Motion Deblurring via Spatial-Temporal Collaboration of Frames and Events

Wen Yang[1,2], Jinjian Wu[1,2*], Jupo Ma[1,2*], Leida Li[1], Guangming Shi[1,2]

[1]School of Artificial Intelligence, Xidian University, Xi'an 710071, China
[2]Pazhou Lab, Huangpu, 510555, China
wen.yang@stu.xidian.edu.cn, {jinjian.wu, majupo, ldli, gmshi}@xidian.edu.cn

Abstract

Motion deblurring can be advanced by exploiting informative features from supplementary sensors such as event cameras, which can capture rich motion information asynchronously with high temporal resolution. Existing event-based motion deblurring methods neither consider the modality redundancy in spatial fusion nor temporal cooperation between events and frames. To tackle these limitations, a novel spatial-temporal collaboration network (STCNet) is proposed for event-based motion deblurring. Firstly, we propose a differential-modality based cross-modal calibration strategy to suppress redundancy for complementarity enhancement, and then bimodal spatial fusion is achieved with an elaborate cross-modal co-attention mechanism to weight the contributions of them for importance balance. Besides, we present a frame-event mutual spatio-temporal attention scheme to alleviate the errors of relying only on frames to compute cross-temporal similarities when the motion blur is significant, and then the spatio-temporal features from both frames and events are aggregated with the custom cross-temporal coordinate attention. Extensive experiments on both synthetic and real-world datasets demonstrate that our method achieves state-of-the-art performance. Project website: https://github.com/wyang-vis/STCNet.

Introduction

Motion blur is commonly inevitable due to camera shake or object motion over the period of exposure time, which not only deteriorates the visual experience for humans but hinders other computer vision tasks such as tracking (Jin, Favaro, and Cipolla 2005; Mei and Reid 2008), video stabilization (Matsushita et al. 2006), etc. To eliminate the adverse effects, the task of motion deblurring has received much research attention recently.

Traditional motion deblurring techniques explicitly utilize image priors and various constraints (Bar et al. 2007; Cho, Wang, and Lee 2012; Wulff and Black 2014; Bahat, Efrat, and Irani 2017; Kotera, Šroubek, and Milanfar 2013; Levin et al. 2009) that are handcrafted with empirical observations. However, it is challenging to design such priors and constraints to model the inherent properties of latent frames and motion blur. Due to the success of deep neural networks (DNNs), some deep convolutional neural network (CNN)-based methods (Zhang et al. 2019; Zamir et al. 2021; Chen et al. 2021; Cho et al. 2021), recurrent neural network (RNN)-based methods (Nah, Son, and Lee 2019; Zhong et al. 2020; Zhou et al. 2019; Zhu et al. 2022) and Transformer-based methods (Liang et al. 2021; Wang et al. 2022b; Liang et al. 2022b,a) have been proposed for motion deblurring, which implicitly learn more general prior information from large-scale training data. Despite their good performance, these learning-based deblurring methods may fail to deal with severe blur. Motion deblurring cannot be solved trivially from the input blur set alone, as it is a highly ill-posed problem with infinite feasible solutions.

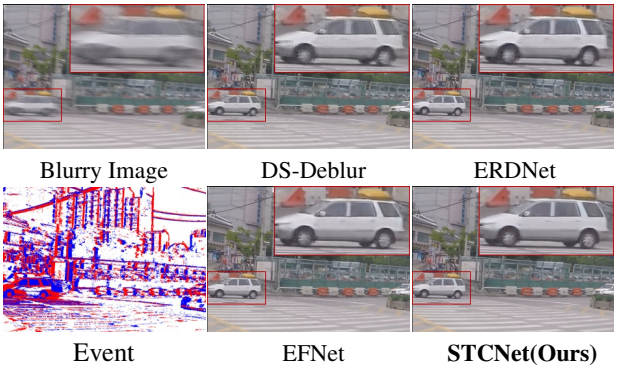

Figure 1: Comparison of visualized deblurring results with state-of-the-art event-based motion deblurring methods DS-Deblur (Yang et al. 2022), ERDNet (Chen et al. 2022a), EFNet (Sun et al. 2022), and our STCNet.

Event cameras are bio-inspired sensors that can record per-pixel intensity changes asynchronously with high temporal resolution and output a stream of *events* encoding time, location and polarity of intensity changes (Vitoria et al. 2023) if the intensity changes surpass a threshold. Understandably, with the attractive properties that offer motion information with microsecond accuracy, event cameras have been attempted to address motion deblurring. Recently, some event-based motion deblurring methods are proposed (Pan et al. 2019; Jiang et al. 2020; Lin et al. 2020;

*Corresponding authors.

Sun et al. 2022; Chen et al. 2022a; Sun et al. 2023), and have achieved promising performance of deblurring. Crucially, these methods, on the one hand, adopt only simple fusion strategy for spatial complementary fusion and also do not consider the modality redundancy; on the other, neglect the role of the event itself and the event-frame interaction in the temporal domain. These insufficient collaboration of events and frames limits the overall performance.

In this paper, we develop a novel spatial-temporal collaboration network (STCNet) to learn the collaborative fusion of frames and events both in spatial and temporal aspects for motion deblurring. Generally, different modalities are usually complementary but also redundant. We present a differential-modality guided cross-modal calibration strategy to enhance complementarity, which leverages the global interaction of differential-modality and two modalities. The calibration operation allows to later fuse the features better, and potentially avoids the modality redundancy. Then considering the disparity of the contributions of different modality features, we elaborate a cross-modal co-attention scheme to balance the contributions of multi-modality features for spatial complementary fusion. Besides, exploiting spatio-temporal dependencies is useful for motion deblurring. There may be errors in estimating the cross-temporal similarities relying only on frames when fast motions are present, i.e., spatio-temporal modeling lacks the guidance of motion information. Fortunately, benefiting from the rich motion information in the event, we propose a frame-event mutual spatio-temporal attention scheme to model the cross-temporal dependencies by conducting the communication of cross-frames and cross-events, alleviating that issue of cross-temporal similarities computation. Based on the mutual spatio-temporal attention, not only spatio-temporal features from frames, but also additional from events, are aggregated to the current feature with a custom cross-temporal coordinate attention. Coupled with the above spatial and temporal collaboration strategy, our framework achieves state-of-the-art performance of event-based motion deblurring (some visual comparisons are shown in Figure 1). The main contributions of our work are as follows.

- We propose a novel spatial-temporal collaboration network (STCNet) for event-based motion deblurring, which facilitates the collaborative fusion of frames and events in both spatial and temporal domains. Extensive experiments show that our model outperforms state-of-the-art event-based and image/video-based methods.
- We present a differential-modality guided cross-modal calibration strategy to enhance complementarity and suppress redundancy of multi-modality features. Then the calibrated multi-modality features are fused by a cross-modal co-attention scheme to adaptively balance the modality contributions.
- We propose a mutual spatio-temporal attention to model the cross-temporal dependencies by enjoying the extra assistance of motion information in events. Based on this, informative features from temporal neighbors of both frames and events are fused with current features via a custom cross-temporal coordinate attention.

Related Work

Motion Deblurring

Early methods focus on explicitly using image priors and constraints (Cho, Wang, and Lee 2012; Hyun Kim and Mu Lee 2015; Bahat, Efrat, and Irani 2017; Kotera, Šroubek, and Milanfar 2013; Levin et al. 2009) that are handcrafted with empirical observations. With the development of deep learning, researchers have made significant progress on motion deblurring. State-of-the-art learning-based deblurring methods use a single image or multiple frames.

Image Deblurring. Contemporary successful deep learning-based image deblurring methods can be roughly categorized as follows. 1) Single-Stage Approaches. These methods are based on a single-stage design, using the convolutional neural network (CNN) (Zhang et al. 2020) or Generative Adversarial Network (GAN) (Kupyn et al. 2018, 2019). 2) Multi-Stage Approaches. These methods aim to recover clean images in a progressive manner with multi-stage (Nah, Hyun Kim, and Mu Lee 2017; Tao et al. 2018; Zamir et al. 2021; Chen et al. 2021), which decompose the image deblurring task into smaller easier subtasks. 3) Coarse-to-Fine Strategies. These methods typically stack sub-networks with multi-scale inputs and gradually improve sharpness of images (Park et al. 2020; Cho et al. 2021). 4) Attention Modules. Attention mechanisms can help learn cross-spatial/channel correlations to better address deblurring (Suin, Purohit, and Rajagopalan 2020; Tsai et al. 2022; Purohit and Rajagopalan 2020; Liang et al. 2021).

Video Deblurring. The spatio-temporal correlation between adjacent inputs is critical for video deblurring. Recurrent neural network (RNN) or convolutional neural network (CNN) are adopted to exploit temporal information (Nah, Son, and Lee 2019; Zhong et al. 2020; Zhou et al. 2019; Su et al. 2017; Zhu et al. 2022). To improve the deblurring performance further, some extra multiple frames aligning methods were proposed to model spatio-temporal correlation, such as optical flow based methods (Pan, Bai, and Tang 2020; Xiang, Wei, and Pan 2020), deformable and dynamic convolutions based methods (Wang et al. 2019; Zhou et al. 2019). Recently, the emergence of Transformer provides an alternative for effective temporal modeling for video deblurring (Liang et al. 2022b,a; Lin et al. 2022), due to its advantages of modeling long-range spatial dependencies.

Event-Based Motion Deblurring

Event cameras provide visual information with low latency and with strong robustness against motion blur, which offers great potential for motion deblurring. Event-based motion deblurring methods can be divided into two categories (Xu et al. 2021), i.e., model driven and data driven algorithms.

Model driven methods formulate the relation from blurry images to sharp images with the physical event generation principle (Scheerlinck, Barnes, and Mahony 2018). Specifically, Pan et al. (Pan et al. 2019) modeled the blur-generation process by associating event to a latent frame with an Event-based Double Integral (EDI) algorithm for deblurring. Scheerlinck et al. (Scheerlinck, Barnes, and Mahony 2018)

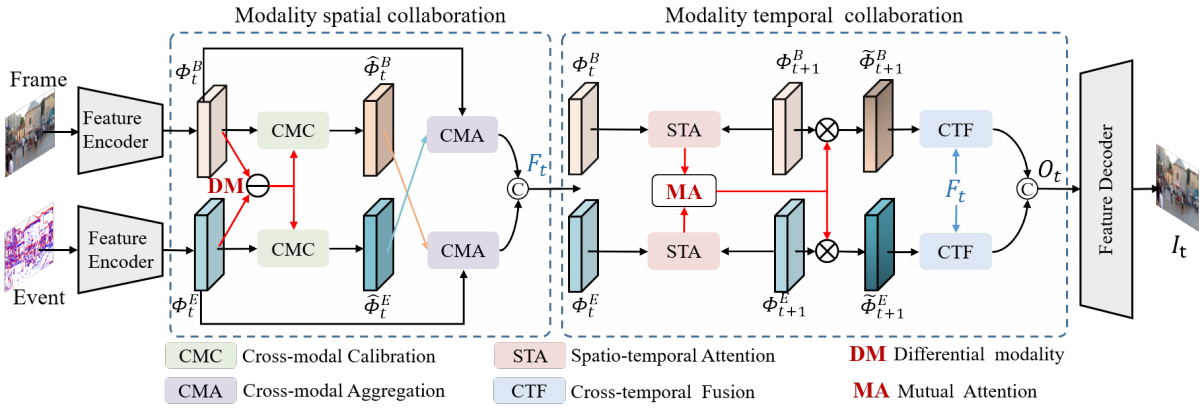

Figure 2: Framework of our STCNet, containing two parts: modality spatial collaboration and modality temporal collaboration.

presented a continuous-time formulation of event-based intensity estimation using complementary filtering to combine frames with events. Regrettably, there is inevitable noise in events due to the non-ideality of physical sensors (Zhang and Yu 2022), resulting in degraded performance.

Data driven methods tackle above limitations by learning-based approaches (Lin et al. 2020). LEMD (Jiang et al. 2020) presented a sequential formulation of event-based motion deblurring, and unfolded its optimization with deep architecture. eSL-Net (Wang et al. 2020) proposed an event enhanced degeneration model for the high-quality image recovery. Shang et al. (Shang et al. 2021) proposed an event fusion module to utilize beneficial information from events, which can be incorporated into existing motion deblurring methods. EFNet (Sun et al. 2022) first introduced a symmetric cumulative event representation and then proposed a cross-modal attention module to fuse image and event. ERDNet (Chen et al. 2022a) proposed a residual learning approach to learn event-based motion deblurring.

Method

Problem Statement

Given a blurry frame B and the corresponding event stream $E_T \triangleq \{(x_i, y_i, p_i, t_i)\}_{t_i \in T}$ containing all events triggered during exposure time T, where $p = \pm 1$ is polarity, which denotes the direction (increase or decrease) of the intensity changes at that pixel (x, y) and time t, the proposed method is to recover a sharp frame I by exploiting both blurry frame B and event stream E_T, which can be modeled as $I = G(B, E_T)$, where G is deep learning model.

Principled Framework of STCNet

In our work, a novel spatial-temporal collaboration network (STCNet) is proposed for event-based motion deblurring, which can facilitate the collaborative fusion of frames and events in both spatial and temporal domains. Figure 2 shows the overview of STCNet. We first use symmetric feature encoder to extract target features Φ_t^B and Φ_t^E from blurry frame and its corresponding events, separately. Next, we conduct modality spatial collaboration with first differential-modality guided cross-modal calibration (CMC) for complementary enhancement and then cross-modal aggregation (CMA) for contribution balance, obtaining F_t. Besides, we conduct modality temporal collaboration with first frame-event mutual spatio-temporal attention (STA) for cross-temporal dependencies modeling and then cross-temporal fusion (CTF) for spatio-temporal features fusion, obtaining O_t. Finally, feature decoder reconstructs the deblurred result I_t. Below we detail the main parts: modality spatial collaboration and modality temporal collaboration.

Modality Spatial Collaboration (MSC)

Generally, different modalities usually have complementary features (discrepancy) for each other and also have their shared features (commonalities). Differential features are what cross-modal fusion focuses on, while common features are redundant information. We advocate first calibrating the modality features to enhance complementarity and suppress redundancy, and then considering the disparity of the contributions of different modality features for multi-modality fusion, shown in Figure 3.

Firstly, considering that differential-modality contains complementary cues, a differential-modality guided cross-modal calibration strategy is presented to enhance complementarity. The main idea is leveraging global interaction of differential-modality and two modalities to infer attention maps, then the attention maps are multiplied to the input features respectively for feature enhancement.

The calibration process is realized by the CMC in Figure 3. Given the frame features Φ_t^B and event features Φ_t^E, we first obtain the differential-modality features F_{dm} by direct subtraction of two modalities:

$$F_{dm} = \Phi_t^B - \Phi_t^E, \tag{1}$$

then, based on the traditional self-attention (Vaswani et al. 2017), we further put forward an efficient cross-attention mechanism applied to Φ_t^B, Φ_t^E and F_{dm} to infer the attention maps. Φ_t^E and Φ_t^B are transformed into Key K_e, Value V_e and Key K_b, Value V_b, respectively. F_{dm} is transformed into Query Q_d. Then attention maps can be calculated as:

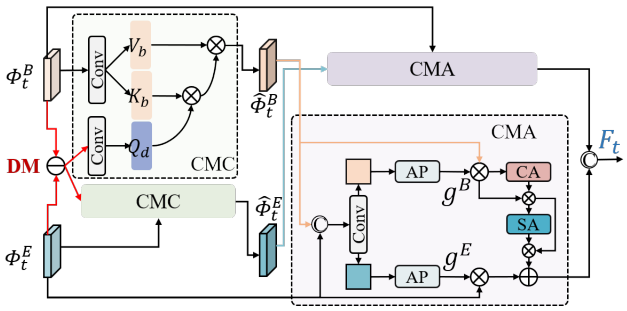

Figure 3: Details of modality spatial collaboration.

$$A^B = Softmax\left(Q_d K_b^T\right), \\ A^E = Softmax\left(Q_d K_e^T\right), \tag{2}$$

where attention maps A^B and A^E contain complementarity clues of frame and event respectively. Thus, the calibrated features can be represented as:

$$\hat{\Phi}_t^B = A^B V_b, \quad \hat{\Phi}_t^E = A^E V_e. \tag{3}$$

Next, considering the disparity of the contributions of different modality features, we elaborate a cross-modal co-attention scheme to balance the contributions of multi-modality features for bimodal spatial complementary fusion. Meanwhile, we use a bi-directional fusion strategy, i.e., from frames to events as well as in the opposite direction.

The aggregation process is realized by the CMA model in Figure 3. Specifically, taking the image-to-event example, given the calibrated feature $\hat{\Phi}_t^B$ and original Φ_t^E, they are first combined using concatenation and convolution operation. Then the combined features are split evenly along the channel dimension into two sub-branches. The sigmoid function and global average pooling are performed on each sub-branch to obtain co-attention scores g^B and g^E, which model the importance of different modal features for the further fusion, which can be formulated as:

$$g = Avg\left(Sig\left(Conv\left(Cat\left(\hat{\Phi}_t^B, \Phi_t^E\right)\right)\right)\right), \tag{4}$$

where $Avg(\cdot)$ is the global average pooling, $Sig(\cdot)$ denotes sigmoid function, $Conv(\cdot)$ refers to the convolution layer and $Cat(\cdot)$ is the concatenation operation. And we apply the co-attention scores to the corresponding features to generate gated features $\overline{\hat{\Phi}_t^B}$ and $\overline{\Phi_t^E}$:

$$\overline{\hat{\Phi}_t^B} = \hat{\Phi}_t^B * g^B, \quad \overline{\Phi_t^E} = \Phi_t^E * g^E. \tag{5}$$

Further, we use channel-wise and spatial-wise attentions to emphasize the supplementary features:

$$\overline{\hat{\Phi}_t^B}' = CA(\overline{\hat{\Phi}_t^B}) * \overline{\hat{\Phi}_t^B}, \quad \overline{\hat{\Phi}_t^B}'' = SA(\overline{\hat{\Phi}_t^B}') * \overline{\hat{\Phi}_t^B}', \tag{6}$$

then we devise the aggregation operation as an element-wise addition of the two modalities:

$$\overline{E} = \overline{\Phi_t^E} + \overline{\hat{\Phi}_t^B}'' \tag{7}$$

Similarly, we can obtain $\overline{B}$. The final fusion feature can be denoted as $F_t = cat(\overline{E}, \overline{B})$.

Modality Temporal Collaboration (MTC)

Exploring the useful information from neighboring inputs is crucial for motion deblurring. There may be errors in estimating the cross-temporal similarities relying only on frames when fast motions are present, i.e., spatio-temporal modeling lacks the guidance of motion information. Fortunately, on the one hand, events contain rich motion information that can assist frames to better model cross-temporal relevance; on the other hand, spatio-temporal dependencies of event sequences can also be explored. Thus, by enjoying the extra assistance of events, we propose a frame-event mutual spatio-temporal attention to model the cross-temporal dependencies and then aggregate informative features from temporal neighbors of both frames and events via a custom cross-temporal coordinate attention, illustrated in Figure 4.

In our work, the features of the two adjacent moments before and after are fused to the features of the current moment. We take time t and time $t+1$ for example. Given the features of frame Φ_t^B and Φ_{t+1}^B, as well as event Φ_t^E and Φ_{t+1}^E, the key of the proposed cross-temporal dependencies capturing is to conduct the communication of cross-frames and cross-events. As shown in STA of Figure 4, we first transform Φ_t^B into Query Q_b, and Φ_{t+1}^B into Key K_b, Value V_b, as well as Φ_t^E into Query Q_e, and Φ_{t+1}^E into Key K_e, Value V_e. Then intra-modality individual cross-temporal attention is first estimated by multiplying the queries from one moment and the keys from the other moment:

$$S^B = Softmax\left(Q_b K_b^T\right), \\ S^E = Softmax\left(Q_e K_e^T\right). \tag{8}$$

Then we joint inter-modality cross-temporal attention to obtain the mutual spatio-temporal attention S^M:

$$S^M = S^B S^E. \tag{9}$$

Then the informative spatio-temporal features $\widetilde{\Phi}_{t+1}^B$ and $\widetilde{\Phi}_{t+1}^E$ from both frame and event domains are obtained with the guidance of mutual attention S^M:

$$\widetilde{\Phi}_{t+1}^B = S^M V_b, \quad \widetilde{\Phi}_{t+1}^E = S^M V_e. \tag{10}$$

Following, we fuse the current features F_t with $\widetilde{\Phi}_{t+1}^B$ and $\widetilde{\Phi}_{t+1}^E$ separately. Taking the fusion of F_t and $\widetilde{\Phi}_{t+1}^B$ as an example, the CTF is developed based on coordinate attention (Hou, Zhou, and Feng 2021), which can concurrently capture channel and location importance and long-range dependencies. Figure 4 shows the structure of CTF.

Specifically, given the F_t and $\widetilde{\Phi}_{t+1}^B$, we use two spatially scoped pooling kernels $(H, 1)$ or $(1, W)$ encode each channel of the two features along the horizontal and vertical orientations. The aggregated features are represented as:

$$F_t^h = XAP(F_t), \quad F_t^w = YAP(F_t), \\ B_{t+1}^h = XAP(\widetilde{\Phi}_{t+1}^B), \quad B_{t+1}^w = YAP(\widetilde{\Phi}_{t+1}^B), \tag{11}$$

where XAP and YAP denote the average pooling along the vertical and horizontal directions, respectively. Then we

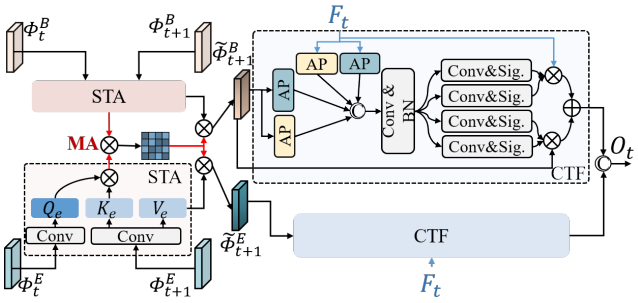

Figure 4: Details of modality temporal collaboration.

join the four aggregated feature by spatial concatenation, followed by convolution and batch normalization (BN):

$$R = BN(Cat(F_t^h, F_t^w, B_{t+1}^h, B_{t+1}^w)). \tag{12}$$

Then, we split R into four separate tensors R_f^h, R_f^w, R_b^h and R_b^w. We use four 1×1 convolution to transform each of the four split tensors into a tensor with the same number of channels as the input features:

$$\begin{aligned} W_f^h &= \sigma(Conv(R_f^h)), \quad W_f^w = \sigma(Conv(R_f^w)), \\ W_b^h &= \sigma(Conv(R_b^h)), \quad W_b^w = \sigma(Conv(R_b^w)), \end{aligned} \tag{13}$$

where $W_f^h, W_f^w, W_b^h, W_b^w$ represent F_t and $\widetilde{\Phi}_{t+1}^B$ coordinate attention weight in the vertical and horizontal directions, respectively. The final weighted features can be defined as:

$$M_f = F_t W_f^h W_f^w, \quad M_b = \widetilde{\Phi}_{t+1}^B W_b^h W_b^w, \tag{14}$$

then we devise the aggregation operation as an element-wise addition of the two features:

$$M_{fb} = M_f + M_b. \tag{15}$$

Similarly, we can obtain M_{fe}. The final fusion feature can be denoted as $O_t = cat(M_{fb}, M_{fe})$.

Loss Function

In this paper, we use the Charbonnier loss (Charbonnier et al. 1994) to train our network in an end-to-end fashion:

$$L_{\text{char}} = \frac{1}{CHW}\sqrt{\|I - G\|^2 + \varepsilon^2}, \tag{16}$$

where I and G is deblurred out and ground truth, respectively, C, H, W are dimensions of frame, and constant ε is empirically set to 10^{-3} as in (Zamir et al. 2021).

Experiments

Experimental Settings

Datasets. Our STCNet is evaluated on 1) Synthetic dataset. *GoPro* (Nah, Hyun Kim, and Mu Lee 2017) and *DVD* (Su et al. 2017) datasets are widely adopted for image-only and event-based deblurring such as (Sun et al. 2022), which contains synthetic blurring images and sharp ground-truth images, as well as synthetic events generated by simulation algorithm ESIM (Rebecq, Gehrig, and Scaramuzza 2018). 2) Real dataset. *REB* dataset is a real event dataset captured by us with the DAVIS346 event camera, including both real events and clear ground-truth images captured under various conditions both indoors and outdoors, that are well-exposed and minimally motion-blurred. The blurring images are generated by using the same strategy as the *GoPro*. There are 60 videos of REB, 40 of which are used for training and 20 for testing. In addition, several sequences are collected under fast camera movement or fast moving scenes for qualitative comparison, without ground truth.

Implementation Details. Our method is implemented using Pytorch on NVIDIA RTX 3090 GPU. The size of training patch is 256×256 with minibatch size of 8. The optimizer is ADAM (Kingma and Ba 2015), and the learning rate is initialized at 2×10^{-4} and decreased by the cosine learning rate strategy with a minimum learning rate of 10^{-6}. For data augmentation, each patch is horizontally flipped with the probability of 0.5. The Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) are adopted as the evaluation metrics.

Comparison With State-of-the-Art Methods

We compare our STCNet to state-of-the-art image/video-only deblurring methods, including MemDeblur (Ji and Yao 2022), MMP-RNN (Wang et al. 2022a), MPRNet (Zamir et al. 2021), MIMO-UNet++ (Cho et al. 2021), Restormer (Zamir et al. 2022), RNN-MBP (Zhu et al. 2022), NAFNet (Chen et al. 2022b), VRT (Liang et al. 2022a), DFFN (Kong et al. 2023), DSTN (Pan et al. 2023), and event-based deblurring methods, including RED* (Xu et al. 2021), eSL-Net* (Wang et al. 2020), D2Nets* (Shang et al. 2021), DS-Deblur* (Yang et al. 2022), ERDNet* (Chen et al. 2022a), EFNet* (Sun et al. 2022), REFID* (Sun et al. 2023).

***GoPro*:** We report the performance of compared motion deblurring approaches on *GoPro* dataset in Table 1. Overall, our method achieves the best performance against other algorithms (1.40dB improvement in terms of PSNR over best image/video-only methods and 0.54dB improvement over best event-based methods). Moreover, we show the qualitative visual quality comparison in Figure 5. Overall, visual quality comparisons demonstrate that our method can recover sharper texture details that are closer to the ground-truth, while the results restored by other methods still suffer from motion blur, losing sharp edge information.

***DVD*:** The STCNet is trained on *GoPro* dataset and tested on *DVD* dataset. Table 2 reports the quantitative results on the *DVD* dataset. Our method significantly outperforms other state-of-the-art competitors (2dB improvement in terms of PSNR over best image/video-only methods and 0.79dB improvement over best event-based methods), demonstrating the superior generalization ability of the proposed framework.

***REB*:** The quantitative performance of real-world dataset *REB* is shown in Table 3. Our method significantly outperforms other competitors (2.69dB improvement in terms of PSNR over best image/video-only methods and 0.53dB improvement over best event-based methods). We show the deblurring visual comparison on real blurs in Figure 6. Our

Method	RED*	eSL-Net*	D2Nets*	MemDeblur	MMP-RNN	MPRNet	MIMO-UNet++	Restormer	DS-Deblur*
PSNR	28.98	30.23	31.76	31.76	32.64	32.66	32.68	32.92	33.13
SSIM	0.8499	0.8703	0.9430	0.9230	0.9359	0.9590	0.9590	0.9610	0.9465
Method	RNN-MBP	NAFNet	DFFN	ERDNet*	VRT	DSTN	EFNet*	REFID*	**STCNet(Ours)***
PSNR	33.32	33.69	34.21	34.25	34.81	35.05	35.46	35.91	**36.45**
SSIM	0.9627	0.9670	0.9692	0.9534	0.9724	0.9733	0.9720	0.9730	**0.9809**

Table 1: Comparison of motion deblurring methods on *GoPro* dataset. * denotes event-based methods.

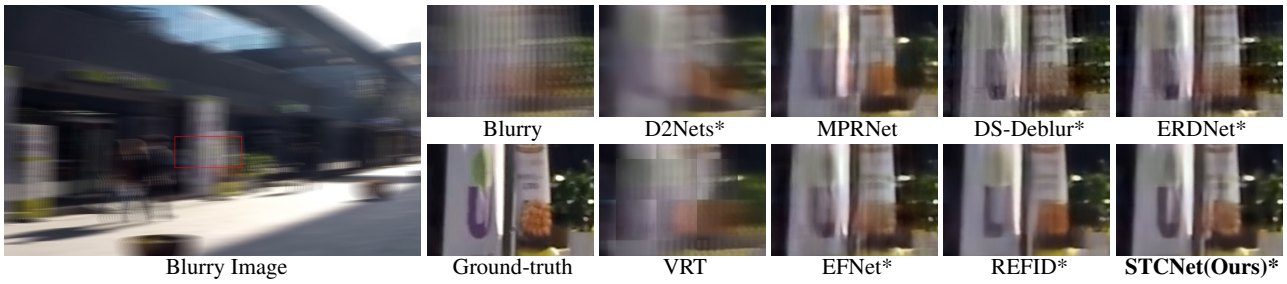

Figure 5: Visual comparisons on *GoPro* datatset. * denotes event-based methods. Best viewed on a screen and zoomed in.

Method	D2Nets*	MPRNet	eSL-Net*	DS-Deblur*	NAFNet	ERDNet*	VRT	EFNet*	REFID*	**STCNet(Ours)***
PSNR	26.64	27.80	27.50	31.63	27.94	32.29	31.94	32.85	33.15	**33.94**
SSIM	0.8819	0.9091	0.8914	0.9436	0.9126	0.9506	0.9602	0.9571	0.9611	**0.9692**

Table 2: Comparison of motion deblurring methods on *DVD* dataset. * denotes event-based methods.

Method	MMP-RNN	Restormer	D2Nets*	NAFNet	DS-Deblur*	ERDNet*	eSL-Net*	REFID*	EFNet*	**STCNet(Ours)***
PSNR	30.66	32.21	32.47	32.75	32.84	34.02	34.55	34.84	34.91	**35.44**
SSIM	0.9122	0.9505	0.9585	0.9570	0.9583	0.9663	0.9710	0.9723	0.9720	**0.9772**

Table 3: Comparison of motion deblurring methods on *REB* dataset. * denotes event-based methods.

method achieves the most visually plausible deblurring results with sharper textures while others produce results with more artifacts and cannot remove severe blur effectively.

Complexity Comparison

We further calculate the parameters and average runtime for complexity analysis. All experiments are conducted with image size of $1280 \times 720 \times 3$. Results in average running time and parameters are presented in Table 4. It is obvious that our method has comparable parameters and running time with consideration of acceptable calculation consumption to achieve promising deblurring performance.

Ablation Study

To evaluate the effectiveness of the key components (MSC and MTC) in our model, we conduct ablation studies on *GoPro* dataset and *REB* dataset. A baseline is first experimented with, which simply concatenates frame features Φ_t^B and event features Φ_t^E and neglects the spatio-temporal correlation between successive inputs. First row of Table 5 shows the performance of baseline.

Effectiveness of MSC Module. We append it to *Baseline* to conduct cross-modal fusion using a calibration-then-aggregation strategy. There is a great performance gap in the first two rows of Table 5, which shows that MSC can efficiently fuse events with frames. Then, we validate the importance of differential-modality guided cross-modal calibration (DM-CMC) strategy in MSC. The DM-CMC is appended to the *baseline* to calibrate Φ_t^B and Φ_t^E and the calibrated bi-modal features are simply concatenated, and the results are shown in the first two rows in Table 7, showing that DM-CMC can enhance complementarity. Further, the validity of CMA in MSC is tested, which is designed to weight different contribution of modalities. The CMA is appended to the *baseline* to adaptively fusion, ignoring modality redundancy problem, and the results are shown in the first and third rows in Table 7. Apparently, CMA can efficiently emphasize modality own importance for better fusion.

Effectiveness of MTC Module. We append MTC module to *baseline* to capture sharp information from temporal neighbors of both frames and events, and the results are shown in the first and third rows in Table 5. Apparently, cross-temporal relevance can be modeled by MTC to improve the deblurring performance. Then the mutual spatio-temporal attention scheme in MTC is validated. We model spatio-temporal relevance with only frames, only events,

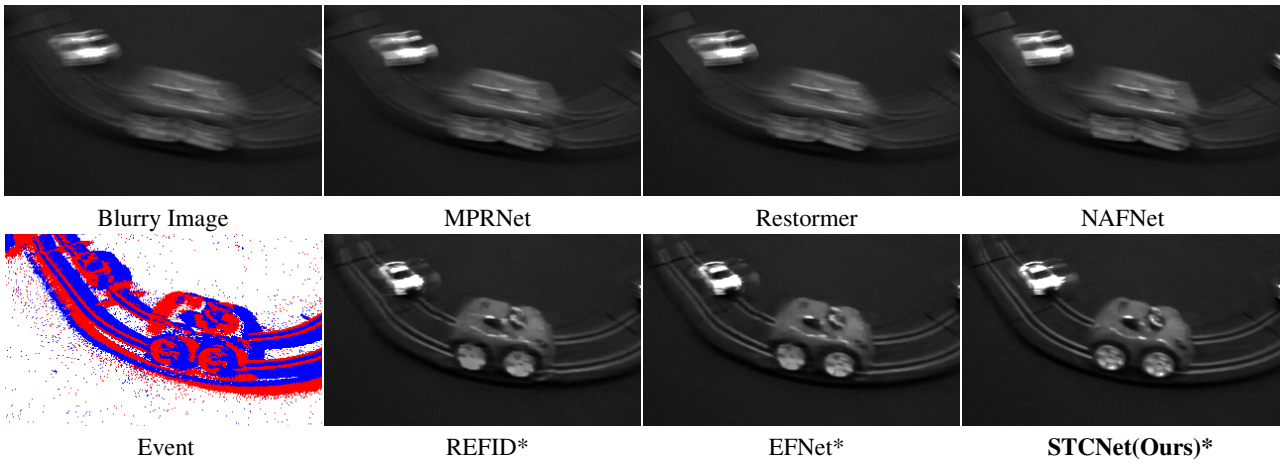

Figure 6: Visual comparison on real blur set of *REB* dataset. * denotes event-based methods. Best viewed on a screen and zoomed in.

Method	eSL-Net*	D2Nets*	MemDeblur	MPRNet	MIMO-UNet++	Restormer	DS-Deblur*	ERDNet*	REFID*	**STCNet(Ours)***
Params (M)	0.19	32.63	6.1020	20.10	16.10	26.09	15.60	18.08	15.9	16.25
Runtime (s)	0.015	1.340	0.9110	0.117	0.025	1.1546	0.292	0.020	0.072	0.098
PSNR(dB)	30.23	31.76	31.76	32.66	32.68	32.92	33.13	34.25	35.91	36.45

Table 4: Complexity comparison with other methods. * denotes event-based methods.

MSC	MTC	Gropo		REB	
		PSNR	SSIM	PSNR	SSIM
✗	✗	33.40	0.9615	33.12	0.9610
✓	✗	35.93	0.9780	35.00	0.9721
✗	✓	35.05	0.9733	34.20	0.9682
✓	✓	**36.45**	**0.9809**	**35.44**	**0.9772**

Table 5: Ablation study on MSC and MTC in STCNet.

Cross-frame	Cross-event	Gropo		REB	
		PSNR	SSIM	PSNR	SSIM
✗	✗	33.40	0.9615	33.12	0.9610
✓	✗	34.37	0.9681	33.58	0.9645
✗	✓	33.96	0.9658	33.15	0.9608
✓	✓	**34.61**	**0.9706**	**33.79**	**0.9652**

Table 6: Ablation study on mutual attention in MTC.

DM-CMC	CMA	Gropo		REB	
		PSNR	SSIM	PSNR	SSIM
✗	✗	33.40	0.9615	33.12	0.9610
✓	✗	34.73	0.9712	33.97	0.9667
✗	✓	35.67	0.9768	34.66	0.9703
✓	✓	**35.93**	**0.9780**	**35.00**	**0.9721**

Table 7: Ablation study on calibration-aggregation in MSC.

CTF	Gropo		REB	
	PSNR	SSIM	PSNR	SSIM
✗	34.61	0.9706	33.79	0.9652
✓	**35.05**	**0.9733**	**34.20**	**0.9682**

Table 8: Ablation study on cross-temporal fusion in MTC.

joint them by STA, and the spatio-temporal features and current features are simply concatenated. Table 6 shows that mutual attention better captures spatio-temporal dependence. Besides, we test the validity of the CTF, which adaptively fuses above mutual attention-guided spatio-temporal features and current features. Table 8 shows the superiority of CTF.

Conclusion

In this work, we explore the complementary fusion of events and frames for motion deblurring. A novel spatial-temporal collaboration network is introduced to facilitate the cross-modal fusion both in spatial and temporal aspects. We first conduct cross-modal spatial fusion with first differential-modality guided cross-modal calibration for complementary enhancement and then co-attention based cross-modal aggregation for adaptive fusion. And then to attach importance to the temporal correlation among adjacent neighbors, we prepose the frame-event mutual spatio-temporal attention for cross-temporal dependencies modeling and then fuse spatio-temporal features with a cross-temporal coordinate attention based cross-temporal fusion. Extensive evaluations show that our method achieves state-of-the-art performance.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under contract 62022063.

References

Bahat, Y.; Efrat, N.; and Irani, M. 2017. Non-uniform blind deblurring by reblurring. In *ICCV*.

Bar, L.; Berkels, B.; Rumpf, M.; and Sapiro, G. 2007. A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *ICCV*.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*.

Chen, H.; Teng, M.; Shi, B.; Wang, Y.; and Huang, T. 2022a. A Residual Learning Approach to Deblur and Generate High Frame Rate Video With an Event Camera. *IEEE TMM*.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022b. Simple baselines for image restoration. *ECCV*.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. HINet: Half instance normalization network for image restoration. In *CVPR*.

Cho, S.; Wang, J.; and Lee, S. 2012. Video deblurring for hand-held cameras using patch-based synthesis. *ACM TOG*.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking Coarse-to-Fine Approach in Single Image Deblurring. In *ICCV*.

Hou, Q.; Zhou, D.; and Feng, J. 2021. Coordinate attention for efficient mobile network design. In *CVPR*.

Hyun Kim, T.; and Mu Lee, K. 2015. Generalized video deblurring for dynamic scenes. In *CVPR*.

Ji, B.; and Yao, A. 2022. Multi-Scale Memory-Based Video Deblurring. In *CVPR*.

Jiang, Z.; Zhang, Y.; Zou, D.; Ren, J.; Lv, J.; and Liu, Y. 2020. Learning event-based motion deblurring. In *CVPR*.

Jin, H.; Favaro, P.; and Cipolla, R. 2005. Visual tracking in the presence of motion blur. In *CVPR*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023. Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring. In *CVPR*.

Kotera, J.; Šroubek, F.; and Milanfar, P. 2013. Blind deconvolution using alternating maximum a posteriori estimation with heavy-tailed priors. In *CAIP*.

Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; and Matas, J. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*.

Kupyn, O.; Martyniuk, T.; Wu, J.; and Wang, Z. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*.

Levin, A.; Weiss, Y.; Durand, F.; and Freeman, W. T. 2009. Understanding and evaluating blind deconvolution algorithms. In *CVPR*.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2022a. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *ICCV*.

Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; and Van Gool, L. 2022b. Recurrent Video Restoration Transformer with Guided Deformable Attention. *NeurIPS*.

Lin, J.; Cai, Y.; Hu, X.; Wang, H.; Yan, Y.; Zou, X.; Ding, H.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. Flow-guided sparse transformer for video deblurring. *ICML*.

Lin, S.; Zhang, J.; Pan, J.; Jiang, Z.; Zou, D.; Wang, Y.; Chen, J.; and Ren, J. 2020. Learning event-driven video deblurring and interpolation. In *ECCV*.

Matsushita, Y.; Ofek, E.; Ge, W.; Tang, X.; and Shum, H.-Y. 2006. Full-frame video stabilization with motion inpainting. *IEEE TPAMI*.

Mei, C.; and Reid, I. 2008. Modeling and generating complex motion blur for real-time tracking. In *CVPR*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*.

Nah, S.; Son, S.; and Lee, K. M. 2019. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*.

Pan, J.; Bai, H.; and Tang, J. 2020. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*.

Pan, J.; Xu, B.; Dong, J.; Ge, J.; and Tang, J. 2023. Deep Discriminative Spatial and Temporal Network for Efficient Video Deblurring. In *CVPR*.

Pan, L.; Scheerlinck, C.; Yu, X.; Hartley, R.; Liu, M.; and Dai, Y. 2019. Bringing a blurry frame alive at high frame-rate with an event camera. In *CVPR*.

Park, D.; Kang, D. U.; Kim, J.; and Chun, S. Y. 2020. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*.

Purohit, K.; and Rajagopalan, A. 2020. Region-adaptive dense network for efficient motion deblurring. In *AAAI*.

Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. ESIM: an open event camera simulator. In *CoRL*.

Scheerlinck, C.; Barnes, N.; and Mahony, R. 2018. Continuous-time intensity estimation using event cameras. In *ACCV*.

Shang, W.; Ren, D.; Zou, D.; Ren, J. S.; Luo, P.; and Zuo, W. 2021. Bringing Events Into Video Deblurring With Non-Consecutively Blurry Frames. In *ICCV*.

Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; and Wang, O. 2017. Deep video deblurring for hand-held cameras. In *CVPR*.

Suin, M.; Purohit, K.; and Rajagopalan, A. 2020. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*.

Sun, L.; Sakaridis, C.; Liang, J.; Jiang, Q.; Yang, K.; Sun, P.; Ye, Y.; Wang, K.; and Van Gool, L. 2022. Event-Based Fusion for Motion Deblurring with Cross-modal Attention. In *ECCV*.

Sun, L.; Sakaridis, C.; Liang, J.; Sun, P.; Cao, J.; Zhang, K.; Jiang, Q.; Wang, K.; and Van Gool, L. 2023. Event-Based Frame Interpolation with Ad-hoc Deblurring. In *CVPR*.

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *CVPR*.

Tsai, F.-J.; Peng, Y.-T.; Tsai, C.-C.; Lin, Y.-Y.; and Lin, C.-W. 2022. BANet: A Blur-aware Attention Network for Dynamic Scene Deblurring. *IEEE TIP*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vitoria, P.; Georgoulis, S.; Tulyakov, S.; Bochicchio, A.; Erbach, J.; and Li, Y. 2023. Event-Based Image Deblurring with Dynamic Motion Awareness. In *ECCVW*.

Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020. Event enhanced high-quality image recovery. In *ECCV*.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*.

Wang, Y.; Lu, Y.; Gao, Y.; Wang, L.; Zhong, Z.; Zheng, Y.; and Yamashita, A. 2022a. Efficient video deblurring guided by motion magnitude. *ECCV*.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022b. Uformer: A general u-shaped transformer for image restoration. In *CVPR*.

Wulff, J.; and Black, M. J. 2014. Modeling blurred video with layers. In *ECCV*.

Xiang, X.; Wei, H.; and Pan, J. 2020. Deep video deblurring using sharpness features from exemplars. *IEEE TIP*.

Xu, F.; Yu, L.; Wang, B.; Yang, W.; Xia, G.-S.; Jia, X.; Qiao, Z.; and Liu, J. 2021. Motion Deblurring with Real Events. In *ICCV*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; Dong, W.; and Shi, G. 2022. Learning for Motion Deblurring with Hybrid Frames and Events. In *ACM MM*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *CVPR*.

Zhang, H.; Dai, Y.; Li, H.; and Koniusz, P. 2019. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*.

Zhang, X.; and Yu, L. 2022. Unifying Motion Deblurring and Frame Interpolation with Events. In *CVPR*.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2020. Residual dense network for image restoration. *IEEE TPAMI*.

Zhong, Z.; Gao, Y.; Zheng, Y.; and Zheng, B. 2020. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*.

Zhou, S.; Zhang, J.; Pan, J.; Xie, H.; Zuo, W.; and Ren, J. 2019. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*.

Zhu, C.; Dong, H.; Pan, J.; Liang, B.; Huang, Y.; Fu, L.; and Wang, F. 2022. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *AAAI*.