ADAPTIVE SINGLE IMAGE DEBLURRING

Maitreya Suin   **Kuldeep Purohit**   **A. N. Rajagopalan**

ABSTRACT

This paper tackles the problem of dynamic scene deblurring. Although end-to-end fully convolutional designs have recently advanced the state-of-the-art in non-uniform motion deblurring, their performance-complexity trade-off is still sub-optimal. Existing approaches achieve a large receptive field by a simple increment in the number of generic convolution layers, kernel-size, which comes with the burden of the increase in model size and inference speed. In this work, we propose an efficient pixel adaptive and feature attentive design for handling large blur variations within and across different images. We also propose an effective content-aware global-local filtering module that significantly improves the performance by considering not only the global dependencies of the pixel but also dynamically using the neighboring pixels. We use a patch hierarchical attentive architecture composed of the above module that implicitly discover the spatial variations in the blur present in the input image and in turn perform local and global modulation of intermediate features. Extensive qualitative and quantitative comparisons with prior art on deblurring benchmarks demonstrate the superiority of the proposed network.

1 Introduction

Blind motion deblurring is an ill-posed problem which aims to recover a sharp image from a given image degraded due to motion based smearing of texture and high-frequency details. Due to its diverse applications in surveillance, remote sensing, and cameras mounted on hand-held and vehicle mounted cameras, deblurring has gathered substantial attention from computer vision and image processing community in past two decades.

Majority of existing deblurring approaches are based on variational model, whose key component is the regularization term. Large literature studies design of priors that are apt for recovering the underlying undistorted image and the camera trajectory. The restoration quality depends on the selection of the prior, its weight, as well as tuning of other parameters involving highly non-convex optimization setups (Nimisha et al. [2017]). A significant number of works have been proposed Paramanand and Rajagopalan [2011], Nimisha et al. [2018a], Rao et al. [2014], Nimisha et al. [2018b], Vasu and Rajagopalan [2017], Paramanand and Rajagopalan [2014], Vijay et al. [2013] where various traditional approaches were adopted for deblurring. Non-uniform blind deblurring for general dynamic scenes is a challenging computer vision problem as blurs arise from various sources including moving objects, camera shake and depth variations, causing different pixels to capture different motion trajectories. Such hand-crafted priors struggle while generalizing across different types of real-world examples, where blur is far more complex than modeled Gong et al. [2017].

Non-uniform blind deblurring for general dynamic scenes is a challenging computer vision problem as blurs arise from various sources including moving objects, camera shake and depth variation, causing different pixels to capture different motion trajectories. Conventional hand-designed formation models would require explicit estimation of all of these independent variables from a single blurred image, which is an extremely ill-posed problem. As a result, applying such algorithms on general dynamic scenes yields images with unpleasant artifacts and incomplete deblurring.

Recent works Purohit et al. [2019], Purohit and Rajagopalan [2020], Mohan et al. [2021, 2019], Vasu et al. [2018] based on deep convolutional neural networks (CNN) have studied the benefits of replacing the image formation model with a parametric model that can be trained to emulate the non-linear relationship between blurred-sharp image pairs. Such works Nah et al. [2017] directly regress to deblurred image intensities and overcome the limited representative capability of variational methods in describing dynamic scenes. These methods can handle combined effects of camera motion and dynamic object motion and achieve state-of-the-art results on single image deblurring task. They have reached a respectable reduction in model size, but still lack in accuracy and are not real-time.
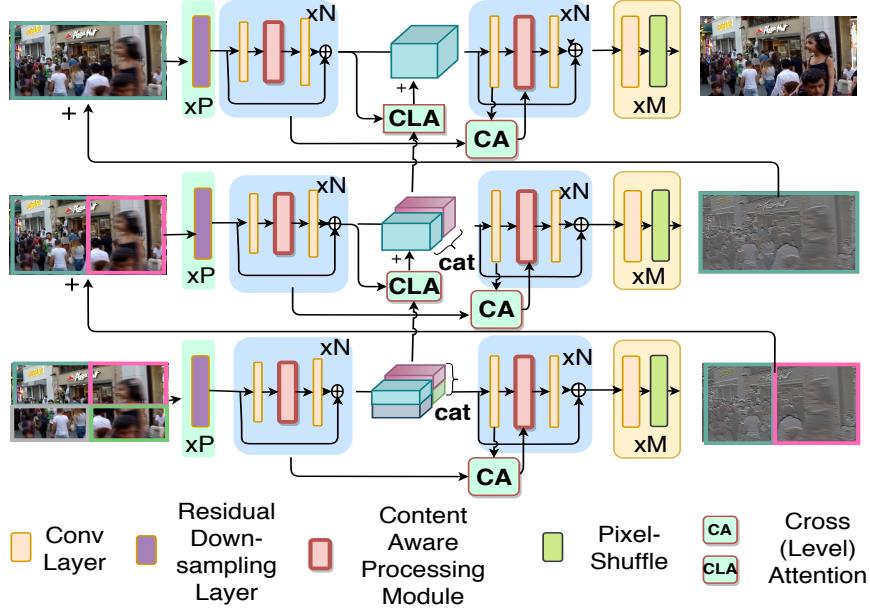
Figure 1: Overall architecture of our proposed network.

Reaching a trade-off between the size and speed of the network, the receptive field and the accuracy of restoration is a non-trivial task. Therefore, our work focuses on the design of efficient modules that achieve performance improvement without requiring a deeper framework. We investigate motion-dependent adaptability within a CNN. This problem is challenging especially since multiple segments of different sizes and motion co-exist in a single image. Since motion blur is essentially a direct aggregation of spatial transformation of the image, a deblurring network can benefit from adapting to the magnitude as well as the direction of motion. We deploy content-aware modules which adjusts the filter to be applied and the context of each pixel depending on the motion information.

Following the state of the art in deblurring, we adopt an multi-patch hierarchical design to directly estimate the restored sharp image. Instead of cascading along depth, we introduce content-aware feature and filter transformation capability through global-local attentive module and residual attention across layers to improve their performance. These modules learn to exploit the similarity in the motion between different pixels within an image and also sensitive to position specific local context.

Our design originates from the intuition that motion blur is essentially a aggregation of various spatially varying transformation of the sharp image, and hence a deblurring network can benefit from implicitly decoding the magnitude as well as the direction of motion accompanied with the global context. The proposed module can dynamically generate global attention map and local filters almost in real-time. The feature transformations and filters estimated by the network are image dependent and hence can be visualized for different images. The efficiency of our architecture is demonstrated through comprehensive comparison on two benchmarks with the state-of-the-art deblurring approaches. Our model achieves superior performance, while being computationally more efficient, resulting in real-time deblurring of HD images on a single GPU. The major contributions in this work are:

1. We propose an efficient deblurring design built on new convolutional modules that learn transformation of features using global attention and dynamic local filters. We show that these two branches complements each other and results in superior deblurring performance. Moreover, the efficient design of attention-module enables us to use it throughout the network without the need of explicit downsampling.

2. We provide extensive analysis and evaluations on static and dynamic scene deblurring benchmarks.

2   Method

To date, the driving force behind performance improvement in deblurring has been the use of a large number of layers and larger filters which assist in increasing the "static" receptive field and the generalization capability of a CNN. However, these techniques offer suboptimal design, since network performance does not always scale with network

depth, as the effective receptive field of deep CNNs is much smaller than the theoretical value (investigated in Luo et al. [2016]).

Although previous multi-scale and scale-recurrent methods have shown good performance in removing non-uniform blur, they suffer from expensive inference time and performance bottleneck while simply increasing model depth. Instead, inspired by Zhang et al. [2019] , we adopt multi-patch hierarchical structure as our base-model, which compared to multi-scale approach has the added advantage of residual-like architecture that leads to efficient learning and faster processing speed.

The overall architecture of our proposed network is shown in Fig. 1. We divide the network into 3 levels instead of 4 as described in Zhang et al. [2019]. We found that the relative performance gain due to the inclusion of level 4 is negligible compared to the increase in inference time and number of parameters. At the bottom level input sliced into 4 non-overlapping patches for processing, and as we gradually move towards higher levels, the number of patches decrease and lower level features are adaptively fused using attention module as shown in Fig. 1. The output of level 1 is the final deblurred image. Note that unlike Zhang et al. [2019], we also avoid cascading of our network along depth, as that adds severe computational burden. Instead, we advocate the use of content-aware processing modules which yield significant performance improvements over even the deepest stacked versions of original DMPHN Zhang et al. [2019]. Major changes incorporated in our design are described next.

Each level of our network consists of an encoder and a decoder. Both the encoder and the decoder are made of standard convolutional layer and residual blocks where each of these residual blocks contains 1 convolution layer followed by a content-aware processing module and another convolutional layer. The content-aware processing module comprises two branches for global and local level feature processing which are dynamically fused at the end. The residual blocks of decoder and encoder are identical except for the use of cross attention in decoder. We have also designed cross-level attention for effective propagation of lower level features throughout the network. We begin with describing content-aware processing module, then proceed towards the detailed description of the two branches and finally how these branches are adaptively fused at the end.

2.1 Attention

Given the input x_m, we generate three attention maps $P \in \mathbb{R}^{C_2 \times HW}$, $Q \in \mathbb{R}^{C_2 \times HW}$ and $M_2 \in \mathbb{R}^C$ using convolutional operations $f_p(\cdot)$, $f_q(\cdot)$ and $f_{M_2}(\cdot)$ where global-average-pooling is used for the last case to get C dimensional representation. We take the first cluster of attention map Q and split it into C_2 different maps $Q = \{q_1, q_2, ..., q_{C_2}\}$, $q_i \in \mathbb{R}^{HW}$ and these represent C_2 different spatial attention-weights. A single attention reflects one aspect of the blurred image. However, there are multiple pertinent properties like edges,textures etc. in the image that together helps removing the blur. Therefore, we deploy a cluster of attention maps to effectively gather C_2 different key features. Each attention map is element-wise multiplied with the input feature map x_{m_1} to generate C_2 part feature maps as

$$x_{m_1}^k = q_k \odot x_{m_1} \quad \text{, with} \sum_{i=1}^{HW} q_{ki} = 1 \quad (k = 1, 2, ..., N) \tag{1}$$

where $x_m^k \in \mathbb{R}^{C \times HW}$. We further extract descriptive global feature by global-sum-pooling (GSP) along HW dimension to obtain k^{th} feature representation as

$$\bar{x}_{m_1}^k = GSP_{HW}(x_{m_1}^k) \quad (k = 1, 2, ..., N) \tag{2}$$

where $\bar{x}_m^k \in \mathbb{R}^C$. Now we have $\bar{x}_{m_1} = \{\bar{x}_{m_1}^1, \bar{x}_{m_1}^2, ..., \bar{x}_{m_1}^{C_2}\}$ which are obtained from C_2 different attention-weighted average of the input x_m. Each of these C_2 representations is expressed by an C-dimensional vector which is a feature descriptor for the C channels. We further enhance these C dimensional vectors by emphasizing the important feature-embeddings as

$$\bar{x}_{m_1 m_2}^k = M_2 \odot \bar{x}_{m_1}^k \tag{3}$$

where M_2 can be expressed as

$$M_2 = f_{m_2}(\bar{x}_{m_1}; \theta_{m_2}) \in \mathbb{R}^C \tag{4}$$

Next we take the set of attention maps $P = \{p_1, p_2, ..., p_{HW}\}$ where $p_i \in \mathbb{R}^{C_2}$ is represents attention map for i^{th} pixel. Intuitively, p_i shows the relative importance of C_2 different attention-weighted average ($\bar{x}_{m_1 m_2}$) for the current pixel and it allows the pixel to adaptively select the weighted average of all the pixels. For each output pixel j, we element-wise multiply these C_2 feature representations $\bar{x}_{m_1 m_2}^k$ with the corresponding attention map p_j, to get

$$y^j = p_j \odot \bar{x}_{m_1 m_2} \text{ with} \sum_{i=1}^{C_2} p_{ji} = 1 \text{ , } (j = 1, 2, ..., HW) \tag{5}$$

where $y^j \in \mathbb{R}^{C \times C_2}$. We again apply global-average-pooling on y^j along C_2 to get C dimensional feature representation for each pixel as

$$\bar{y}^j = GAP_{C_2}(y^j) \tag{6}$$

where $\bar{y}^j \in \mathbb{R}^C$ represent the accumulated global feature for the jth pixel.

(a) Blurred Image   (b) Blurred patch   (c) MS-CNN   (d) DelurGAN   (e) SRN   (f) DelurGAN-V2   (g) Stack(4)-DMPHN   (h) Ours
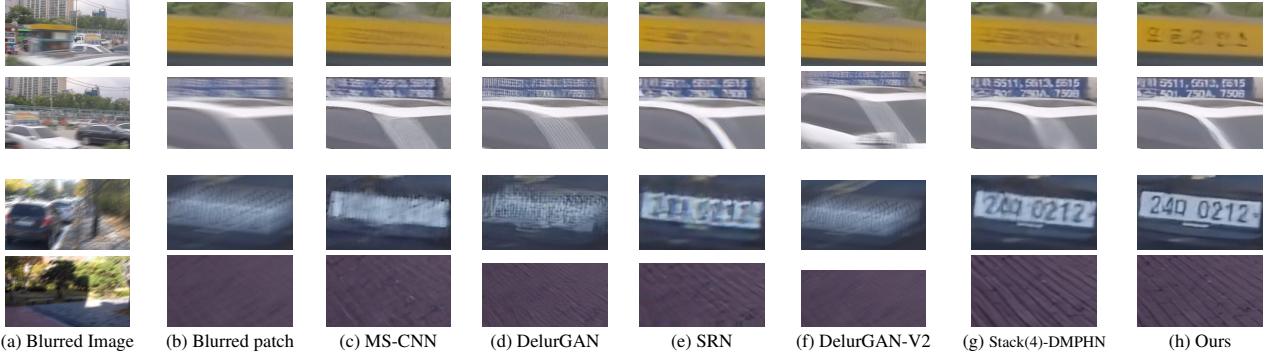
Figure 2: Visual comparisons of deblurring results on images from the GoPro test set Nah et al. [2017]. Key blurred patches are shown in (b), while zoomed-in patches from the deblurred results are shown in (c)-(h).

2.2 Filtering

We design a content-aware "global-local" processing module which depending on the input, deploys 2 parallel branch to fuse global and local features. The global branch is made of the the attention module described above. For decoder this includes both self and cross module attention whereas for encoder only self-attention is used. In contrary to Bello et al. [2019], for the local branch we use a content-aware convolutional layer where the filter is spatially varying and changes at runtime depending on the input image to handle spatially-varying dynamic motion blur effectively. Our work is based on Su et al. [2019] as we use a 'meta-layer' to generate pixel dependent spatially varying kernel to implement spatially variant convolution operation. Given the input feature map $x \in \mathbb{R}^{C \times H \times W}$, we apply an kernel generation funtcion to generate a spatially varying kernel K and do the convolution operation as:

$$x_i^{dyn} = \sum_{j \in \Omega(i)} K_{p_i, p_j} W[p_i - p_j] x_j + b \quad (i = 1, 2, ..., HW) \tag{7}$$

where $x_i^{dyn} \in \mathbb{R}^C$, $p_i = (x_i, y_i)^T$ are pixel coordiantes, $\Omega()$ defines the convolution window, b denotes biases and K_{p_i, p_j} is the pixel dependent kernel generated. Standard spatial convolution can be seen as a special case of the above with adapting kernel being constant $K_{p_i, p_j} = 1$. Finally we sum-fuse the output of these two branches as

$$x^{GL} = x^{att} + x^{dyn} \tag{8}$$

3 Experiments

3.1 Implementation Details

Datasets: We follow the configuration of Zhang et al. [2019], Kupyn et al. [2019], Tao et al. [2018], Kupyn et al. [2017], Nah et al. [2017], which train on 2103 images from the GoPro dataset (Nah et al. [2017]).

Training settings and implementation details: All the convolutional layers within our proposed modules contain 128 filters. The hyper-parameters for our encoder-decoder backbone are $N = 3$, $M = 2$, and $P = 2$, and filter size in PDF modules is 5×5. Following Zhang et al. [2019], we use batch-size of 6 and patch-size of 256×256. Adam optimizer Kingma and Ba [2014] was used with initial leaning rate 10^{-4}, halved after every 2×10^5 iterations. We use PyTorch (Paszke et al. [2017]) library and Titan Xp GPU.

3.2 Performance comparisons

The main application of our work is efficient deblurring of general dynamic scenes. Due to the complexity of the blur present in such images, conventional image formation model based deblurring approaches struggle to perform well. Hence, we compare with only two conventional methods Whyte et al. [2012], Xu et al. [2013] (which are

Table 1: Performance comparisons with existing algorithms on 1103 images from the deblurring benchmark GoPro Nah et al. [2017].

Method	PSNR	SSIM	Time
Xu et al. [2013]	21	0.741	3800
Whyte et al. [2012]	24.6	0.846	700
Hyun Kim et al. [2013]	23.64	0.824	3600
Gong et al. [2017]	26.4	0.863	1200
Nah et al. [2017]	29.08	0.914	6
Kupyn et al. [2017]	28.7	0.858	1
Tao et al. [2018]	30.26	0.934	1.2
Zhang et al. [2018]	29.19	0.931	1
Gao et al. [2019]	30.90	0.935	1.0
Zhang et al. [2019]	31.20	0.940	0.98
Kupyn et al. [2019]	29.55	0.934	0.48
Ours	31.85	0.948	0.34

selected as representative traditional methods for non-uniform deblurring, with publicly available implementations). We provide extensive comparisons with state-of-the-art learning-based methods, namely MS-CNNNah et al. [2017], DeblurGANKupyn et al. [2017], DeblurGAN-v2Kupyn et al. [2019], SRNTao et al. [2018], and Stack(4)-DMPHNZhang et al. [2019]. We use official implementation from the authors with default parameters.

Quantitative Evaluation We show performance comparisons on two different benchmark datasets. The quantitative results on GoPro testing set are listed in Table 1.

The average PSNR and SSIM measures obtained on the GoPro test split is provided in Table 1. It can be observed from the quantitative measures that our method performs better compared to previous state-of-the-art.

Qualitative Evaluation: Visual comparisons on different dynamic and 3D scenes are shown in Figs. 2. Visual comparisons are given in Fig. 2. We observe that the results of prior works suffer from incomplete deblurring or artifacts. In contrast, our network is able to restore scene details more faithfully which are noticeable in the regions containing text, edges, etc. An additional advantage over Hyun Kim et al. [2013], Whyte et al. [2012] is that our model waives-off the requirement of parameter tuning during test phase. The proposed method achieves consistently better PSNR, SSIM and visual results with lower inference-time than DMPHN (Zhang et al. [2019]).

4 Conclusion

We proposed a new content-adaptive architecture design for the challenging task of removing spatially-varying blur in images of dynamic scenes. Efficient self-attention is utilized in all the encoder-decoder to get better representation whereas cross-attention helps in efficient feature propagation across layers and levels. Proposed dynamic filtering module shows content-awareness for local filtering. The proposed method is more interpretable which is one of its key strengths. Our experimental results demonstrated that the proposed method achieved better results than state-of-the-art methods on two benchmarks both qualitatively and quantitatively. We showed that the proposed content-adaptive approach achieves an optimal balance of memory, time and accuracy and can be applied to other image-processing tasks. Refined and complete version of this work appeared in CVPR 2020.

References

TM Nimisha, Akash Kumar Singh, and AN Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *Proceedings of the IEEE E International Conference on Computer Vision (ICCV)*, 2017.

Chandramouli Paramanand and Ambasamudram N Rajagopalan. Depth from motion and optical blur with an unscented kalman filter. *IEEE Transactions on Image Processing*, 21(5):2798–2811, 2011.

Thekke Madam Nimisha, Kumar Sunil, and AN Rajagopalan. Unsupervised class-specific deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 353–369, 2018a.

Makkena Purnachandra Rao, AN Rajagopalan, and Guna Seetharaman. Harnessing motion blur to unveil splicing. *IEEE transactions on information forensics and security*, 9(4):583–595, 2014.

TM Nimisha, AN Rajagopalan, and Rangarajan Aravind. Generating high quality pan-shots from motion blurred videos. *Computer Vision and Image Understanding*, 171:20–33, 2018b.

Subeesh Vasu and AN Rajagopalan. From local to global: Edge profiles to camera motion in blurred images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4447–4456, 2017.

Chandramouli Paramanand and AN Rajagopalan. Shape from sharp and motion-blurred image pair. *International journal of computer vision*, 107(3):272–292, 2014.

Channarayapatna Shivaram Vijay, Chandramouli Paramanand, Ambasamudram Narayanan Rajagopalan, and Rama Chellappa. Non-uniform deblurring in hdr image reconstruction. *IEEE transactions on image processing*, 22(10): 3739–3750, 2013.

Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, AVD Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

Kuldeep Purohit, Anshul Shah, and AN Rajagopalan. Bringing alive blurred moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2019.

Kuldeep Purohit and AN Rajagopalan. Region-adaptive dense network for efficient motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11882–11889, 2020.

MR Mahesh Mohan, GK Nithin, and AN Rajagopalan. Deep dynamic scene deblurring for unconstrained dual-lens cameras. *IEEE Transactions on Image Processing*, 30:4479–4491, 2021.

MR Mohan, Sharath Girish, and AN Rajagopalan. Unconstrained motion deblurring for dual-lens cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7870–7879, 2019.

Subeesh Vasu, Venkatesh Reddy Maligireddy, and AN Rajagopalan. Non-blind deblurring: Handling kernel uncertainty with cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2018.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, volume 1, page 3, 2017.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems*, pages 1673–1682, 2018.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. *arXiv preprint arXiv:1904.09925*, 2019.

Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3106–3115, 2019.

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019.

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.

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.

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *arXiv preprint arXiv:1711.07064*, 2017.

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

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013.

Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *International journal of computer vision*, 98(2):168–186, 2012.

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.

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.

Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019.