Event-driven Video Deblurring via Spatio-Temporal Relation-Aware Network

Chengzhi Cao , Xueyang Fu*, Yurui Zhu, Gege Shi and **Zheng-Jun Zha**
University of Science and Technology of China, China
chengzhicao@mail.ustc.edu.cn, xyfu@ustc.edu.cn, {zyr, sgg19990910}@mail.ustc.edu.cn, zhazj@ustc.edu.cn

Abstract

Video deblurring with event information has attracted considerable attention. To help deblur each frame, existing methods usually compress a specific event sequence into a feature tensor with the same size as the corresponding video. However, this strategy neither considers the pixel-level spatial brightness changes nor the temporal correlation between events at each time step, resulting in insufficient use of spatio-temporal information. To address this issue, we propose a new Spatio-Temporal Relation-Attention network (STRA), for the specific event-based video deblurring. Concretely, to utilize spatial consistency between the frame and event, we model the brightness changes as an extra prior to aware blurring contexts in each frame; to record temporal relationship among different events, we develop a temporal memory block to restore long-range dependencies of event sequences continuously. In this way, the complementary information contained in the events and frames, as well as the correlation of neighboring events, can be fully utilized to recover spatial texture from events constantly. Experiments show that our STRA significantly outperforms several competing methods, e.g., on the HQF dataset, our network achieves up to 1.3 dB in terms of PSNR over the most advanced method. The code is available at https://github.com/Chengzhi-Cao/STRA.

1 Introduction

As an important data source in the computer vision community, video usually contain inevitable blur due to movement of objects [Gallego *et al.*, 2021; Zou *et al.*, 2021]. To eliminate the adverse effects, video deblurring have attracted considerable attentions [Touvron *et al.*, 2021; Li and Xu, 2021]. Recently, a new sensor, called event camera for recording and capturing scene intensity changes at the microsecond level, has been recommended to promote video blurring [Zhu *et al.*, 2019].

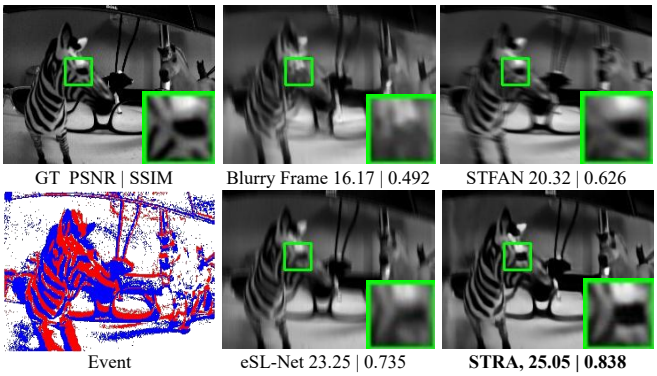


Figure 1: Quantitative and qualitative comparisons on deblurring results by state-of-the-art video deblurring STFAN [Zhou *et al.*, 2019], eSL-Net [Wang *et al.*, 2020] and our STRA.

Due to the success of convolutional neural networks (CNNs) [Wang *et al.*, 2021; Nikzad *et al.*, 2021], event-driven deblurring has been extensively developed and achieved promising performance [Xu *et al.*, 2021]. However, these methods still have some limitations. On the one hand, existing video deblurring networks directly utilize events as an extra prior without considering the correlation among different events [Wang *et al.*, 2019] [Zhou and Teng, 2021]. These networks fulfill an independent feature map by compressing the intensity changes into one time step, so the temporal information will be lost and the high temporal resolution of events cannot be fully utilized. On the other hand, event-driven video recovery networks rely heavily on the deployment of events [Zhu *et al.*, 2019]. However, these networks simply concatenate features maps of both the blurring frames and events as the input of CNNs, ignoring the rich brightness change information as well as spatial consistency between events and frames [Zou, 2020]. These problems limit the further development of principled work on event-based video deblurring.

In this paper, we develop a unique framework for the event-driven video deblurring, where the spatial consistency between the event and frame is fully utilized to recover spatial textures, while the temporal correlation in event sequences is represented to record long-range dependencies of them continuously. To achieve these two goals, we design a frame-

*Corresponding author

event Spatial Fusion Block (SFB) to combine the two types of features from frames and events. With the help of the brightness prior, this block calculates the non-local features to capture spatial consistency between each frame and event. Then, to utilize high temporal information provided by events, we introduce a Temporal Memory Block (TMB) to restore temporal information between different sequences of events continuously. This block computes the long-range dependency of different events to restore temporal event correlation. The final deblurring network is constructed based on the two blocks and trained in an end-to-end fashion. Furthermore, our proposed spatial fusion block and temporal memory block can cope with synthetic and real-world video deblurring datasets to achieve favorable performance, substantiating the effectiveness of our spatio-temporal relation-aware structure.

The main contributions of our work are three-fold:

- We propose a spatio-temporal relation-aware network for accurate event-driven video deblurring. Our network achieves better performance through fusing features of frames and events properly.
- A novel spatial fusion block is proposed for modeling the spatial relationship between frames and events, which greatly makes use of the spatial consistency.
- A temporal memory block is developed to record long-range dependencies of event sequences in each time step to utilize temporal information from events effectively.
- Extensive experiments show that our method can yield high-quality deblurred results, e.g., our network achieves superior performance by 1.3 dB compared with [Xu *et al.*, 2021] on the HQF dataset.

2 Related Work

2.1 Video Deblurring

Due to the ill-posed nature of video deblurring, traditional methods always constrain this problem by using some assumptions or priors, which cannot adapt to complex dynamic scenes including moving things. The blurring effects come from several different scenes, such as depth variation, moving objects, camera shaking, etc [Li *et al.*, 2020]. Recently, learning-based methods have been introduced to solve this problem and made considerable achievements. In particular, [Deng *et al.*, 2021] proposes a separable-patch structure with channel spatial attention blocks to utilize multi-scale integration to obtain larger receptive field. To make better use of the consecutive sharp frames, [Li and Xu, 2021] designs a correlative module relying on spatial relations among blurring frames to enhance the correlations. [Durand, 2018] regards all frames as equal and construct them order-independently in the burst by putting them in arbitrary size. [Zhang and Luo, 2019] applies 3D convolution blocks in temporal domains to restore sharp frame details and use a GAN-based generator for effective adversarial training. Although the above-mentioned methods achieve considerable performance, they always attach great importance to network structures but ignore other helpful information to improve the deblurring performance.

2.2 Event-based Video Deblurring

The event camera measures the pixel-level brightness change and outputs events when the change exceeds a pre-defined threshold. By introducing events into the deblurring issue, it is easier to handle the blurry texture erasure. Most event-based video deblurring methods utilize neural networks and directly learn the relation from a blurry image to a sequence of sharp images with the aid of events. So the most important question is, how to make full use of its high temporal resolution. [Pan *et al.*, 2019] finds that blur frames are always around in sharp frames. Based on this observation, the authors proposed a flexible fusion module to detect nearest sharp frames and obtain similarity in event and blur videos. To solve the degradation problem caused by the inconsistent data, [Xu *et al.*, 2021] utilizes the self-supervised learning strategy to exploit latent information in events. D2Net [Shang *et al.*, 2021] detects blurry frames as a binary classification task, adopt bidirectional LSTM (BiLSTM) to classify sharp frames and blurry frames, by which temporal correlations of adjacent frames in both forward and backward directions are leveraged. [Zou, 2020] finds that the integrals of events among sharp and blurry frames, and adopted events to predict the residual for video deblurring and interpolation. [Wang *et al.*, 2020] applies sparse learning on the sparsity data of events, and perform super-resolution, deblurring and denoising simultaneously in an general network.

The existing methods, however have not considered the use of rich brightness change information by mining the spatial consistency between image features and event features. In addition, the long-time dependency in different event sequences is also not fully exploited. Therefore, we propose a novel framework to solve the event-based video deblurring by constantly being aware of the spatial consistency and temporal correlation between frames and events.

3 Methodology

3.1 Network Structure

Figure 2 presents the workflow of our proposed network. Given a sequence of video frames B and their corresponding events E, the network processes frames in temporal order with brightness prior of events to generate the final output O. To complete a target sharp frame, our network firstly takes three consecutive frames into an U-Net structure to extract features. Then, the spatial fusion block is adopted to fuse two types (frames and events) of features properly by calculating non-local operation with the help of brightness prior. To take advantages of high temporal information provided by events, the temporal memory block will restore temporal information between different sequences of events continuously. Since the core components of our network are the spatial fusion block and temporal memory block, below we detail these two parts.

3.2 Spatial Fusion Block

Inspired by the asymmetric fusion between intra-scale features from different scales [Cho *et al.*, 2021], we take three different feature maps as input and combine multi-scale features by resizing three different maps to the same scale and

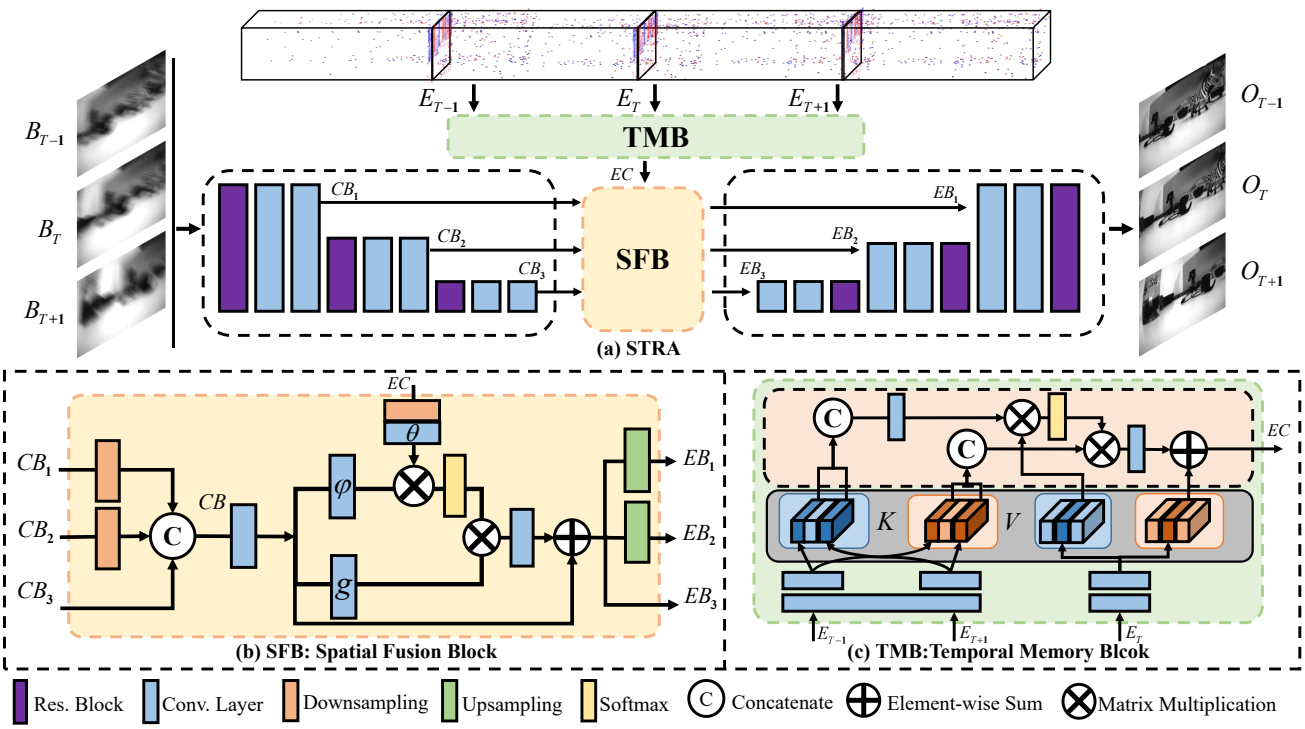


Figure 2: Our proposed network architecture, which is based on the U-Net structure, contains two specifically designed blocks: spatial fusion block (SFB) and temporal memory block (TMB). Given the triplet blurring consecutive frames and their corresponding event sequences, SFB is utilized to calculate the spatial consistency between frames and events, and then fuse them properly; TMB restores the long-range dependencies of different event sequences continuously and records them in temporal order, providing event features for SFB.

concatenating them for fusion with event features. Specifically, this process is formulated as:

$$CB = Conv(Concat(CB_1^{\downarrow}, CB_2^{\downarrow}, CB_3)), \tag{1}$$

where CB_1, CB_2 and CB_3 represent the three different maps from U-Net encoder, CB represents the output of feature fusion. Downsampling ($\downarrow$) is applied in order to concatenate features from different scale. For events, we use a downsampling layer to keep the same scale with the frame feature. Then, with consideration of spatial correlation between all positions in each frame and the corresponding brightness changes in events, we apply the non-local attention operation to fuse spatial feature [Wang *et al.*, 2018; Li *et al.*, 2019] from frames and events as:

$$Feat_i = \frac{1}{C(E)} \sum_{\forall j} f(E_i, F_j) g(F_j), \tag{2}$$

where i is the index of the output position and j is the index that enumerates all possible positions. F and E represent the input frame feature maps and event feature maps, respectively. $Feat$ is the output of the same size as F. The function $g(\cdot)$ calculates the representation of the input frame feature maps at the position j. The result will be normalized by factor $C(x) = N$, where N is the number of positions in E. The function $f(\cdot, \cdot)$ is defined as follows:

$$f(E_i, F_j) = \theta(E_i)^T \varphi(F_j). \tag{3}$$

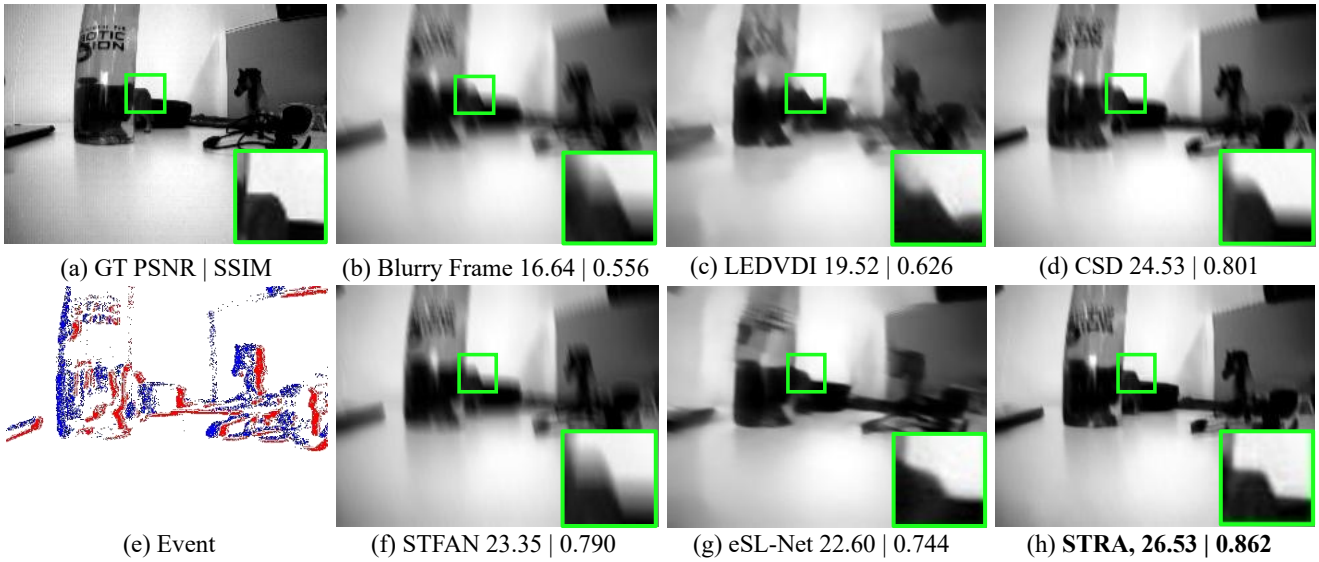
We use softmax operation as activation function. Finally, to make the connection with U-net decoder feature maps, we also upsample the output to three maps (EB_1,EB_2, EB_3) with different scales. The whole structure of SFB is shown in Figure 2(b). Different from the traditional non-local network that only calculates the weighted sum of features at all positions in a single feature map, our SFB fuses three frame feature maps with different scales and captures correlation with corresponding event sequences.

3.3 Temporal Memory Block

Motivated by the spatial and temporal non-local correlation to generate segmentation to current frame [Xie *et al.*, 2021], our temporal memory block consists of a memory encoder and a reader. For $T-1$ and $T+1$ event sequence, we use one common-used and two special-used convolutional layers to obtain the key and the value. Each key and value will be obtained by deploying their own convolutional layer on the common feature map F_m so that all keys and values of each frame will be stored by temporal order [Yao *et al.*, 2021]. The equations are shown as follows:

$$K_{T-1}, V_{T-1} = Conv_{T-1}(F_m), \tag{4}$$
$$K_{T+1}, V_{T+1} = Conv_{T+1}(F_m). \tag{5}$$

In the memory reader, the keys and values of previous events and next events are concatenated, and the similarities between query and keys are used to measure temporal non-local correspondence with current events, which will generate



(a) GT PSNR | SSIM (b) Blurry Frame 16.64 | 0.556 (c) LEDVDI 19.52 | 0.626 (d) CSD 24.53 | 0.801

(e) Event (f) STFAN 23.35 | 0.790 (g) eSL-Net 22.60 | 0.744 (h) **STRA, 26.53 | 0.862**

Figure 3: Visual results of event-driven video deblurring in HQF dataset. Quantitative results are presented by PSNR | SSIM values.

Metrics \ Methods	LEDVI	D2Net	eSL-Net	CSD	STFAN	RED-Net	**STRA**	Ideal value
PSNR	22.22	26.16	25.42	24.71	24.17	25.72	**27.54**	$+\infty$
SSIM	0.687	0.864	0.754	0.724	0.711	0.763	**0.834**	1.00

Table 1: Average PSNR and SSIM results on the HQF dataset.

the corresponding value to aware temporal changes [Zhang and Zhu, 2021]. Then, the value is processed based on the non-local attention operation [Hu *et al.*, 2018] as shown in Figure 2(c). In this way, the temporal correlation between the current event sequence and its neighbor sequences will be fully utilized to record long-range dependencies.

3.4 Loss Function

In this paper, we use Mean squared error (MSE) to train our network in an end-to-end fashion:

$$L_{MSE} = ||f(B) - GT||_2, \tag{6}$$

where B and GT is the blurry frame and its ground truth counterpart, respectively. $f(\cdot)$ represents the network.

4 Experiments

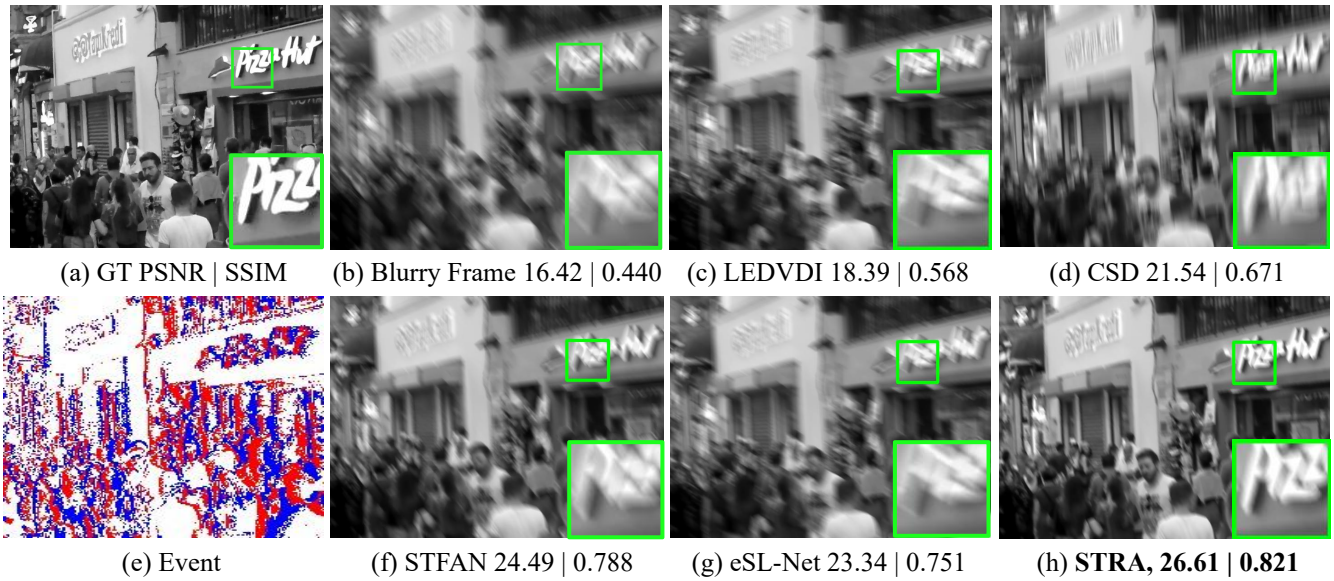
4.1 Experimental Settings

Dataset. Our Spatio-Temporal Relation-Aware network (STRA) is trained based on the benchmark GoPro dataset [Nah *et al.*, 2017], composed of synthetic events, 2,103 pairs of blurring frames and sharp clear ground-truth frames. The blurry image is offered by averaging nearby (the number varies from 7 to 13) frames. To increase the noise diversity, V2E [Hu *et al.*, 2021] is utilized to generate the corresponding event sequences with consideration of different contract thresholds for pixel-level from Gaussian distribution $N(0.18, 0.03)$ [Zou, 2020]. For evaluation in real-world events, we utilize HQF dataset [Stoffregen *et al.*, 2020], including both real-world events and ground-truth frames captured from a DAVIS240C [Brandli *et al.*, 2014], which is a dynamic event-based vision sensor to report brightness changes. The blurring frames are generated by using the same strategy as the GoPro dataset. We also test our network on the GoPro testing datasets, where the number of frame pairs is 1,111.

Implementation Details. Our network is implemented using Pytorch on a single NVIDIA RTX 2080Ti GPU. In the training process, we randomly cropped the sampled frames with the size of 256×256. For data augmentation, each patch was horizontally flipped with the probability of 0.5. We use a batch size of 8 training pairs and ADAM optimizer [Kingma and Ba, 2017] with parameter $\beta_1 = 0.9, \beta_2 = 0.999$. The maximum training epoch is set to 200, with the initial learning rate 10^{-4}, then decays by 25% every 50 epochs.

4.2 Comparison with State-of-the-art Methods

We compare our proposed STRA with several state-of-the-art event-driven video deblurring methods, including LEDVI [Zou, 2020], eSL-Net [Wang *et al.*, 2020], CSD [Wang *et al.*, 2021], STFAN [Zhou *et al.*, 2019], and RED-Net [Xu *et al.*, 2021] D2Net [Shang *et al.*, 2021]. The quantitative results in GoPro testing dataset and HQF dataset are presented in Tables 1 and 2. It is clear that our network achieves outstanding improvements compared with the state-of-the-arts, on average 0.91 dB in GoPro dataset and 1.3 dB in HQF dataset with real-world events. This is because our STRA can benefit from



(a) GT PSNR | SSIM (b) Blurry Frame 16.42 | 0.440 (c) LEDVDI 18.39 | 0.568 (d) CSD 21.54 | 0.671

(e) Event (f) STFAN 24.49 | 0.788 (g) eSL-Net 23.34 | 0.751 (h) **STRA, 26.61 | 0.821**

Figure 4: Visual results of event-driven video deblurring in GoPro dataset. Quantitative results are presented by PSNR | SSIM values.

Metrics \ Methods	LEDVI	D2Net	eSL-Net	CSD	STFAN	RED-Net	**STRA**	Ideal value
PSNR	22.86	28.10	22.59	27.54	28.07	28.98	**29.73**	$+\infty$
SSIM	0.733	0.897	0.750	0.834	0.836	0.849	**0.927**	1.00

Table 2: Average PSNR and SSIM results on the GoPro dataset.

Methods	# Params	FLOPS	Throughput (images/s)
LEDVI	4.73M	1.54G	11.57
eSL-Net	1.32M	4.97G	17.21
CSD	2.07M	3.32G	20.03
STFAN	3.64M	1.18G	14.78
STRA	2.36M	1.10G	22.54

Table 3: Comparison of different backbones on parameter numbers, FLOPS and Thoughput. The Throughput is measured as the number of images processed per second on a RTX 2080Ti GPU.

Methods	frame-based		frame-event-based	
	D2Net	STRA	D2Net	STRA
PSNR	25.93	26.82	28.10	29.73
SSIM	0.770	0.792	0.897	0.927

Table 4: Comparisons of different baselines between D2Net and STRA in GoPro dataset.

the temporal correlation between different event sequences, whose brightness changes can be used to reconstruct sharp frames. Moreover, our model can fully utilize spatial consistency between each frame and the corresponding event sequence.

Figures 3 and 4 present deblurring results of STRA and other comparing methods. Visual quality comparisons demonstrate that the proper fusion of frames and events data in spatial and temporal dimensions can accomplish high-quality video deblurring with more structural details. For example, in Figures 3 (e), the event data can capture much more brightness changes of one scene and contribute to the final deblurring results. In Figure 4, the letter 'P' in the GoPro testing dataset can be restored better by our STRA. Other comparison methods, such as eSL-Net [Wang *et al.*, 2020] and LEDVI [Zou, 2020], however, did not make full use of high temporal information and spatial consistency in events and frames, so the reconstructed frames are unstable on both synthetic and real-world event datasets.

4.3 Runtime and Parameter Numbers

We use 200 images with the size of 180×320 for testing on a RTX 2080Ti GPU. Results in average running time and parameters are presented in Table 3. It is obvious that our STRA has comparable FLOPS and running time with consideration of acceptable storage consumption to achieve promising video deblurring performance.

4.4 Ablation Study

To find out what contributes to the superior performance of our approach, we conducted ablation study to demonstrate the effectiveness of each component.

Effect of Events. We test the importance of events by training D2Net [Shang *et al.*, 2021] and our STRA without events over the synthetic GoPro dataset, and the quantitative results are shown in Table 4, which shows that the frame-event-based

Temporal Memory Block	Spatial Fusion Block	GoPro	HQF
✗	✗	27.81	25.03
✓	✗	29.04	25.94
✗	✓	28.50	25.38
✓	✓	**29.73**	**27.54**

Table 5: Quantitative ablation study on the two relation-aware blocks. We evaluate the effect of two blocks on PSNR values.

Channel # \ Res-unit #	$R = 4$	$R = 8$	$R = 12$
$C = 16$	28.02	28.51	28.73
$C = 32$	29.29	29.73	29.11
$C = 48$	28.71	29.08	29.45

Table 6: PSNR values on different parameter settings.

method has better performance than other baselines. Moreover, our STRA also achieves superior performance than STFAN in all baselines.

Effect of TMB and SFB. We validate the importance of temporal memory block by training STRA without neighboring correlation of events over the synthetic GoPro dataset, and there is a great performance gap of quantitative results over two datasets in the first two rows of Table 5. It shows that temporal dependency between different events can efficiently improve the deblurring performance by 1.23 dB in GoPro dataset and 0.91 dB in HQF dataset.

We also test the effect of spatial fusion block in the same way, and results are shown in the first and third row in Table 5. Apparently, spatial fusion with events and frames achieves PSNR gains up to 0.69 dB and 0.35 dB on GoPro and HQF dataset, respectively. Most importantly, when both spatial fusion block and temporal memory block are embedded in STRA, we achieve even higher deblurring performance than inserting only one block.

Effect of the ResNet Depth. Note that in Figure 2, each res-block has several basic res-units, and the number of channels attaches importance to the performance of STRA. So we also evaluate the influence of the number of channels and res-units in STRA. Specifically, we set channel number $C \in (16, 32, 48)$ and res-unit number $R \in (4, 8, 12)$. Table 6 shows the PSNR and SSIM performance in different settings. To achieve a good trade-off between efficiency and performance, we set $C = 32$ and $R = 8$ as the default setting.

4.5 Visualization

We first visualize some attention maps to see the representation of the spatial non-local operation between frames and events [Mou *et al.*, 2021]. In Figures 5 (b), it is clear that similar brightness can contribute more to query patches, which demonstrates that our spatial fusion block can also present long-range spatial correlations to particular patches.[1]

[1]More results and visualizations can be found in the supplement.

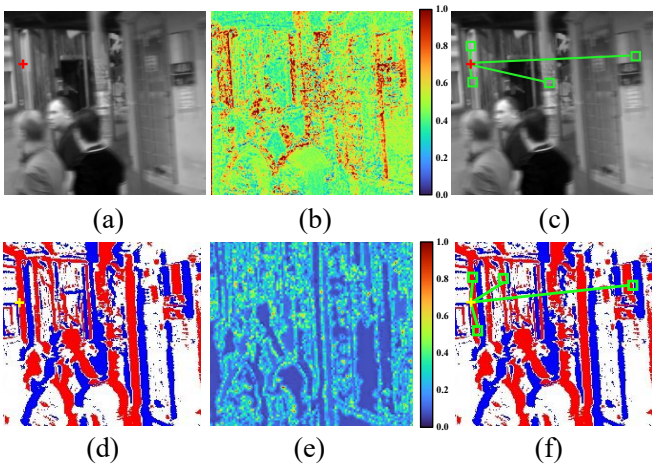


Figure 5: Visualizations of similarity matrixes in the spatial and temporal blocks. Regions of query points are labeled with red and yellow cross. The similarity matrixes of query points are presented in the form of heat maps in (b) and (e). For illustration purpose, we present some greatly correlated neighbors (labeled with green boxes) in (c) and (f).

For temporal dependency between different events, we show some similarity matrixes with neighbor events in Figure 5 (e). One can see that regions of frequent brightness changes are more important for event patches, which illustrates that temporal connection between neighbor event sequences has been built in this way.

5 Conclusion

We proposed a new relation-aware network for event-driven video deblurring by rethinking the problem at both the spatial and temporal level. One advantage of this method over most deep-learning-based deblurring models is that we attach importance to the temporal correlation between different event sequences, and restore it continuously to achieve better deblurring performance in consecutive frames. While other event-driven deblurring models take a single frame and corresponding events as input with no consideration of high temporal information in events. On the other hand, our network calculates spatial consistency between events and frames by improving non-local operations to capture blurring contexts. Our network learns to reconstruct sharp edges by relying on sufficient temporal and spatial information in events to create high-quality frames. Both subjective and objective results on synthetic and real-world datasets have demonstrated the effectiveness of our proposed network.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, the National Natural Science Foundation of China (NSFC) under Grants U19B2038 and 61901433, the University Synergy Innovation Program of Anhui Province under Grants GXXT-2019-025, the Fundamental Research Funds for the Central Universities under Grant WK2100000024.

References

[Brandli *et al.*, 2014] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 dB 3 μs Latency Global Shutter Spatiotemporal Vision Sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[Cho *et al.*, 2021] Sung-Jin Cho, Seo-Won Ji, and Jun-Pyo Hong. Rethinking Coarse-to-Fine Approach in Single Image Deblurring. *CVPR*, 2021.

[Deng *et al.*, 2021] Senyou Deng, Wenqi Ren, and Yanyang Yan. Multi-Scale Separable Network for Ultra-High-Definition Video Deblurring. *ICCV*, 2021.

[Durand, 2018] Frédo Durand. Burst Image Deblurring Using Permutation Invariant Convolutional Neural Networks. In Vittorio Ferrari and Martial Hebert, editors, *ECCV*. 2018.

[Gallego *et al.*, 2021] Guillermo Gallego, Tobi Delbrück, and Garrick Orchard. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2018.

[Hu *et al.*, 2021] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2e: From Video Frames to Realistic DVS Events. *CVPR*, 2021.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *NeurIPS*, 2017.

[Li and Xu, 2021] Dongxu Li and Chenchen Xu. ARVo: Learning All-Range Volumetric Correspondence for Video Deblurring. *CVPR*, 2021.

[Li *et al.*, 2019] Xiang Li, Wenhai Wang, and Xiaolin Hu. Selective Kernel Networks. In *CVPR*, 2019.

[Li *et al.*, 2020] Ang Li, Jianzhong Qi, and Rui Zhang. Generative Image Inpainting with Submanifold Alignment. *IJCAI*, 2020.

[Mou *et al.*, 2021] Chong Mou, Jian Zhang, and Zhuoyuan Wu. Dynamic Attentive Graph Learning for Image Restoration. *ICCV*, 2021.

[Nah *et al.*, 2017] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. In *CVPR*, 2017.

[Nikzad *et al.*, 2021] Mohammad Nikzad, Yongsheng Gao, and Jun Zhou. Attention-based Pyramid Dilated Lattice Network for Blind Image Denoising. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[Pan *et al.*, 2019] Liyuan Pan, Cedric Scheerlinck, and Yuchao Dai. Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera. In *CVPR*, 2019.

[Shang *et al.*, 2021] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *ICCV*, pages 4531–4540, 2021.

[Stoffregen *et al.*, 2020] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the Sim-to-Real Gap for Event Cameras. *ECCV*, 2020.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, and Abhinav Gupta. Non-local Neural Networks. *CVPR*, 2018.

[Wang *et al.*, 2019] Zihao W. Wang, Weixin Jiang, and Kuan He. Event-Driven Video Frame Synthesis. In *ICCVW*, 2019.

[Wang *et al.*, 2020] Bishan Wang, Jingwei He, and Wen Yang. Event Enhanced High-Quality Image Recovery. *ECCV*, 2020.

[Wang *et al.*, 2021] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards Compact Single Image Super-Resolution via Contrastive Self-distillation. *IJCAI*, 2021.

[Xie *et al.*, 2021] Fei Xie, Wankou Yang, and Kaihua Zhang. Learning Spatio-Appearance Memory Network for High-Performance Visual Tracking. In *ICCVW*, 2021.

[Xu *et al.*, 2021] Fang Xu, Lei Yu, and Bishan Wang. Motion Deblurring With Real Events. *ICCV*, 2021.

[Yao *et al.*, 2021] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-Wise Attention Spiking Neural Networks for Event Streams Classification. In *ICCV*, 2021.

[Zhang and Luo, 2019] Kaihao Zhang and Wenhan Luo. Adversarial Spatio-Temporal Learning for Video Deblurring. *IEEE Trans. on Image Process.*, 2019.

[Zhang and Zhu, 2021] Wendong Zhang and Junwei Zhu. Context-Aware Image Inpainting with Learned Semantic Priors. *IJCAI*, 2021.

[Zhou and Teng, 2021] Chu Zhou and Minggui Teng. DeLiEve-Net: Deblurring Low-light Images with Light Streaks and Local Events. In *ICCVW*, 2021.

[Zhou *et al.*, 2019] Shangchen Zhou, Jiawei Zhang, and Jinshan Pan. Spatio-Temporal Filter Adaptive Network for Video Deblurring. In *ICCV*, 2019.

[Zhu *et al.*, 2019] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. In *CVPR*, 2019.

[Zou *et al.*, 2021] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. EventHPE: Event-Based 3D Human Pose and Shape Estimation. *ICCV*, 2021.

[Zou, 2020] Dongqing Zou. Learning Event-Driven Video Deblurring and Interpolation. In Andrea Vedaldi and Horst Bischof, editors, *ECCV*. 2020.