



Adaptive Spatio-Temporal Convolutional Network for Video Deblurring

Fengzhi Duan and Hongxun Yao(✉)

Harbin Institute of Technology, Harbin, China
h.yao@hit.edu.cn

Abstract. Video deblurring is a challenging task due to the spatially variant blur caused by camera shake, object motions, and depth variations, etc. However, for the blurred area in the current video frame, the corresponding pixels of its neighboring video frames are often clear. Based on this observation, we propose an Adaptive Spatio-Temporal Convolutional Network (ASTCN) to compensate for blurry pixels in the current frame by using clear pixels in adjacent frames. In order to use the spatial information of adjacent frames in the current frame, the video frames must be aligned first. Existing methods usually estimate optical flow in the blurry video to align consecutive frames. However, they tend to generate artifacts when the estimated optical flow is not accurate. In order to overcome the limitations of optical flow estimation, we use deformable convolution in ASTCN to complete multi-scale adjacent frame alignment at the feature level. Secondly, we propose an adaptive spatio-temporal feature fusion module based on dynamic filters, which uses the features of the clear regions of adjacent frames to perform adaptive feature transformation on the intermediate frame to remove the blur. Extensive experimental results show that the proposed algorithm has shown superior performance on the benchmark datasets as well as real-world videos.

Keywords: Video deblurring · Pixel quality compensation · Dynamic filter

1 Introduction

1.1 Research Background

Video deblurring is a basic problem in the field of computer vision, which aims to recover clear frames from blurred video sequences. In recent years, with the explosion of short videos and the widespread popularity of handheld and airborne video capture devices, this issue has received active attention and research from related professionals. The blur in the video is usually caused by target motion, camera shake, and depth changes. This kind of blur is a spatial change blur, that is, different frames in a video sequence and different areas of the same frame have different degrees of blur. Therefore, the use of a globally unified fuzzy kernel cannot eliminate this spatial change blur. Blurred video will not only lead to the degradation of the human visual sensory experience, but also hinder some advanced visual tasks, such as target detection, visual tracking and SLAM. Therefore, it is very meaningful to study an effective video deblurring algorithm.

Furthermore, in order to allow non-professionals to use the algorithm and improve the portability and ease of use of the algorithm, it is necessary to develop a video deblurring system based on the algorithm.

1.2 Image Deblurring

The task of image deblurring is to output the restored clear image for the input blurred image. However, due to different reasons and different types of image blur in the real world, it is difficult to use a unified model or algorithm to deal with all blurs. Recent algorithms try to deal with the spatial change blur caused by camera shake, high-speed target movement, and depth changes in dynamic scenes. Most of these methods are based on the following image degradation models:

$$B = L \otimes K + N \quad (1)$$

B , L , and N represent blurred image, clear image and additional noise, respectively, and K is the fuzzy kernel, which represents the convolution operation.

Finding the blur kernel for each pixel is a highly ill-conditioned problem. Traditional methods try to model the fuzzy model by making a priori assumptions about the fuzzy source. [1, 2] assume that the blur is only caused by 3D camera movement. However, in dynamic scenes, due to the presence of multiple moving objects and camera movement, kernel estimation is more challenging. [3] proposed a dynamic scene deblurring method. However, the estimation of these blur kernels is still inaccurate, especially in the case of sudden object movement and severe camera shake. This erroneous blur kernel estimation directly affects the quality of the latent image, resulting in ringing artifacts in the restored image.

In recent years, CNN has been applied to image processing problems and has shown good results [4–7]. Since there is no real blurred image in the real world and the clear image of its corresponding label can be used for supervised learning, related researchers usually use the fuzzy kernel to convolve the clear image to generate the corresponding blurred image. [4, 5, 7] use blurry images with unified blur kernel synthesis for training. Since the advent of image deblurring data set [8, 9], by directly estimating the deblurring output without kernel estimation, an end-to-end learning method [8, 10, 11] has been proposed. These end-to-end methods only need to input the blurred image, and after the model is processed, the final clear image after deblurring can be obtained.

1.3 Video Deblurring

Early research on video deblurring only regarded it as a simple extension of image deblurring, and the redundant information between video frames was not fully utilized. Recent research divides the video restoration task into four parts: feature extraction, adjacent frame alignment, feature fusion, and feature reconstruction through a more complex process, making full use of the spatial information of the video between frames. Recent methods have made significant progress due to the use of clearer areas from adjacent frames [12, 13] or optical flow from consecutive frames [14, 15] to compensate the quality of intermediate video frames. However, directly using the clear area of adjacent

frames to compensate for the area corresponding to the intermediate frame usually produces obvious artifacts because the adjacent video frames are not completely aligned. Most existing methods [16–18] perform alignment by explicitly estimating the optical flow field between the reference frame and its adjacent frames. However, the optical flow estimation has a large amount of calculation and is difficult to be accurate, which often causes artifacts in the reference frame.

In recent years, with the successful application of CNN in the field of image processing, video deblurring methods based on deep neural networks have been extensively studied by relevant personnel. These methods utilize the timing relationship between input frames in various ways. DBN [9] stacks 5 consecutive frames in the channel dimension, and the convolutional neural network aggregates spatiotemporal information between adjacent frames. KIM [19] et al. used a deep loop network to connect multiple frame features to restore the current image. But these methods all perform adjacent frame alignment through optical flow estimation. Another research area achieves implicit motion compensation through dynamic filtering [20] or deformable convolution [21], which achieves a better alignment effect. It can perform the alignment of adjacent frames well without requiring a large amount of calculation, and the accuracy is higher than that of the optical flow estimation method.

1.4 ASTCN

To overcome the above limitations, we propose an Adaptive Spatio-Temporal Convolutional Network (ASTCN) for video deblurring. ASTCN includes three modules: a multi-scale deformable convolution alignment module, an adaptive spatiotemporal feature fusion module, and a feature reconstruction module. The input of the model is N consecutive video frames, and the output is the reconstructed intermediate frame.

The main contributions are summarized as follows:

- We propose a multi-scale adjacent frame alignment module based on deformable convolution, which overcomes the large and inaccurate problems of traditional optical flow estimation for alignment.
- We propose an adaptive spatiotemporal feature fusion module based on dynamic filters, which performs pixel-level feature transformation on the intermediate frame.
- We quantitatively and qualitatively evaluated our network on the benchmark dataset and proved its superior performance.

2 Related Work

2.1 Single-Image Deblurring

Recently, researchers have proposed many end-to-end CNN models [8, 10, 23, 24] for image deblurring. In order to obtain a large receptive field for processing large blur, a multi-scale strategy is used in the literature [8, 10]. In order to deal with the non-uniform blur of the dynamic scene, Zhang et al. used the spatial variant RNN [25] to eliminate the blur by using the RNN weights generated by the neural network in the feature space.

In order to generate clear images with more details, the adversarial loss is used in the literature [8, 24] to train the network, so that the restored images are more in line with human visual perception.

2.2 Video Deblurring

Unlike single image deblurring, video deblurring methods can utilize redundant spatial information that exists across adjacent frames. Therefore, CNN-based methods usually use multiple consecutive frames in the video sequence as model input to restore the intermediate video frame.

[26] proposed a quality enhancement method MFQE for compressed video. The core idea is to compensate poor quality (blurred) frames with good quality (clear) frames: First, MFQE uses a classifier to find the good quality frames in the video. For each bad frame, use its adjacent two frames. A good frame for quality enhancement. Before the quality is enhanced, the two adjacent good frames need to be motion compensated to align to the state of the bad frame at the moment. This method achieves better performance and improves the recovery effect of poor quality frames in a video sequence. But there are two problems at the same time: (1) MEQE relies on accurate motion compensation (optical flow estimation). If the motion compensation is not accurate, subsequent quality enhancement methods will fail. (2) Different adjacent frames and different areas of the same video frame have different degrees of blur, so while MFQE uses good quality frames to make up for poor quality frames, it may also introduce blurred areas in good quality frames. Therefore, good quality areas in adjacent frames should be used at the pixel level to compensate for the corresponding poor quality areas in the current frame.

In the NTIRE 2019 Video Enhancement Challenge, [27] proposed the EDVR algorithm. To compensate for the quality of adjacent frames, the EDVR algorithm uses a spatio-temporal attention fusion module to assign pixel-level aggregation weights on each video frame, so that the same area of different adjacent frames and different areas of the same video frame contribute to the recovery of intermediate frames. All are different. However, the algorithm has too many parameters and a large amount of calculation, which leads to a long model reasoning time, which is difficult to apply to actual production.

STFAN [22] proposed a filter adaptive convolution layer, which applies an element-wise convolution kernel to the video frame that needs to be restored, and it adaptively transforms the features of the video frame at the pixel level according to the input. Adjacent frame alignment and deblurring are both spatial mutation tasks, because the blur in a dynamic scene is non-uniform blur. STFAN uses the proposed filter adaptive convolution layer, and regards the two processes of alignment and deblurring as two filter adaptive convolutions in the feature domain. STFAN integrates adjacent frame alignment and deblurring into the same frame without the need for displayed motion estimation.

3 Proposed Algorithm

In this section, we first give an overview of our algorithm in Sec. 3.1. Then the three components of the algorithm are explained in Sect. 3.2–3.4: multi-scale deformable convolution alignment module, adaptive spatiotemporal feature fusion module and feature reconstruction module.

3.1 Overview

As shown in Fig. 1, ASTCN consists of three parts: a multi-scale deformable convolution alignment module, an adaptive spatio-temporal feature fusion module and a feature reconstruction module. The model uses five consecutive frames as input to recover a clear intermediate frame. The intermediate frame of five consecutive frames is represented as a reference frame, and the remaining frames are represented as adjacent frames.

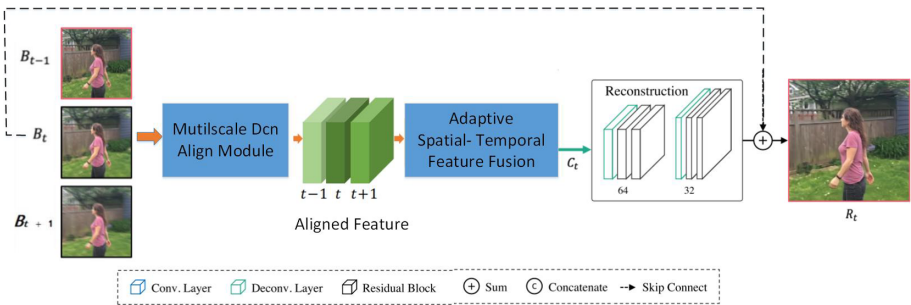


Fig. 1. Proposed network structure.

3.2 Multi-scale Deformable Convolution Alignment Module

Since the adjacent video frames in the video sequence have a certain amount of jitter, it is necessary to align the adjacent frames before the spatio-temporal feature fusion in order to use the clear areas in the adjacent frames to make up for the corresponding fuzzy areas in the reference frame. Compensate to the state of the reference frame at the moment. Traditional alignment methods are based on motion estimation, such as optical flow estimation. This type of method requires a large amount of calculation, and the optical flow estimation is difficult to be accurate, especially when there is a large jitter between adjacent video frames. Inaccurate motion estimation may introduce artifacts in the recovered video frames, seriously affecting the performance of the network.

Because there are different degrees of jitter between different adjacent frames and reference frames, and CNN has limitations in geometric transformation modeling due to its fixed convolution kernel configuration, so it is impossible to use CNN to handle this different degree of jitter. [15] proposed a deformable convolution operation to enhance the geometric transformation modeling capabilities of conventional CNN. Later [21] applied the deformable convolutional network to the task of video super-resolution

reconstruction to align adjacent frames. Inspired by [21, 29], we proposed a multi-scale adjacent frame alignment module based on deformable convolutional network as shown in Fig. 2.

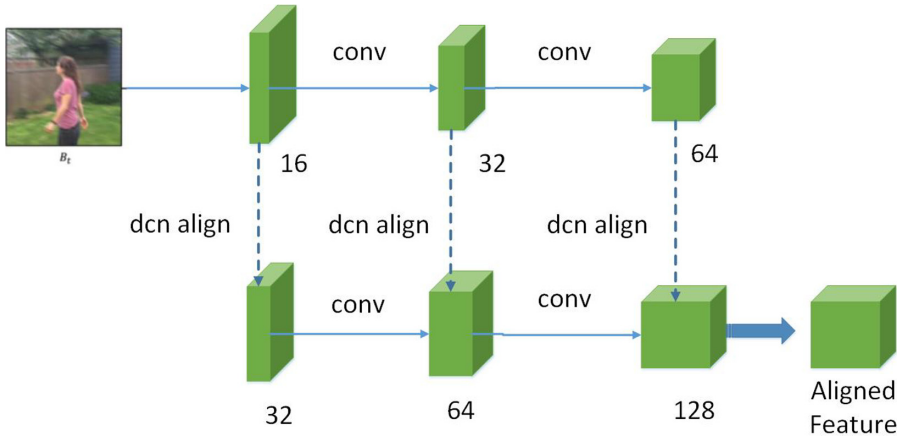


Fig. 2. Multi-scale deformable convolution alignment module

We use a multi-scale strategy to align adjacent video frames from coarse to fine. The calculation process of deformable convolution to perform alignment is as follows:

$$Ft + i(x, y) = \sum_{k=1}^K wk \bullet Ft + i(p + pk + offsetk) \bullet maskk \tag{2}$$

Where mask is the spatial mask of size $H \times W \times 1$. Spatial mask is obtained by multiplying the occlusion mask and the pixel quality mask. The value of the mask is a hard value of 0 or 1. 0 represents the occluded area in adjacent frames which has no effect on compensating for intermediate frames. The pixel quality mask reflects the clarity of pixels at different spatial positions of adjacent video frames. Its value is soft and lies between 0 and 1. The higher the value, the better the pixel quality. Spatial mask realizes pixel compensation in the spatial domain and solves the occlusion problem of adjacent video frames. Finally, we get the aligned features as the input of the adaptive spatiotemporal feature fusion module.

3.3 Adaptive Spatio-Temporal Feature Fusion Module

In the video deblurring task, alignment and feature fusion are the two most critical parts. How to efficiently fuse the spatio-temporal features of the aligned consecutive frames is the core of this part. Because the blur in the dynamic scene is non-uniform blur, the degree of blur is different at different spatial positions.

Motivated by the Kernel Prediction Network (KPN) [28], which applies the generated spatially variant filters to the input image, we propose the adaptive spatio-temporal

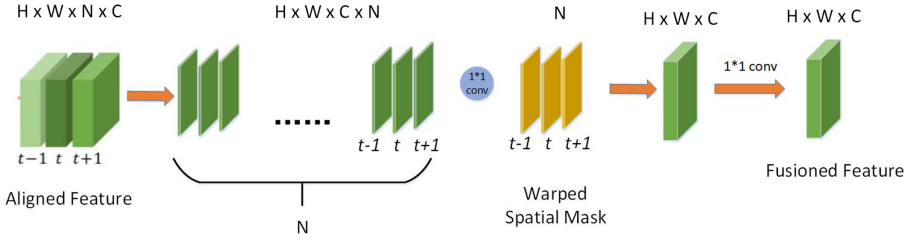


Fig. 3. Adaptive spatio-temporal feature fusion module

feature fusion module which applies generated element-wise convolutional filters to the aligned features, as shown in Fig. 3.

This module conforms to the dynamic filter architecture and includes two branches: filter generation module and dynamic filter convolution module. The filter generation module warps the Spatial Mask of each video frame with optical flow to get the warped spatial mask. The dynamic filter convolution module performs convolution operations on the generated warped spatial mask and aligned feature. So the convolution kernel here has no parameters to learn, but the intermediate calculation results. The parameters that the model needs to learn are greatly reduced, and the model inference speed is accelerated. The last 1×1 convolution is an ordinary convolution operation, the purpose is to merge the feature information of different channels on the features of the transformed intermediate frame. Finally, we get the fused feature map of intermediate frame.

3.4 Feature Reconstruction Module

The feature reconstruction network is responsible for reconstructing the features transformed by the spatio-temporal feature fusion module into a residual image with the same resolution as the intermediate frame, and finally adding the reconstructed residual image and the intermediate frame to obtain a clear frame that is restored.

The reconstruction network is composed of two deconvolution modules, and each deconvolution module is composed of a deconvolution layer and two residual blocks. The deconvolution layer up-samples the input features, and the residual module can well prevent the network from disappearing due to the depth of the gradient. The deformable convolution alignment module and the feature reconstruction network as a whole are an Encoder-Decoder structure, which itself is conducive to image enhancement tasks such as image denoising.

4 Experiments

4.1 Implementation Details

Dataset. In our experiments, we train the proposed network using the video deblurring dataset from [9]. It contains 71 videos (6,708 blurry-sharp pairs), splitting into 61 training videos (5,708 pairs) and 10 testing videos (1,000 pairs).

Data Augmentation. We perform several data augmentations for training. We first divide each video into several sequences with length 5. To add motion diversity into the training data, we reverse the order of sequence randomly. For each sequence, we perform the same image transformations. It consists of chromatic transformations such as brightness, contrast as well as saturation, which are uniformly sampled from [0.8, 1.2] and geometric transformations including randomly flipping horizontally and vertically and randomly cropping to 256×256 patches. To make our network robust in real-world scenarios, a Gaussian random noise from $N(0, 0.01)$ is added to the input images.

Loss Function. To effectively train the proposed network, we use the mean squared error (MSE) loss that measures the differences between the restored frame R and its corresponding sharp ground truth S :

$$\mathcal{L}_{mse} = \frac{1}{CHW} \|R - S\|^2, \quad (3)$$

where C, H, W are dimensions of image, respectively; R and S respectively denote the restored image and the corresponding ground truth.

Experimental Settings. We use the Xavier initialization method to initialize the ASTCN network and use the Adam optimizer to train it, where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Since the alignment module is unstable in the process of learning feature offset, the initial learning rate of the multi-scale deformable convolution alignment module is set to a small $1e-6$, and the initial learning rate of the remaining modules is set to $1e-4$, and the entire network is trained end-to-end together. The learning rate of all modules decays by 0.1 times every 400 k iterations, and the ASTCN network finally converges after 850 k iterations. We use the Pytorch framework to implement the network, and train and test it on NVIDIA GeForce 2080Ti.

4.2 Experimental Results

Quantitative Evaluations. In order to evaluate the performance of the ASTCN algorithm, we compared it with the video deblurring algorithm in recent years on the DVD [9] data set (the running results of the remaining algorithms are from the original paper). In the experiment, PSNR and SSIM are used as evaluation indicators, and Time (sec) represents the time required for each algorithm to process a video frame. Params (M) is the parameter quantity of the model, and its unit is millions. PSNR and SSIM reflect the accuracy of each algorithm, and Time (sec) and Params (M) reflect the efficiency of the algorithm. It can be seen from Table 1 that ASTCN is superior to the latest algorithm Pan [30] in terms of performance and efficiency. Although Kupyn [24] and STFAN [22] have less processing time and model parameters, their deblurring effect is far worse than that of ASTCN. It can be said that this algorithm takes into account both performance and efficiency, and can efficiently process blurred videos.

Qualitative Evaluations. In order to further verify the generalization ability of the ASTCN algorithm, we conducted a qualitative evaluation on the real fuzzy video

Table 1. Quantitative evaluation on the video deblurring dataset [36], in terms of PSNR, SSIM, running time (sec) and parameter numbers ($\times 10^6$) of different networks.

Method	Kupyn [24]	Nah [8]	Su [9]	STFAN [22]	Pan [30]	Ours
PSNR	26.78	29.51	30.05	31.15	32.13	32.15
SSIM	0.848	0.912	0.920	0.905	0.927	0.928
Time(sec)	0.22	4.78	6.88	0.15	0.87	0.53
Params(M)	11.38	11.38	16.67	5.37	16.19	13.16



Fig. 4. Unprocessed real blurred image

sequence in the DVD [9] dataset. As shown in the Figs. 4, 5, 6, ASTCN recovers the detailed information in the blurred image, which can powerfully handle the unknown real blur in the dynamic scene, which further proves the superiority of the algorithm.



Fig. 5. Image processed by ASTCN



Fig. 6. Blurred image (left) and deblurred image (right) in the same area detail comparison

5 Conclusion

We have proposed a novel adaptive spatio-temporal network for video deblurring based on deformable convolution and dynamic filters. The entire model is end-to-end, and you only need to input a blurred video sequence to get the deblurred video. Our proposed multi-scale deformable convolution alignment module can perform the alignment of adjacent frames well without displaying the motion estimation, and solves the problem of large and inaccurate alignment of traditional optical flow estimation. The adaptive spatio-temporal feature fusion module adaptively performs pixel-level feature transformation on the reference frame according to the input continuous video frames, which greatly improves the efficiency of feature fusion and shortens the model inference time. The experimental results demonstrate the effectiveness of the proposed method in terms of accuracy, speed as well as model size.

References

1. Gupta, A., Joshi, N., Zitnick, C.L., et al.: Single image deblurring using motion density functions. In: European Conference on Computer Vision, vol. 68, no. 1, pp. 562-573 (2010). https://doi.org/10.1007/978-3-642-15549-9_13

2. Whyte, O., Sivic, J., Zisserman, A., Ponce, J.: Non-uniform deblurring for shaken images. In: *International Journal of Computer Vision*, vol. 98, no. 2, pp. 168–186 (2012)
3. Kim, T.H., Ahn, B., Lee, K.M.: Dynamic scene deblurring. In: *International Conference on Computer Vision*, vol. 53, no. 2, 065–074 (2013)
4. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. In: *Advances in Neural Information Processing Systems*, vol. 46, no. 1, pp. 1790–1798 (2014)
5. Schuler, C.J., Hirsch, M., Harmeling, S., Schölkopf, B.: Learning to deblur. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1439–1451 (2016)
6. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 63, no. 1, pp. 769–777 (2015)
7. Leibe, B., Matas, J., Sebe, N., et al.: A neural approach to blind motion deblurring. 221–235 (2016). [https://doi.org/10.1007/978-3-319-46487-9\(Chapter 14\)](https://doi.org/10.1007/978-3-319-46487-9(Chapter 14))
8. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 13, no. 1, pp. 623–634 (2017)
9. Su, S., Delbracio, M., Wang, J., et al.: Deep video deblurring for hand-held cameras. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 86, no. 7, pp. 130–139 (2017)
10. Tao, X., Gao, H., Wang, Y., et al.: Scale-recurrent network for deep image deblurring. 20(4), 231–240 (2018)
11. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, vol. 35, no. 8, pp. 302–312 (2018)
12. Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.-Y.: Full-frame video stabilization with motion inpainting. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(7), 1150–1163 (2006)
13. Cho, S., Wang, J., Lee, S.: Video deblurring for hand-held cameras using patch-based synthesis. *ACM Trans. Graph.* **31**(4), 64 (2012)
14. Hyun Kim, T., Mu Lee, K.: Generalized video deblurring for dynamic scenes. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 16, no. 1, pp. 323–334 (2012)
15. Ren, W., Pan, J., Cao, X., Yang, M-H.: Video deblurring via semantic segmentation and pixelwise non-linear kernel. In: *International Conference on Computer Vision*, vol. 12, no. 3, pp. 32–41 (2017)
16. Caballero, J., et al.: Realtime video super-resolution with spatio-temporal networks and motion compensation. In: *The IEEE Conference on Computer Vision and Pattern Recognition*, vol. 27, no. 2, 372–382 (2018)
17. Xue, T., Chen, B., Wu, J., et al.: Video enhancement with task-oriented flow. *Int. J. Comput. Vision* **10**(3), 23–32 (2017)
18. Kim, T.H., Sajjadi, M.S.M., Hirsch, M., Schölkopf, B.: Spatio-temporal transformer network for video restoration. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11207, pp. 111–127. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_7
19. Kim, T.H., Mu Lee, K., Schölkopf, B., Hirsch, M.: Online video deblurring via dynamic temporal blending network. In: *Conference on Computer Vision and Pattern Recognition*, vol. 45, no. 2, pp. 120–129 (2017)
20. Jo, Y., Oh, S.W., Kang, J., et al.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: *IEEE*, vol. 13, no. 1, pp. 13–21 (2018)
21. Tian, Y., Zhang, Y., Fu, Y., et al.: TDAN: temporally deformable alignment network for video super-resolution. 13(1), 23–33 (2018)

22. Zhou, S., Zhang, J., Pan, J., et al.: Spatio-temporal filter adaptive network for video deblurring. 35(1), 36–47 (2019)
23. Zhang, J., et al.: Dynamic scene deblurring using spatially variant recurrent neural networks. In: Conference on Computer Vision and Pattern Recognition, vol. 15, no. 1, pp. 78–89 (2018)
24. Kupyn, O., Budzan, V., Mishkin, D., Matas, J.: Deblurgan: blind motion deblurring using conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition, vol. 25, no. 2, pp. 78–89 (2018)
25. Liu, S., Pan, J., Yang, M.-H.: Learning recursive filters for low-level vision via a hybrid neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 560–576. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_34
26. Yang, R., Xu, M., Wang, Z., et al.: Multi-frame quality enhancement for compressed video. In: IEEE vol. 19, no. 2, 765–773 (2018)
27. Wang, X., Chan, K.C.K., Yu, K., et al.: EDVR: video restoration with enhanced deformable convolutional networks. 21(5), 5563–5575 (2019)
28. De Brabandere, B., Jia, X.: Dynamic filter networks. 23(2), 980–993 (2016)
29. Dai, J., Qi, H., Xiong, Y., et al.: Deformable convolutional networks. 27(2), 213–224 (2017)
30. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. 18(2), 1534–1542 (2020)