Efficient Dynamic Scene Deblurring Using Spatially Variant Deconvolution Network with Optical Flow Guided Training

Yuan Yuan, Wei Su, Dandan Ma*
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an, Shaanxi, P. R. China
{y.yuan1.ieee, npusuwei, madandanhello}@gmail.com

Abstract

In order to remove the non-uniform blur of images captured from dynamic scenes, many deep learning based methods design deep networks for large receptive fields and strong fitting capabilities, or use multi-scale strategy to deblur image on different scales gradually. Restricted by the fixed structures and parameters, these methods are always huge in model size to handle complex blurs. In this paper, we start from the deblurring deconvolution operation, then design an effective and real-time deblurring network. The main contributions are three folded, 1) we construct a spatially variant deconvolution network using modulated deformable convolutions, which can adjust receptive fields adaptively according to the blur features. 2) our analysis shows the sampling points of deformable convolution can be used to approximate the blur kernel, which can be simplified to bi-directional optical flows. So the position learning of sampling points can be supervised by bi-directional optical flows. 3) we build a light-weighted backbone for image restoration problem, which can balance the calculations and effectiveness well. Experimental results show that the proposed method achieves state-of-the-art deblurring performance, but with less parameters and shorter running time.

1. Introduction

Due to relative motion during exposure time, motion blur always occurs when taking pictures. Many factors such as camera shake, object motion and depth variation could result in blur artifacts. Blur artifacts downgrade the image quality, which is harmful to the computer vision tasks such as object detection, text recognition and object tracking because of the blurry structures of objects. Dynamic scene deblurring is to recover clear image from the observed blurry image. The blurring in dynamic scenes can be modeled as non-uniform blur, which is usually formulate as:

$$\mathbf{b} = \mathbf{K}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{b}$, $\mathbf{s}$ and $\mathbf{n}$ represent the vectorized blurry image, clear latent image, and additional noise respectively. $\mathbf{K}$ is the non-uniform blur matrix, each row of which represents a local blur kernel attached to sharp image to generate a blurry pixel. The solution space of non-uniform blur is very large, which causes it hard to solve the $\mathbf{s}$ and $\mathbf{K}$ with $\mathbf{b}$.

To constraint the solution space of non-uniform blur, some hand-crafted priors such as dark channel prior [22], heavy-tailed gradient prior [24], hyper-Laplacian prior [16], extreme channel prior [33], and *etc.*, are introduced. However, the traditional deblurring process involves expensive non-convex optimization, which is time and memory consuming. Once the priors are not suitable, artifacts such

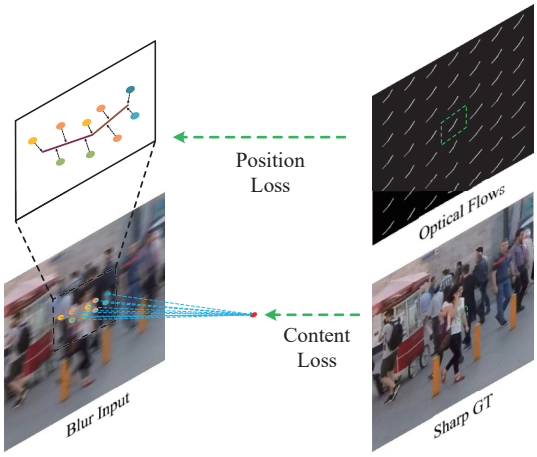

Figure 1. The overview of the proposed method. The sampling points of deformable convolution are used to approximate the local blur kernel, and supervised by the bi-directional optical flows, which can be easily obtained from the deblurring datasets. The position loss is used to train the sampling points getting closer to the optical flows.

*Corresponding Author.

as ringing artifacts would appear. Kim *et al.* [9] design a segment-based non-uniform debulr framework, in which the blur kernel is shared within one segmentation. Some works [10, 28, 6] approximate the blur kernel to be a local linear kernel, and estimate the latent image and linear kernel jointly. Li *et al.* [20] learn the natural image prior with a convolution neural network, and use it as a regularization term.

With the development of deep learning, many researchers [21, 23, 27, 18, 29] try to build deblurring neural network to deblur images or videos in an end-to-end manner and achieve state-of-the-art performance. Without the blur kernel estimation process, these methods directly generate clear images from blurry input images. However, owing to the complexities of dynamic scene deblurring, a large number of convolution layers should be added into the deblurring network to ensure enough large receptive field for handling severe motion blur situations. In addition, the degrees of blur vary among different blurry images but the parameters and structures of deblurring network are fixed. So the current deblurring networks contain amounts of parameters to deal with various blurry images, which causes the networks large in size and massive in calculation.

In this paper, we propose a novel dynamic scene deblurring network. We start from the deblurring deconvolution operation and then try to model the deconvolution operation with stacked deformable convolutional layers. The deformable convolutional layers can automatically adjust the distributions and weights of the sampling points based on blur features contained in the blurry image. Therefore, the entire network has the ability to automatically adjust the receptive fields and weights according to the blurry inputs. What's more, the deformable convolution layers can use optical flows as the auxiliary supervising. Experiments show that the proposed method achieves state-of-the-art deblurring performance, but with less parameters and shorter running time. The overview of the proposed method is shown in Figure 1.

In general, the main contributions of this paper are summarized as below:

- We design a feature deconvolution module to approximate the deblurring deconvolution at feature level using modulated deformable convolutions. The distribution of the sampling points of the deformable convolution and the magnitude of the weights can be automatically adjusted according to the direction and degree of blur, which achieves the adaptation of receptive fields and weights. Compared with the other deblurring networks based on regular convolution, the proposed deblurring network has less parameters and a simpler structure.

- We introduce the bi-directional optical flows to guide the learning of deformable convolutions, whose sampling points can be used to approximate blur kernels. Without directly using artificially synthesized optical-image pairs, we calculate optical flows directly from the deblurring dataset. Experiment shows that better results can be achieved when using the optical flow guided training.

- We design a light-weighted backbone for image restoration problem. In order to reduce the loss of spatial information, we only use downsampling operation once and apply dilated convolutions with different dilation rates to ensure that the network's receptive field is unchanged. Since there is no reduction in spatial resolution, it is not necessary to increase the number of channels, so the convolutional layers have fewer parameters than networks which use multiple downsampling operations.

2. Related Work

2.1. Deep Image Deblurring

With the development of deep learning and the emergence of large numbers of datasets for deblurring problem, many researchers [21, 23, 29, 18, 19] have designed end-to-end deblurring networks achieving excellent performance. These deep learning based methods directly predict the clear images from the blurry images without the need for blur kernel estimation process, which makes them more efficient than those with blur kernel estimations. Nah *et al.* [21] propose a multiscale convolutional neural network to remove blurring in dynamic scenes and conduct excellent deblurring results. The blur is removed from the coarse scale and then refined to the original scale. The networks at each scale contain 40 convolutional layers and do not share parameters across scales. So the total parameters of the network are huge in number, which leads to an increase in computations and inference time, and great difficulties to train. In order to solve these problems, Tao *et al.* [29] use an encoder-encoder structure with skip connections and parameter sharing at three scales, which can stablize the training process and achieve better deblurring performance with impressive results. The network structure is simpler and the number of parameters is smaller.

2.2. Deconvolutional Neural Network

Some methods [35, 32] try to model the deblurring problem as a deconvolution problem and design the networks to approximate the deconvolution operations. Zhang *et al.* [35] analyzed that deconvolution can be implemented by using a spatial recurrent neural network (RNN). In order to obtain large receptive fields and fuse features from different filtering directions, they use four RNNs and add a

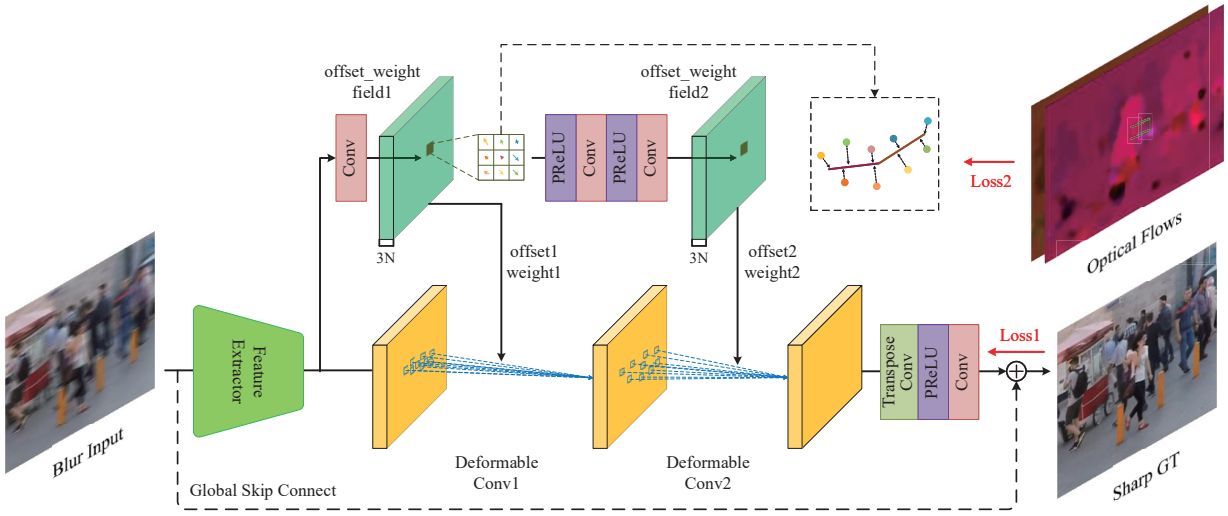

Figure 2. The overall architecture of the proposed deblurring network, which contains four components, *i.e.* feature extractor, neck, head and global skip connection. The neck is composed of two modulated deformable convolutions, and the head is used to reconstruct the clear image. The offsets and weights can be adjusted adaptively according to the blur features extracted from blurry image.

convolution layer after each RNN. The pre-trained VGG16 [26] sub-network is then used to predict the spatially variant weights for the RNNs. However, the RNNs can not calculate parallelly along the spatial dimension, so the inference time of this method is still not reduced, and VGG16 is used as the weight generation network, which increases the parameters and calculations of the network. Although the deconvolution kernel can be approximated using large-kernel convolution, a large amount of parameters are introduced. Xu *et al.* [32] use separable convolutions to approximate the deconvolution kernel, and then design a deconvolution neural network for image deblurring. However, this method can only remove uniform blurs, and it needs to train different network parameters for different blur kernels, which restricts the application in dynamic scenes.

2.3. Deblurring with Optical Flow

Some dynamic scene deblurring methods [14, 1] use bi-directional optical flows to approximate the blur kernel and generate reliable deblurring results. Kim *et al.* [14] propose a segmentation-free dynamic scene deblurring method, and apply bi-directional optical flows calculated from the previous and next frames to approximate the blur kernel. This assumption reduces the solution space and makes it easier to solve. However, it is usually not true because the motions in dynamic scenes are very complex, and the blur kernels usually have complex shapes rather than linear shapes. Chen *et al.* [1] design a self-supervised learning framework to fine-tune the existing deblurring networks and achieve significant improvements. They use optical flow prediction networks [3, 11] to obtain the bi-directional optical flows which are used as local blur kernels, and then blur the recovered image. The loss is calculated by the supervision of the original blurry picture. This technique improves the performance of existing methods and makes the deblurring results more faithful to the latent clear image.

3. Methodology

In this section, we introduce the architecture of the proposed deblurring network in detail. The overall architecture is illustrated in Figure 2. In addition, the feature deconvolution module, optical flow guided training and loss functions used for training are described seperately.

3.1. Network Architecture

The deblurring network consists of four parts, *i.e.* backbone, neck, head and global skip connection. The backbone, also refered as feature extractor, takes blurry images as input and extracts content features and blur features. The neck is a feature deconvolution module which contains two modulated deformable convolutions. The head is used to upsample the feature maps from the neck and reconstruct the RGB image. And with the help of global skip connection, the network only needs to learn the residual between blurry image and clear image.

The Dilated Backbone. The backbone is used to extract the blur feature and encode the content feature from the input blur image. The structure of the backbone is illustrated in Figure 3. To decrease the difficulty of reconstruction, we only use the downsampling operation once. In order to ensure that there are large enough receptive fields for deblurring, we use dilated convolutions with different

dilation rates. The dilation rates are set to 1, 2 and 4 in the network. Since there is no downsampling, the spatial resolution of feature maps are unchanged, so the number of the output channels does not need to be increased, which only needs to remain the same as that of the input channels. At the same time, in order to make better use of the features of different scales, a fusion module is introduced to fuse the output feature maps of different layers, which can also accelerate the training process benefited from the skip connection. The feature maps after the fusion module are divided into two parts as outputs of the backbone. One of the outputs is the blur feature, which is used to generate the offsets and weights for the deformable convolutions in the neck. And the other is the content feature, which is used as the input of the deformable convolution.

The Neck. The neck is a feature deconvolution module, which is a blur-adaptive component and contains two modulated deformable convolutions [37]. The blur feature extracted by the backbone is used to generate offsets and weights for the two deformable convolutions. The offsets would be small when there are slight blur in the image, but bigger when there are severe blur. Different from the model [21, 29, 19] constructed by regular convolutions, the receptive field of our network can be adaptively adjusted according to the blurring degrees of the input images.

The Head. The head is used to reconstruct the RGB image which has the same size with input image. There is a downsampling operation in the backbone, so the resolution of the feature maps is half of that of the input images. Therefore an upsampling operation should be added in the neck to enlarge the spatial size of feature maps. Instead of bilinear interpolation, we use transpose convolution to upsample the feature maps.

The Global Skip Connection. The global skip connection is usually applied in learning-based image restoration tasks [13, 36, 23, 19]. Instead of restoring the image directly, the network only needs to learn the residual between the blurred images and the ground truth, with the help of global skip connection, which reduces the learning difficulty significantly. Therefore, we add global skip connection in the deblurring network like [18, 19].

3.2. Feature Deconvolution Module

In this subsection, we start from a simple image deconvolution equation, and then try to approximate it using deformable convolutions. The deblurring process of uniform blur in Fourier domain can be described as

$$\mathcal{F}[S] = \frac{1}{\mathcal{F}[K]} \cdot \mathcal{F}[B], \tag{2}$$

where $\mathcal{F}$ is the Fourier transformation, and S, B and K are the sharp image, blurry image and blur kernel seperately. The $\frac{1}{\mathcal{F}[K]}$ is the inverse kernel for deconvolution, which

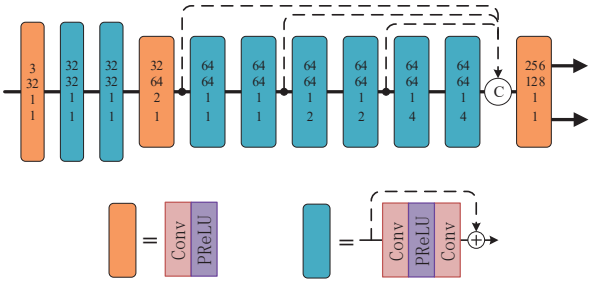

Figure 3. The architecture of feature extractor. The 4 numbers in each module indicate in channels, out channels, stride and dilation rate seperately. The size of all convolution kernels is 3. The dashed lines represent skip connections. The ⓒ is concatenation along channel dimension, and the ⊕ is element-wise summation.

varies according to the blur degree. It means that different blur images have different inverse kernels. Moreover, [35] shows that the non-zero region of a inverse filter is larger than that of the blur kernel, which means that a much larger receptive field is needed to deconvolution. To approximate the inverse filter, regular convolution layer is not efficient, restricted by the regular sampling grids and fixed weights. Conversely, the deformable convolution [2, 37] has flexible sampling points, whose offsets can be learned from the input feature and adjusted adaptively. Morever, the weights can also be changed by multipling a mask. Based on above, we use modulated deformable convolutions to approximate the deconvolution operation. Due to the sparse sampling, deformable convolutions are used at feature level instead of image level in this paper. In addition, to enlarge the receptive field and make the module more interpretable, we change the Equ. 2 to

$$\mathcal{F}[S] = \frac{1}{\mathcal{F}[K]^2} \cdot \mathcal{F}[K] \cdot \mathcal{F}[B]. \tag{3}$$

So the feature deconvolution module can be divided into two parts, which are used to approximate the $\mathcal{F}[K]$ and $\frac{1}{\mathcal{F}[K]^2}$ in spatial domain seperately. It is different from the usage in [30], *i.e.* we do not stack deformable convolutions in cascade to deblur gradually. Specifically, the deformable sampling parameter of the second part is derived from that of the first part, instead of output feature of the first part, as shown in Figure 2. It should be noted that in this module, we only use one deformable convolution to approximate the $\frac{1}{\mathcal{F}[K]^2}$ in spatial domain, more deformable convolutions can be added to get better approximation and better deblurring performance. The prediction of offsets and masks for these two deformable convolutions can be formulated as:

$$\mathbf{ow1} = Conv(\mathbf{x}), \tag{4}$$

$$\mathbf{ow2} = Conv(f(Conv(f(\mathbf{ow1})))), \tag{5}$$

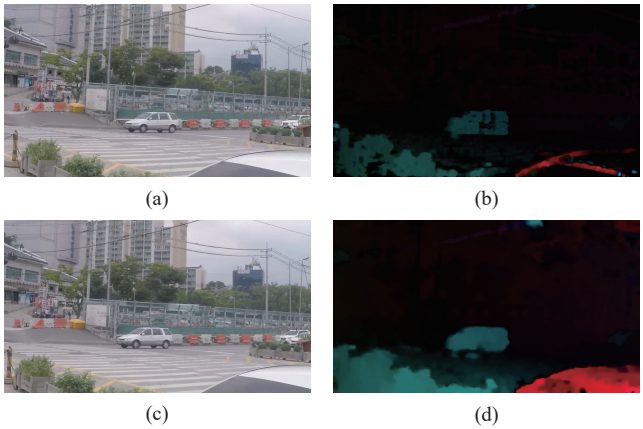

Figure 4. Visual comparision of the optical flows. (a) and (c) are two consecutive frames. (b) and (d) are the optical flows calculated by Farnback [4] and DIS [17] respectively.

where $Conv$ represents the regular convolution layer, f is the activate function, which is Parametric ReLU (PReLU) [8] in the proposed deblurring network. **x** is the blur feature extracted by the backbone. **ow1** and **ow2** are the generated offsets and weights, used for the two stacked deformable convolutional layers.

3.3. Optical Flow Guided Training

According to the previous section, we need to constrain the sampling points of the first deformable convolution to be close to the distribution of the blur kernel. But usually the datasets only contain blurry images and the corresponding clear images, there is no blur kernel available. So we use the same method as [14] to approximate the blur kernel with optical flows of the current frame to the previous and next frames, that is, the blur kernel can be simplified into bi-directional optical flows. Therefore, it only needs to make the spatial distribution of the deformable sampling points close to the two optical flow strips. In order to obtain more accurate dense optical flows, as shown in Figure 4, we use the DIS algorithm [17] to calculate the optical flows, instead of Farnback algorithm [4].

There are two methods to measure the matching degree between the sampling points and the two optical line segments here. The first way is to fit a two-dimensional curve with the sampling points firstly, and then calculate the matching error between the curve and the bi-directional optical flows. However, it is difficult to fit a curve by using two-dimensional point sets, so this method is very hard to realize. Another way is to use the mean of the shortest distances from all sampling points to the bi-directional optical flows, which is formulated as:

$$md = \frac{1}{n}\sum_{i=1}^{n} min(d_{i1}, d_{i2}), \tag{6}$$

Algorithm 1 Algorithm for calculating the shortest distance from sampling points to bi-directional optical flows.

Input:
The coordinate of the sampling point (x, y);
The parameters of the optical flow (u, v);
Output:
The shortest distance min_dis;

Calculate the abscissa of foot point from the sampling point to optical flow $x_0 = \frac{uvy+v^2x}{u^2+v^2}$;
if $x_0 \in [min(0,u), max(0,u)]$ **then**
  $min\_dis = \frac{\|ux-vy\|}{\sqrt{u^2+v^2}}$;
else
  $d_1 = \sqrt{x^2+y^2}$;
  $d_2 = \sqrt{(x-u)^2+(y-v)^2}$;
  $min\_dis = min(d_1, d_2)$;
end if
return min_dis;

where d_{ij} represents the shortest distance from the *i-th* point to the *j-th* optical flow. n is the number of the sampling points, which is 25 in our experiments. md is the mean of the n shortest distances. The details of calculating the shortest distance from sampling points to optical flows are shown in Algorithm 1.

3.4. Loss Functions

The loss function used for training is composed of pixel-wise loss, perceptual loss and position loss, which is formulated as:

$$\mathcal{L} = \mathcal{L}_{pixel} + \lambda_1 \cdot \mathcal{L}_{percep} + \lambda_2 \cdot \mathcal{L}_{position}, \tag{7}$$

where λ_1 is set to 0.01, and λ_2 is set to 0.0001 in our experiments.

Pixel loss. Two classical loss functions used for pixel-level are MAE and MSE loss, which are also referred as L1 and L2 loss respectively. The L2 loss has been used in many deblurring problems [29, 34] achieving impressive results. Therefore, we also use the L2 loss as the pixel-wise loss, which can be formulated as:

$$\mathcal{L}_{pixel} = \frac{1}{2N_p}\left\|L - S\right\|_F^2, \tag{8}$$

where S and L denote the ground truth clear image and the model output respectively. N_p is the number of elements of S and L.

Perceptual loss. For generating deblurred images with sharp structures, some methods [21, 23, 18] calculate loss on semantic feature, such as Patch-GAN loss and perceptual loss [12]. Each element in the semantic feature corresponds to a local region of the input image, therefore this loss can focus on restoring the general content and conducting to the

Table 1. Comparison results on GOPRO testing dataset in terms of performance and efficiency

Method	Sun *et al.* [28]	DeepDeblur [21]	Zhang *et al.* [35]	SRN [29]	DeblurGAN-v2 [19]	Ours(A)	Ours(B)
PSNR	24.64	29.08	29.19	**30.10**	29.55	29.57	29.81
SSIM	0.8429	0.9135	0.9306	0.9323	0.9340	0.9338	**0.9368**
Runtime	12.1 min	3.1 s	1.4 s	0.4 s	0.35 s	**0.01 s**	**0.01 s**
Model Size	54.1 MB	303.6 MB	37.1 MB	33.6 MB	15.0 MB	**3.1 MB**	**3.1 MB**

recovery of image structure. However, GAN [7] needs to alternately train the discriminator and generator and carefully trade off the training times of discriminator and generator. Conversely, the perceptual loss is based on the pre-trained VGG-Net, which requires no training, so the perceptual loss is easier to use. Therefore, in this paper we use the perceptual loss as the content loss, which can be formulated as:

$$\mathcal{L}_{percep} = \frac{1}{2N_c} \left\| \phi_i(L) - \phi_i(S) \right\|_F^2, \tag{9}$$

where ϕ_i represents the feature maps of the i-th layer of VGG-16, which is set to 12 in our experiments, and N_c is the number of elements of $\phi_i(L)$ and $\phi_i(S)$.

Position loss. To better train the deblurring network, we add bi-directional optical flow for auxliary supervision. The distributions of sampling points of the first deformable convolution layer should be close to the optical flows. Therefore, the position loss is the mean value of the shortest distances from points to optical flows. In addition, to reduce the impact of miscalculating optical flows and restriction on model fitting capabilities, a margin is added to the loss, *i.e.* only points with a distance greater than the margin are penalized. So the final position loss is formulated as:

$$\mathcal{L}_{position} = \frac{1}{n} \sum_{i=1}^{n} max(min(d_{i1}, d_{i2}), M), \tag{10}$$

where n is the number of sampling points, and M represents the margin value.

4. Experiments

In this section we make a comparision with state-of-the-art image deblurring methods and carry out abalation experiments to evaluate the effectiveness of the proposed optical flow guided deep deblurring network. All the experiments are conducted with an i7-6800K CPU and four NVIDIA Geforce GTX 1080Ti GPUs. The models are implemented with the Pytorch 1.1.0 Library.

4.1. Datasets

For training the proposed end-to-end deblurring network, a large training dataset containing blurry and clear image pairs should be created. A classical way to generate the blur images is convolving the sharp images and generated blur kernels. But the scenes to be simulated are limited and still different from the real blur images captured by camera. Another way is to avarage short-exposure frames captured by high-speed cameras to similuate long-exposure blurry images which are more realistic and have complex blurs. For fair comparison, we use the GOPRO dataset proposed by [21], which contains 3214 image pairs. Similarly we use 2103 image pairs for training and 1111 image pairs for testing here.

Table 2. Results of abalation experiments

Dilated Backbone		✓	✓	✓
FD Module			✓	✓
OG Training				✓
PSNR	28.34	29.01	29.57	29.81
SSIM	0.9124	0.9237	0.9338	0.9368

4.2. Experiment Setting

We use Xavier [5] to initialize the parameters and Adam [15] optimizer to train the deblurring network. The β_1, β_2 and ϵ are set to 0.9, 0.9 and 10^{-8} seperately. The polynomial-decay strategy is used to decay the learning rate from 10^{-3} to 10^{-6} at 2000 epochs, and the power of the decay strategy is set to 0.9 in our experiments. The training process takes about 44 hours. Experiments show that 2000 epochs are enough to converge the model. Each training batch contains 32 blur-ground truth image patches. Every patch is augmented by flipping, rotation, and permutation of RGB channels, and then cropped to 256×256 size.

4.3. Comparisons with State-of-the-art Methods

We evaluate the proposed deblurring network and make comparisons with state-of-the-art deblurring methods [28, 21, 35, 29, 19] in terms of PSNR, SSIM [31], model size and inference time for 720p images. The quantitative results are shown in Table 1, where Ours(B) and Ours(A) represent the network with and without optical flow guided training

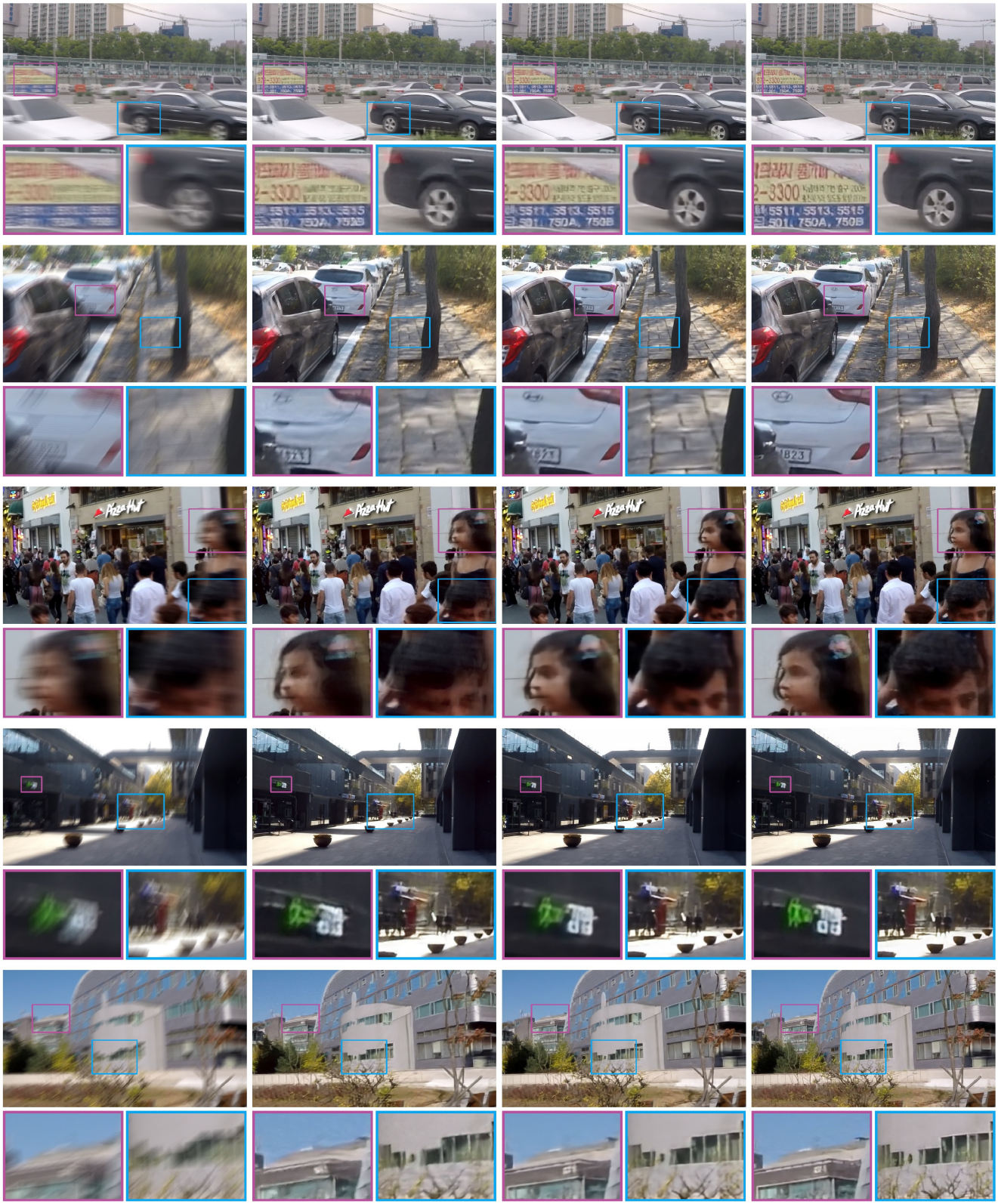

Figure 5. Visual comparisons on the GOPRO testing dataset. There are 5 blurry images from different scenes. From Left to Right: input blurry images, results of Nah *et al.* [21], results of Tao *et al.* [29], and results of the proposed method.

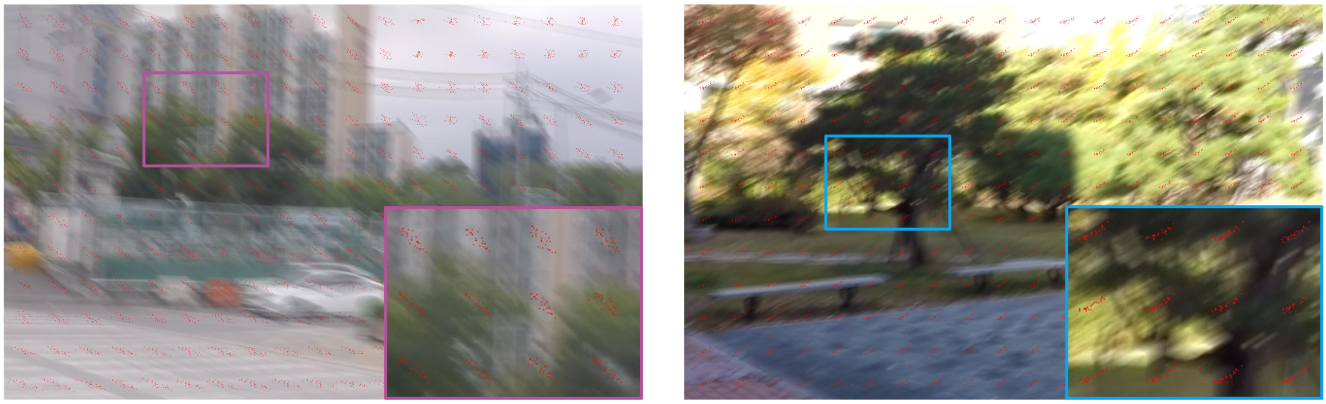

Figure 6. Visualization of the sampling points generated by the proposed deblurring network.

respectively. Our deblurring method achieves the best performance among state-of-the-art methods, with the highest SSIM value and the second highest PSNR value. Moreover, the proposed network can deblur a 1280×720 image with the fastest speed, nearly **0.01s** per image, which is **40**× and **35**× faster than Tao *et al.* [29] and Kupyn *et al.* [19] respectively. Overall, our method achieves the real-time deblurring without performance degradation. In addition, the proposed network owns the smallest model size, nearly 0.8M parameters, which is **100**× and **10**× smaller compared with Nah *et al.* [21] and Tao *et al.* [29] respectively. The visualization results are shown in Figure 5. Compared with the deep learning based methods [21, 29], the restored images of our method are clearer and sharper at the edges. The content of the deblurred image is more faithful, *e.g.* the numbers of the license plate are deblurred perfectly, while [21] and [29] fail to do that.

4.4. Ablation Experiments

We make ablation experiments to evaluate the effectiveness of the proposed components, including the dilated backbone, the feature deconvolution module (FD Module) and the optical flow guided training (OG Training). The results of ablation experiments are summarized in Table 2. By setting all the dilation rates of backbone to 1 and replacing the neck with a residual block, we build a baseline deblurring model, which achieves 28.34 dB on PSNR and 0.9124 on SSIM. When replacing the backbone with dilated backbone, it achieves an increase of 0.67 dB on PSNR and 0.0113 on SSIM, which means the dilated backbone could extract better features benefitting from the larger receptive field. The effectiveness of FD Module can be evaluated by comparing Column 3 and Column 4. It achieves better performance benefitting from the adaptive adjustment of receptive field and weights, which means deformable convolution is more suitable for deblurring task than norm convolution. When using optical flow guided training (Column 5), it gets a better result thanks to the additional supervision and efficient training.

4.5. Effectiveness of Feature Deconvolution Module

Figure 6 shows two example distributions of sampling points used for the first deformable convolution of feature deconvolution module. It can be seen from the visualization results that the distributions of sampling points can be adjusted adaptively according to the blur patterns contained in the blurry input image. However, the deblurring methods with regular convolution are hard to do that restricted by the fixed structures and parameters, which need to stack more layers for larger receptive fields and stronger fitting capabilities. In addition, the intensive degree of sampling points can vary with blur degrees. Inspired by the observation, the distributions of the sampling points could be used as discriminative blur detection features [25], *i.e.* the larger variations of the distributions of the sampling points, the blurrier the image regions. So the sub network for predicting the sampling points in our deblurring network could be used to detect blur regions when fine-tuned.

5. Conclusion

In this paper, we propose a novel spatially variant deconvolutional neural network for dynamic scene deblurring. The deblurring network is powered by two modulated deformable convolutions and a light-weighted feature extractor. For better training the network, we use bi-directional optical flows as the auxilary supervision. Experimental results show that the proposed method achieves state-of-the-art performance but has less parameters and shorter running time compared with representative deep learning based deblurring methods.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61632018 and 61825603.

References

[1] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *Proceedings of the IEEE International Conference on Computational Photography*, pages 1–9, 2018.

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017.

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[4] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis*, pages 363–370, 2003.

[5] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[6] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2017.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[9] Tae Hyun Kim, Byeongjoo Ahn, and Kyoung Mu Lee. Dynamic scene deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3160–3167, 2013.

[10] Tae Hyun Kim and Kyoung Mu Lee. Segmentation-free dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2766–2773, 2014.

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[14] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. Dynamic video deblurring using a locally adaptive blur model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2374–2387, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1033–1041, 2009.

[17] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast optical flow using dense inverse search. In *Proceedings of the European Conference on Computer Vision*, pages 471–488, 2016.

[18] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[19] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8878–8887, 2019.

[20] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Learning a discriminative prior for blind image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6616–6625, 2018.

[21] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[22] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016.

[23] Sainandan Ramakrishnan, Shubham Pachori, Aalok Gangopadhyay, and Shanmuganathan Raman. Deep generative filter for motion deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2993–3000, 2017.

[24] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *Acm transactions on graphics (tog)*, 27(3):73, 2008.

[25] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1279–1288, 2017.

[28] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015.

[29] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[30] Hua Wang, Dewei Su, Chuangchuang Liu, Longcun Jin, Xianfang Sun, and Xinyi Peng. Deformable non-local network for video super-resolution. *IEEE Access*, 7:177734–177744, 2019.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[32] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1790–1798, 2014.

[33] Yanyang Yan, Wenqi Ren, Yuanfang Guo, Rui Wang, and Xiaochun Cao. Image deblurring via extreme channels prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4003–4011, 2017.

[34] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019.

[35] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2521–2529, 2018.

[36] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[37] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.