

# Recurrence-in-Recurrence Networks for Video Deblurring

Joonkyu Park  
jkpark0825@snu.ac.kr

Seungjun Nah  
seungjun.nah@gmail.com

Kyoung Mu Lee  
kyoungmu@snu.ac.kr

Department of ECE, ASRI  
Seoul National University  
Korea

---

## Abstract

State-of-the-art video deblurring methods often adopt recurrent neural networks to model the temporal dependency between the frames. While the hidden states play key role in delivering information to the next frame, abrupt motion blur tend to weaken the relevance in the neighbor frames. In this paper, we propose recurrence-in-recurrence network architecture to cope with the limitations of short-ranged memory. We employ additional recurrent units inside the RNN cell. First, we employ inner-recurrence module (IRM) to manage the long-ranged dependency in a sequence. IRM learns to keep track of the cell memory and provides complementary information to find the deblurred frames. Second, we adopt an attention-based temporal blending strategy to extract the necessary part of the information in the local neighborhood. The adaptive temporal blending (ATB) can either attenuate or amplify the features by the spatial attention. Our extensive experimental results and analysis validate the effectiveness of IRM and ATB on various RNN architectures.

## 1 Introduction

Videos often suffer from motion blur caused by the relative motions between the camera and the objects. From the motion blur that varies both spatially and temporally, recovering the sharp and rich details is a challenging task. Such complex blur makes the scene hard to be identified, hindering the following scene recognition algorithms to be applied in practice. As motion blur is one of the most common artifacts in videos, many efforts were made to address the video deblurring problem.

From the temporal variation of scenes, conventional video deblurring methods tried to extract the motion information to remove it from the scenes. Motion flow was used to infer the blur trajectories in the frames [6, 8] in the joint optimization process with the latent frames. In [9, 17, 20], the displacement of the neighboring frames are modeled by optical flow to help the learning of the following neural network models. The frames are aligned by warping from the estimated flow so that the relevant information could be better aggregated



Figure 1: Visual comparison of RNN-based video deblurring results (top) and our RIRN-applied results (bottom). RIRN recovers the image details where the baseline methods fail.

in the learning process. However, trying to find such point-to-point correspondences from blurry frames are prone to be erroneous and cause misalignment errors.

Recurrent neural networks, on the other hand, try to handle the temporal propagation of scenes in an implicit manner. Rather than explicitly finding the motion flow, the hidden states convey the information from the past frames to the future frames. Thus, the way to handle hidden states has played key roles in designing recurrent networks for video deblurring [0, 16, 24, 29, 30].

As the hidden states are obtained from the past frame but blindly to the current frame, the hidden state should be carefully used with the input frames. While [24] designed a multi-scale architecture to deliver hierarchical information, [0, 16] modified the hidden states adaptively to the input in order to better focus on the target frame. However, motion-blurred sequences often suffer from abrupt change of scenes that loosens the correlation between the adjacent frames. Extending the neighborhood from the single previous frame to a predetermined range of past and future frames, [29] keeps a set of features and uses them to deblur the center frame. Such an approach could look wider to find more relevant information from the saved features but requires extra RAM proportional to the size of the neighborhood. Also, the optimal number of frames to look could vary by the local blur dynamics and scene contents.

Instead of saving multiple features or hidden states, we propose to model the long-range information of video in an extra memory with our recurrence-in-recurrence network (RIRN). On top of an RNN architecture, we design an inner-recurrence module (IRM) that generates an auxiliary state with complementary information from the hidden state. Different from [29], we do not drop the oldest memory but let IRM learn to keep necessary information. Also, we propose an adaptive temporal blending (ATB) method that finds the relevant part of the information from the local neighborhood. ATB generates attention maps on the image features from the current and the previous frames. While we adopt the idea of temporal blending from [0], we relax the attention constraint so that features could be either attenuated or emphasized by the necessity in reconstructing the deblurred image via learning.

By conducting ablation study with a baseline method, we analyze the behavior of the proposed IRM and ATB and show the effectiveness of the architectural designs. Furthermore, we apply the RIRN architecture to various RNN-based video deblurring methods as well as the gated architectures, LSTM [5] and GRU [0] that manipulates the update of hidden states. Our extensive experimental results exhibit the consistent improvements from our method both quantitatively and qualitatively.

## 2 Related Works

**Video Deblurring with Explicit Motion Modeling.** While single image deblurring methods typically investigate the spatial information to handle the intrinsic blur trajectories [4, 24, 19, 22, 27], early studies on video deblurring tried to aggregate the sharper information from the temporally neighboring frames [2, 13] by finding relevant patches. To cope with complex blurry scenes, most of the optimization-based approaches estimated the blur kernel and used it to find the deblurred frames [6, 8, 12, 25, 26] through the joint optimization process. However, finding spatially non-uniform kernel for every pixel requires heavy computational resources. In [20], learning-based method was introduced by feeding a set of frames into CNN to deblur the center frame. However, optical flow had to be computed from the blurry input frames to align them in the preprocessing. Due to the blurriness, it is difficult to find the pixel-level correspondences and the error could cause inaccuracies in restoring the frames. In [9], spatio-temporal flow estimation was adopted to selectively capture the temporal dependencies while alleviating the occlusion problems from optical flow. Later, [17] proposed to improve the optical flow accuracy by estimating the flow and the latent frames simultaneously with neural networks using temporal sharpness prior. In a recent work [18], the proposed ARVo method first uses an optical flow to align the frames. Then, they construct a correlation volume pyramid among all pixel pairs in the neighboring frames to aggregate relevant information.

**Video Deblurring with Implicit Motion Modeling.** On the other hand, the motion information is implicitly employed in the approaches using recurrent neural networks. Proposing a recurrent neural network architecture, RDN [24] employed temporal skip connections in multiple feature scales to transfer information to the next frames. In OVD [7], the intermediate features were computed with dynamic temporal blending to improve the deblurring performance. As such hidden states are generated from the past frames without knowing the target blurry frame, IFI-RNN used intra-frame iteration using the recurrence architecture to adapt the hidden states to better remove blur in the frame. In a different approach, STFAN [30] proposed a filter-adaptive network to align the features from multiple inputs within the neural network architecture from learning. While [7, 16, 30] tried to better use the given hidden states generated from the previous time steps, ESTRNN [29] proposed to use the features from both the past and the future frames. From a set of features that are kept in memory, the spatio-temporal attention module fused the features to reconstruct the latent frame. In recently proposed [21], attention-based aggregation modules were proposed. From the reinforcement learning-based keyframe batch selection, the gated spatio-temporal attention block uses non-local attention to gather sharper information. By the attention-based aggregation process, the more useful information can be adaptively selected both spatially and temporally to better deblur a frame. In contrast to [29] or [21], our recurrence-in-recurrence architecture does not necessarily require future information or cached history of features. We let our inner recurrence module to keep track of the temporal variance of hidden states from learning.

**Long-Term Dependency Modeling in Sequential Data.** LSTM [5] is one of the earliest and widely-used architecture by solving vanishing gradient problem of RNNs. The gates in LSTM determines the information to be forgotten in the hidden states. GRU [10] presents an architecture with the reset and the update gates to manipulate hidden states. While the gates are meant to model the long-term dependencies in sequential data, every information is saved in a single state which may eventually lead to short-term optimization [28]. Transformer architectures [23] instead, stores the hidden activation at every time step and uses an

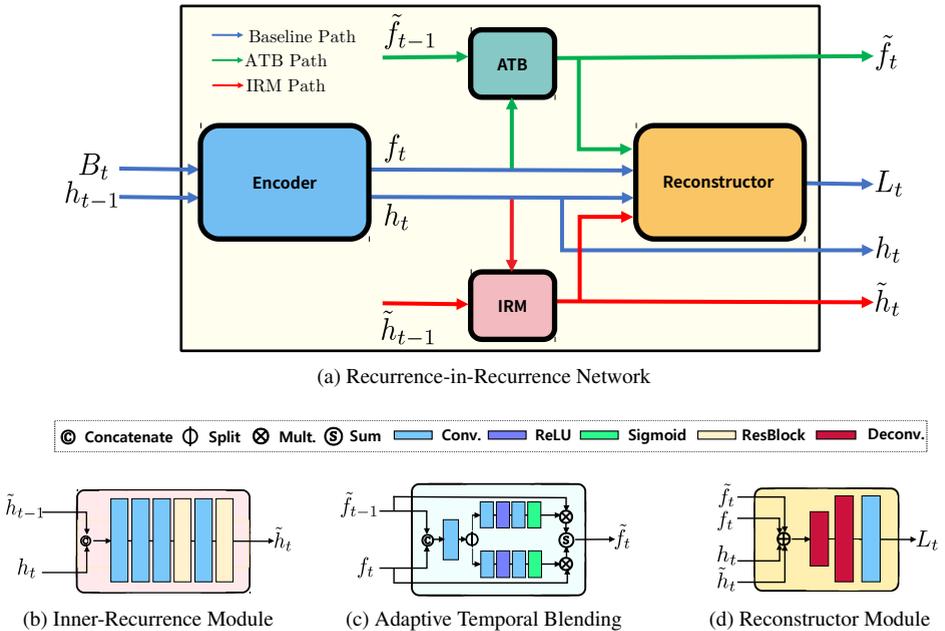


Figure 2: The architecture of Recurrence-in-Recurrence Network and the components

attention module to integrate them, requiring large memory footprint. To model longer-range dependencies in language modeling, Transformer-XL [9] reuses the states in segment-level recurrences, however, sufficiently old states are discarded determined by the size of memory. Recently, [18] introduced compressed representation to preserve old memories instead of discarding or saving in raw forms. They extended the baseline transformer with an additional compression function. Our recurrence-in-recurrence architecture shares a similar idea with [18] in terms of augmenting the short-term memory from an auxiliary module generating additional hidden states.

## 3 Proposed Method

The main goal of this paper is to improve the video deblurring quality of RNN-based methods by supplementing the hidden states by providing an additional long-ranged memory. Starting from the baseline RNN structure, we introduce an inner-recurrence module and the adaptive temporal blending scheme to better handle the temporal dependency. The overall architecture is shown in Figure 2.

### 3.1 Video Deblurring with Recurrent Networks

We design a baseline recurrent network architecture with a feature extractor and a image reconstructor module. From the input blurry frame  $B_t$  and the hidden state  $h_{t-1}$  from the previous time step, the feature encoder emits an updated hidden state  $h_t$  and a feature  $f_t$  as

$$f_t, h_t = \text{Encoder}(B_t, h_{t-1}). \quad (1)$$

Then, the outputs from the encoder is used to reconstruct the deblurred image  $L_t$  as

$$L_t = \text{Reconstructor}(f_t, h_t). \quad (2)$$

The recurrent model is trained with supervised loss function by comparing the output  $L_t$  with the ground-truth sharp image  $S_t$  as  $\|L_t - S_t\|$ .

In our proposed recurrence-in-recurrence network, the inner-recurrence module and adaptive temporal blending each supplement the hidden state  $h_t$  and the feature  $f_t$  to better deblur video frames.

## 3.2 Inner-Recurrence Module

RNNs basically use a single hidden state at each time step to store information from the past frames. The hidden state at each step is optimized to maximize the deblurring performance of the corresponding frame. At every time step, less relevant information to the target frame is forgotten and replaced by new information from the input. However, dynamic videos tend to suffer from abrupt scene changes (i.e. camera shakes) so that long-ranged distant frames are often more related than direct neighbors. In order to find and provide complementary information with long-term relation, we further exploit the recurrence operation to model the temporal variations of the hidden states. Motivated that the set of hidden states,  $\{h_t\}$ , is another sequential data, we propose to design a new type of RNN whose inputs are the hidden states. As the recurrence operates inside the standard architecture of RNNs, we term the module as Inner-Recurrence Module (IRM) where

$$\tilde{h}_t = \text{IRM}(h_t, \tilde{h}_{t-1}). \quad (3)$$

The outputs of IRM,  $\{\tilde{h}_t\}$  is generated by looking into the history and the temporal changes of  $\{h_t\}$ , serving as a sequence of higher-order memory states. In contrast to saving multiple features in [29], IRM does not require additional memory to store the cached set of states at inference. Similarly to Compressive Transformer [18] storing the compressed memory from the memory sequence, our IRM preserves long-ranged information from learning. In the following image reconstructor module,  $\tilde{h}_t$  is jointly used with  $h_t$  together, supplementing the deblurring performance. The IRM architecture is shown in Figure 2b.

## 3.3 Adaptive Temporal Blending

Besides using the hidden state  $h_{t-1}$  to deliver information, the previous input frame  $B_{t-1}$  is often used as well as  $B_t$  in video deblurring RNNs [29, 30] by concatenating the features. Instead, we choose to adaptively blend the features with an attention-based recurrent module, inspired by dynamic temporal blending [7]. Different from the IRM that manages the long-range dependency of frames, our Adaptive Temporal Blending (ATB) focuses on the feature  $f_t$  that is more specific to the target image at time  $t$  as it is not propagated to the next frames in the baseline architecture. DTB [2] assumes the features at all pixels to be equally important, forcing the sum-to-one constraints in the blending weights. However, as the image feature is used together with the hidden states, there could be redundancy causing spatial variance in feature importance. Thus, we adopted an attention mechanism so that the feature importance could be predicted from learning. Instead of updating the input feature directly via recurrence as IRM, ATB produces the attention map to extract necessary information in

Table 1: Effect of ATB applied to RDBN

Architecture	GOPRO		REDS		time(sec)
	PSNR	SSIM	PSNR	SSIM	
RDBN [14]	29.82	0.9043	32.29	0.9222	0.095
RDBN + GSA	30.10	0.9064	32.52	0.9233	0.170
RDBN + IRM	<b>30.14</b>	<b>0.9072</b>	<b>32.59</b>	<b>0.9304</b>	0.161

Table 2: Effect of IRM applied to IFI-RNN (C1H1)

Architecture	GOPRO		REDS		time(sec)
	PSNR	SSIM	PSNR	SSIM	
IFI-RNN [16]	28.30	0.8668	30.01	0.8762	0.049
IFI-RNN + DTB [16]	28.31	0.8697	30.12	0.8763	0.059
IFI-RNN + ATB	<b>28.65</b>	<b>0.8779</b>	<b>30.61</b>	<b>0.8800</b>	0.065

$f_t$ . The spatial attention maps are multiplied to the features as conical combination,

$$\tilde{f}_t = \tilde{w}_{t-1} \times \tilde{f}_{t-1} + w_t \times f_t, \quad (4)$$

where  $\tilde{w}_{t-1} \geq 0$  and  $w_t \geq 0$ . Different from DTB, ATB does not require sum-to-one constraint. We let the attention weights to adaptively attenuate or emphasize the corresponding features without being tied to each other. The design of ATB module is shown in Figure 2c.

### 3.4 Reconstructor

The reconstruction module aggregates all the features extracted from the encoder, inner-recurrence module, and the adaptive temporal blending module. With all features  $f_t$ ,  $h_t$ ,  $\tilde{f}_t$ ,  $\tilde{h}_t$  concatenated, the deblurred image is obtained from several convolutional layers. The reconstructor architecture is shown in Figure 2d.

## 4 Experimental Results

### 4.1 Experimental Setup

To train and validate the performance of every model, we used GOPRO [14], REDS [15], and DVD [10] datasets. DVD dataset contains 61 training sequences and 10 test sequences. GOPRO dataset provides 22 training and 11 test sequences and REDS dataset has 240 training and 30 validation sequences. All the video frames are in  $1280 \times 720$  resolution. To validate the effect of the proposed modules, every experiment was done in a unified setting, training from scratch with L1 loss and ADAM [17] optimizer with batch size 16. On GOPRO and REDS datasets, each model was trained for 500 and 200 epochs, respectively, annealing the learning rate at 300th and 100th epochs from the initial learning rate  $1 \times 10^{-4}$ . On DVD dataset, models were trained for 500 epochs and the learning rate was halved after every 200 epochs. For the model architectures used in the experiments and the implementation details, please refer to the supplementary material.

### 4.2 Ablation Study: Effect of IRM and ATB

We validate the effect of our proposed IRM and ATB by applying them to the baseline RNN architecture based on RDBN [14] and IFI-RNN [16]. The simplest form of IFI-RNN without additional iterations, C1H1 is used.

In Table 1, we show the effect of applying IRM on the baseline model on GOPRO and REDS datasets. Our IRM improves the deblurring accuracy by large margin at a similar degree of computing time. We compare the effect of IRM and GSA [14] by saving a set of features from different time steps. As we do not use future frames for deblurred frame estimation, we used 3 features,  $\{f_{t-2}, f_{t-1}, f_t\}$ , before updating the hidden state  $h_t$ . We find that IRM is not only more accurate than GSA but also faster in computation.

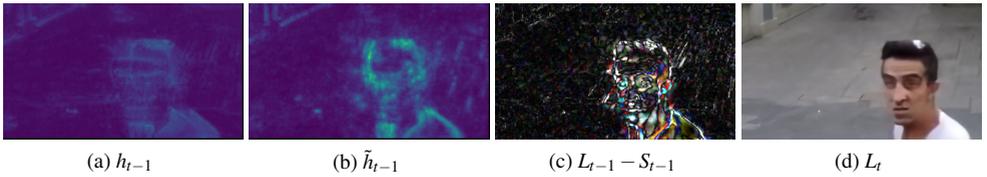


Figure 3: Visualization of hidden state and the IRM output

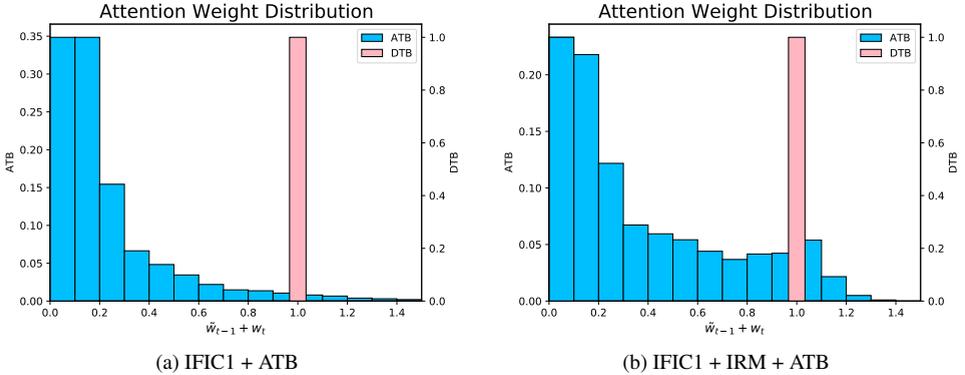


Figure 4: Distribution of sum of attention weights in ATB and DTB on GOPRO [14] dataset.

In Figure 3, we visualize the hidden state and the long-term memory obtained from the IRM as well as the error map of the deblurred image from the previous time step. While the facial texture requires detailed information, the hidden state  $h_{t-1}$  fails to serve as an informative cue due to the erroneous estimation in the previous frame as shown in Figure 3c. It shows that  $\tilde{h}_{t-1}$  from IRM brings a complementary information from the further past frames to help recover  $L_t$ .

In Table 2, we validate the effect of blending features by using ATB in the RNN architecture. Our ATB successfully improves the deblurring performance of the baseline while DTB provides marginal gains over the baseline. The different behavior of ATB and DTB in handling the features is shown in Figure 4. By removing the sum-to-one constraint of DTB in our adaptive temporal blending, we find that majority of the features are suppressed by the learned attention. It shows that our ATB could better find the useful constituents in the features while discarding the unnecessary information. Interestingly, when IRM is used jointly, ATB tends to makes more use of the  $f_t$  and  $\tilde{f}_{t-1}$  as shown in Figure 4b. We further visualize the effectiveness of our ATB in Figure 6 by showing the blended feature. While DTB fails to gather necessary information from blending in Figure 6c, ATB shows clearer edges in Figure 6d, helping the reconstruction from the walking person’s blurry legs. It indicates that our relaxed condition from removing the sum-to-one constraint leads to successful selection on necessary information from the spatial attention.

In Figure 5, we visualize the effect of IRM, ATB in the deblurred results. The carplate numbers are better recognizable from the results using both modules.



Figure 5: Visual ablation showing the effect of ATB and IRM on GOPRO [14] dataset.

Table 3: Application of RIRN on existing RNN architectures

Architecture	DVD		GOPRO		REDS		FPS
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
IFI-RNN [16]	30.53	0.9069	28.30	0.8668	30.01	0.8762	20.15
IFI-RNN + RIRN	<b>30.97</b>	<b>0.9168</b>	<b>29.14</b>	<b>0.8894</b>	<b>31.08</b>	<b>0.8905</b>	12.48
STRCNN [9]	29.15	0.8728	28.72	0.8460	30.23	0.8708	7.69
STRCNN + RIRN	<b>30.17</b>	<b>0.9019</b>	<b>28.87</b>	<b>0.8781</b>	<b>30.76</b>	<b>0.8902</b>	5.94
STFAN [30]	30.93	0.9087	28.77	0.8776	31.26	0.8864	6.66
STFAN + RIRN	<b>31.03</b>	<b>0.9096</b>	<b>29.24</b>	<b>0.8876</b>	<b>31.44</b>	<b>0.8951</b>	4.42
RDBN [29]	31.44	0.9188	29.82	0.9043	32.29	0.9222	10.50
RDBN + RIRN	<b>31.83</b>	<b>0.9227</b>	<b>30.17</b>	<b>0.9120</b>	<b>32.71</b>	<b>0.9322</b>	8.53
GRU	27.53	0.8335	25.11	0.7890	26.69	0.7956	22.24
GRU + RIRN	<b>28.58</b>	<b>0.8773</b>	<b>26.36</b>	<b>0.8217</b>	<b>28.60</b>	<b>0.8428</b>	20.94
LSTM	26.94	0.8365	25.22	0.7948	26.87	0.8046	19.14
LSTM + RIRN	<b>29.16</b>	<b>0.8813</b>	<b>27.24</b>	<b>0.8400</b>	<b>29.12</b>	<b>0.8584</b>	16.90

### 4.3 Effect of RIRN in State-of-The-Art Methods

Finding both the IRM and ATB to be effective in recurrent networks, we apply the recurrence-in-recurrence architecture to various video deblurring methods. In Table 3, we validate the generalizability of RIRN across different architectures by showing the consistent improvements from RIRN-applied models. We use IFI-RNN [16] C1H1 model, STRCNN [9], STFAN [30], RDBN. RDBN is derived from ESTRNN [29] but does not use the cached memory and GSA module. Also, we apply our RIRN to the gated RNN architectures such as LSTM and GRU, where the updates of hidden states are controlled by the gates. While the LSTM and GRU use the gates to control memory of RNNs, our RIRN successfully brings gains in deblurring accuracy from our IRN and ATB. In Figure 8, 9 and 7, we show the visual effect of our RIRN-applied results on various RNNs on GOPRO [14], REDS [15], and DVD [20] datasets. By applying RIRN, the textures are better deblurred and better recognizable compared with the results obtained from the baseline methods. We also show the generalizability of our RIRN on the real blurry videos [2] in Figure 10.

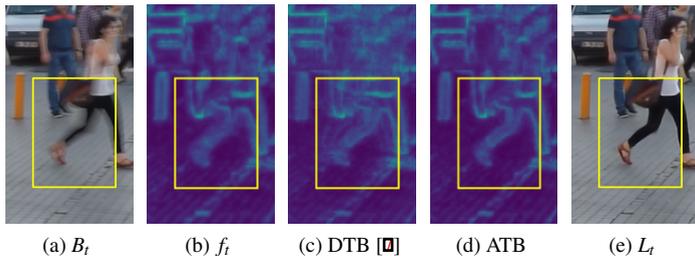


Figure 6: Visualization of  $\tilde{f}_t$  in baseline model(IFI-RNN) with using DTB and ATB.



Figure 7: Comparison of baselines and RIRN-applied method results in DVD [20] dataset

## 5 Conclusion

In this paper, we proposed a new method to improve the existing recurrent neural networks with our recurrence-in-recurrence network architecture. The recurrence-in-recurrence network consists of inner-recurrence module and the adaptive temporal blending method to augment the baseline recurrent networks. The inner-recurrence module learns to model the temporal variation of hidden states and provides a complementary information that is often unseen at the previous time step by looking into further past frames. The adaptive temporal blending of features can selectively extract necessary information and suppress unwanted part of the information. We show that our RIRN consistently improves the video deblurring performance on various RNN-based methods and exhibits state-of-the-art performance.

## 6 Acknowledgment

This work was supported by IITP grant funded by the Korea government(MSIT) [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]



Figure 8: Comparison of the baselines and the RIRN-applied results on GOPRO dataset



Figure 9: Comparison of the baselines and the RIRN-applied results on REDS dataset



Figure 10: Visual comparison of RNN-based video deblurring results (top) and our RIRN-applied results (bottom) on real blurry videos.

## References

- [1] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [2] Sunghyun Cho, Hojin Cho, Yu-Wing Tai, and Seungyong Lee. Registration based non-uniform motion deblurring. 31(7):2183–2192, 2012.
- [3] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.
- [4] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, 2019.
- [5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [6] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, 2015.
- [7] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Schölkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *ICCV*, 2017.
- [8] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. Dynamic video deblurring using a locally adaptive blur model. *IEEE TPAMI*, 40(10):2374–2387, 2017.
- [9] Tae Hyun Kim, Mehdi S. M. Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, 2018.
- [10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [11] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *CVPR*, 2021.
- [12] Yunpeng Li, Sing Bing Kang, Neel Joshi, Steve M Seitz, and Daniel P Huttenlocher. Generating sharp panoramas from motion-blurred videos. In *CVPR*, 2010.
- [13] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE TPAMI*, 28(7):1150–1163, 2006.
- [14] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.
- [15] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenges on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, 2019.

- [16] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*, 2019.
- [17] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*, 2020.
- [18] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *ICLR*, 2020.
- [19] Wenqi Ren, Jiawei Zhang, Jinshan Pan, Sifei Liu, Jimmy Ren, Junping Du, Xiaochun Cao, and Ming-Hsuan Yang. Deblurring dynamic scenes via spatially varying recurrent neural networks. *IEEE TPAMI*, 2021.
- [20] Shuo Chen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017.
- [21] Maitreya Suin and A. N. Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *CVPR*, 2021.
- [22] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018.
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- [24] Patrick Wieschollek, Michael Hirsch, Bernhard Schölkopf, and Hendrik P. A. Lensch. Learning blind motion deblurring. In *ICCV*, 2017.
- [25] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *ECCV*, 2014.
- [26] Haichao Zhang and Lawrence Carin. Multi-shot imaging: Joint alignment, deblurring and resolution-enhancement. In *CVPR*, 2014.
- [27] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson W.H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018.
- [28] Jingyu Zhao, Feiqing Huang, Jia Lv, Yanjie Duan, Zhen Qin, Guodong Li, and Guangjian Tian. Do rnn and lstm have long memory? In *ICML*, 2020.
- [29] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*, 2020.
- [30] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, 2019.