S., Jia, J., 2013. Unnatural l0 sparse representation for natural image deblurring, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 1107–1114.

Yang, L., Ji, H., 2019. A variational em framework with adaptive edge selection for blind motion deblurring, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 10167–10176.

Zhang, H., Dai, Y., Li, H., Koniusz, P., 2019. Deep stacked hierarchical multi-patch network for image deblurring, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 5978–5986.

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: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 2521–2529.

Zhang, K., Zuo, W., Gu, S., Zhang, L., 2017. Learning deep cnn denoiser prior for image restoration, in: Proc. IEEE Conf. Comput. Vision Pattern Recognition, pp. 3929–3938.