Event-based Motion Deblurring with Modality-Aware Decomposition and Recomposition

Wen Yang
wen.yang@stu.xidian.edu.cn
School of Artificial Intelligence, Xidian University

Jinjian Wu*
jinjian.wu@mail.xidian.edu.cn
School of Artificial Intelligence, Xidian University

Leida Li
ldli@xidian.edu.cn
School of Artificial Intelligence, Xidian University

Weisheng Dong
wsdong@mail.xidian.edu.cn
School of Artificial Intelligence, Xidian University

Guangming Shi
gmshi@xidian.edu.cn
School of Artificial Intelligence, Xidian University

ABSTRACT

Event cameras, offering visual information with microsecond accuracy and having strong robustness against motion blur, provide a new perspective to address motion deblurring. How to effectively exploit the collaboration of events and images for motion deblurring is a challenging endeavor. Existing event-based motion deblurring methods perform cross-modal fusion with modality-specific features (complementarity), while ignoring features shared by modalities (correlation), which may lead to insufficient fusion of event and image, resulting in limited performance. To address the above issues, following the idea of divide and conquer, we tackle the challenge in modeling cross-modality fusion with the modality-specific and modality-shared features decomposition and recomposition. Therefore, we propose a novel event-image fusion network (EIFNet) based on modality-aware decomposition and recomposition. Specifically, in the decomposition stage, modality-shared and modality-specific feature separation clues are inferred in parallel by exploring the global correlation of common-mode, differential-mode and two modalities with dual cross-attention. In the recomposition stage, the divided modality-shared and modality-specific features are merged with bi-directional supplement information exchanging via long-range interaction. Extensive experiments demonstrate that our method outperforms state-of-the-art event-driven and image-only methods. Project website: https://github.com/wyang-vis/EIFNet.

CCS CONCEPTS

• **Computing methodologies** → **Computational photography**; *Reconstruction.*

KEYWORDS

Event-based vision, motion deblurring, cross-modal fusion, deep neural network

ACM Reference Format:
Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi. 2023. Event-based Motion Deblurring with Modality-Aware Decomposition and Recomposition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612505

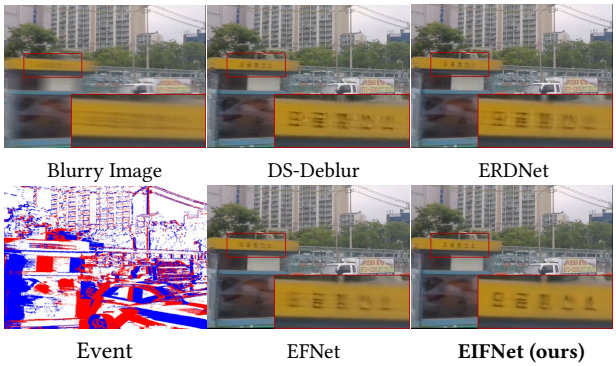


Figure 1: Visual comparison of deblurring with state-of-the-art event-based methods DS-Deblur [50], ERDNet [5], EFNet [37], and our EIFNet.

1 INTRODUCTION

Motion blur often occurs due to the relative motion between the camera and scene during the image integration time. Motion deblurring is one of the critical and challenging topics, which aims to restore the clean image from its blurry version. It is an ill-posed inverse problem, due to the existence of many possible solutions. Conventional methods are mainly based on hand-crafted priors and assumptions [1, 2, 10, 12, 15, 18, 19, 22, 45], which limit the model capacity. With the development of deep learning in many vision tasks, deep neural network (DNN)-based approaches have been proposed to learn the implicit relation from blurry images to sharp images under the supervision of a large-scale dataset of

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612505

blurry-sharp image pairs [7, 8, 29, 31, 36, 38, 52, 53]. Although the abovementioned methods have achieved considerable performance, they may fail to deal with severe blur due to the significant loss of motion information.

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 [13] if the intensity changes surpass a threshold. Understandably, with the attractive properties that offer visual information with microsecond accuracy and have strong robustness against motion blur, event cameras can be attempted to address motion deblurring. Naturally, how to design an efficient cross-modality feature fusion mechanism is the most important step in event-based motion deblurring. Recently, both model-driven [30] and data-driven [3, 5, 16, 27, 35, 37] algorithms have been proposed to recover the sharp image from a blurry image with the aid of events. Generally, different modalities are expected to have some shared features and also specific features [11]. Despite the remarkable progress, existing event-based deblurring methods exploit cross-modal fusion with modality-specific information exploration, while without taking the exploration and exploitation of modality-shared information into consideration, which may lead to insufficient fusion between event and image, resulting in limited performance, even inferior to image-only algorithms.

In this work, following the idea of divide and conquer, we consider decomposing modality-specific and modality-shared features from the two modalities and then recomposing them to tackle the challenges in cross-modal fusion. To this end, a novel event-image fusion network (EIFNet) based on modality-aware decomposition and recomposition is proposed for motion deblurring. To the best of our knowledge, this is the first time that joint modality-specific and modality-shared features are leveraged to event-based image deblurring. Specifically, intermediate features of image and event are first extracted with feature extractor respectively. Then, a modality-aware decomposition (MAD) module is built to divide modality-shared and modality-specific features, in which modality-shared features are separated by exploring the correlation of common-mode and two modalities features with mutual cross-attention, while modality-specific features are detached by exploring the correlation of differential-mode and two modalities features with separate cross-attention. Next, a modality-aware recomposition (MAR) module is built to merge the divided modality-shared and modality-specific features, where different types of features are interacted and aggregated in a bi-directional propagation manner, which transfers supplement information from shared-modality to specific-modality and vice versa by long-range interaction, i.e., shared-induced specific complement and specific-induced shared complement. Finally, a reconstruction model is adopted to reconstruct the target image from the output of the MAR module. Extensive experiments on both synthetic and real-world datasets demonstrate our method achieves state-of-the-art performance (some visual comparisons are shown in Figure 1).

The main contributions of our work are as follows.

- We design a novel modality-aware decomposition and recomposition based event-image fusion network (EIFNet), which properly fuses the events and images with joint shared-modality and specific-modality attentions. Extensive experiments show that our model outperforms state-of-the-art event-driven and image-only methods.
- We propose a novel MAD module to enforce the modality-shared and modality-specific features decomposition with dual cross-attention, which parallel infers shared feature cues with mutual cross-attention between common-mode and two modalities, and specific feature cues with separate cross-attention across differential-mode and two modalities.
- We propose a new MAR module to realize the divided modality-shared and modality-specific features recomposition in a bi-directional supplement manner, which transfers specificity through shared-induced specific complement and specific-induced shared complement with long-range interaction.

The remainder of this paper is structured as follows. In Section 2, we review the related works on motion deblurring. Section 3 describes the details of the proposed model. Experimental results and analysis are presented in Section 4. Finally, we conclude this paper with a discussion in Section 5.

2 RELATED WORK

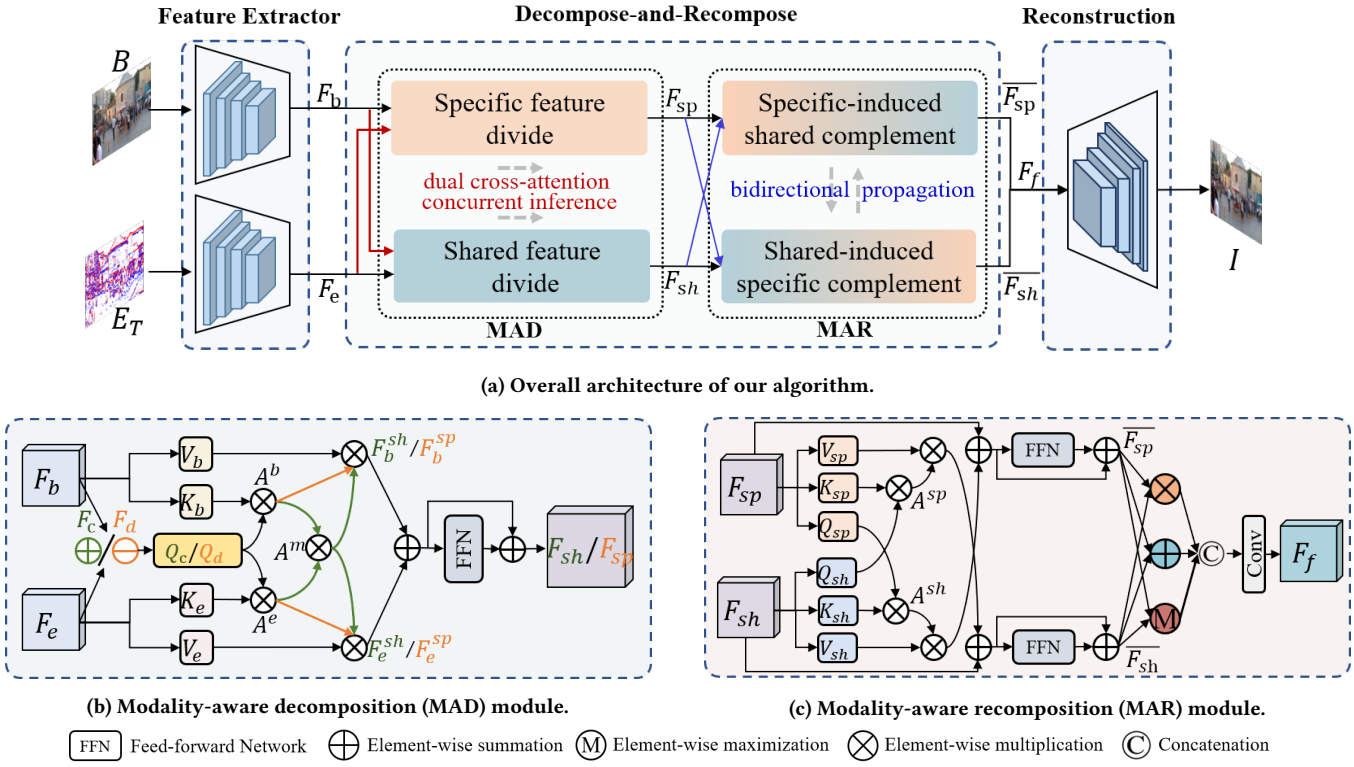
In this section, we briefly review the literature related to our method, including some Image/frame-based and Event-based motion deblurring methods.

2.1 Image Deblurring

Image deblurring is a highly ill-posed problem, which aims to recover a latent sharp image from a blurry image. Conventional image deblurring approaches are usually based on hand-crafted priors and assumptions [1, 2, 10, 12, 15, 18, 19, 22, 45], which lead to limited generality and representing capacity. Recently, many deep neural network (DNN)-based methods have been proposed and brought significant changes to image deblurring. They implicitly learn the relation from blurry images to sharp images. Several novel components and techniques have been proposed: 1) Single-Scale Networks. The single-scale deblurring methods [20, 21, 56] aim to recover highly-realistic images mainly based on well-developed network blocks for high-level vision tasks. 2) Multi-Scale Networks. These methods decompose deblurring task into smaller easier subtasks to recover clean image in a progressive manner [7, 29, 36, 38, 52, 53]. 3) Coarse-to-Fine Strategies. The coarse-to-fine schemes can gradually restore a sharp image with multiple input images on different resolutions [8, 31]. 4) Attention mechanism. Spatial attention modules, and channel attention modules or both have also been incorporated to selectively attend to relevant information for image deblurring [24, 32, 36, 39, 42, 44, 51]. Although the abovementioned methods have achieved considerable performance, image deblurring is a highly ill-posed problem with infinite feasible solutions that cannot be trivially addressed from only the blur set of input.

2.2 Event-based 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



(a) Overall architecture of our algorithm.

(b) Modality-aware decomposition (MAD) module.

(c) Modality-aware recomposition (MAR) module.

Figure 2: Framework of the proposed modality-aware decomposition and recomposition based event-image fusion network (EIFNet). (a) Overall architecture of our algorithm. (b) Details of the modality-aware decomposition (MAD) module. (c) Details of the modality-aware recomposition (MAR) module.

methods can be divided into two categories [47], i.e., model-driven and data-driven algorithms.

Model-driven methods use the physical event generation principle to relate events, blurry images and the latent sharp images [30, 34]. Specifically, BHA [30] proposed an Event-based Double Integral (EDI) algorithm to model the blur-generation process by associating events to a latent frame. CF [34] proposed a continuous-time formulation of event-based intensity estimation with complementary filtering. However, events are essentially with noise both in temporal and spatial domains [54], which inevitably degrades performance.

Data-driven methods directly learn the relation from blurry images to sharp images with the aid of events [27]. LEMD [16] presented a sequential formulation of event-based motion deblurring, and unfolded its optimization with deep network. eSL-Net [43] proposed an event-enhanced degeneration model for high-quality image recovery. D2Net [35] proposed an event-frame fusion module, which can be incorporated into existing image deblurring methods. MADANET+ [49] predicted the high blur image areas guided by events for deblurring. EFNet [37] proposed a cross-modal attention module to fuse image and event. ERDNet [5] presented a residual learning approach to learn event-based motion deblurring. DS-Deblur [50] proposed a dual-stream based event-image fusion framework for motion deblurring.

Nevertheless, some of the above methods only achieve slight performance gains compared to image-only methods, due to the insufficient cross-modal complementary fusion of events and images. In this work, a novel event-image fusion network is proposed for motion deblurring with modality-specific and modality-shared features decomposition and recomposition.

2.3 Vision Transformer

Transformer is proposed in [41] for machine translation and has become the most advanced method in natural language processing (NLP) tasks. Due to the powerful capability for capturing long-range dependencies in the data by the global self-attention, numerous researchers try to combine the transformer structure in computer vision tasks such as object detection [28], segmentation [46]. Motivated by the great success in high-level vision, Transformer models have also been studied for low-level vision problems such as denoising [25, 44], deblurring [23, 26, 48]. In this work, we take the advantage of Transformer in global dependency capturing for modality-aware decomposition and recomposition.

3 METHOD

Problem formulation: Given a blurry image $\mathbf{B}$ and the corresponding event stream $\mathbf{E}_T \triangleq \{(x_i, y_i, p_i, t_i)\}_{t_i \in T}$ containing all asynchronous 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 image $\mathbf{I}$ by exploiting both blurry image $\mathbf{B}$ and event stream $\mathbf{E}_T$, which can be modeled as $\mathbf{I} = g_{\theta^*}(\mathbf{B}, \mathbf{E}_T)$, where g_{θ^*} is deep learning model.

3.1 System Overview of EIFNet

We present EIFNet, a modality-aware decomposition and recomposition based event-image fusion network for event-based image deblurring. The overview of EIFNet is shown in Figure 2(a). We first use two parallel backbones, which contain Channel Attention Blocks (CABs) [55] and down-sampling layers, to extract features F_b and F_e from blurry image $\mathbf{B}$ and corresponding event $\mathbf{E}_T$, separately. Next, a decomposition and recomposition strategy is proposed to divide modality-shared and modality-specific features via a modality-aware decomposition (MAD) module, and then the divided modality-shared and modality-specific features are merged with a modality-aware recomposition (MAR) module, obtaining F_f. Finally, a reconstruction module, containing Channel Attention Blocks (CABs) and up-sampling layers, is used to transform F_f to the final deblurred result $\mathbf{I}$. Below we detail the main part: modality-aware decomposition and recomposition.

3.2 Modality-aware Decomposition and Recomposition

Generally, different sensing modalities usually have some shared features and also have specific features. Following the idea of divide and conquer, and inspired by differential amplifier circuits in which the common-mode signals are suppressed and the differential-mode signals are amplified, we consider decomposing different features from multi-modalities and then recomposing them.

3.2.1 *Modality-aware Decomposition (MAD) Module.*

According to the principle of the differential amplifier, common-mode part F_c and differential-mode part F_d can be represented as follows:

$$\begin{aligned} F_c &= F_b + F_e, \\ F_d &= F_b - F_e. \end{aligned} \tag{1}$$

Intuitively, for cross-modal fusion, both common-mode and differential-mode should be selected and enhanced. We can select the most effective features of image and event by exploring the relevance of common-mode and differential-mode with image and event features, and remix them into new enhanced modality-shared and modality-specific features.

To this end, a new MAD module is proposed to leverage dual cross-attention between common-mode and differential-mode with event and image features for correlation calculation to select modality-shared and modality-specific features. Figure 2(b) shows the details of MAD (the **black path** is the shared backbone). First, following the basic idea of transformer, the image features F_b and event features F_e are transformed into Key K_b, Value V_b and Key K_e, Value V_e, respectively. And the common-mode and differential-mode (F_c and F_d) are transformed into Query Q_c and Q_d, respectively. Then, we can conduct feature separation by communicating the Query from F_c or F_d and the Key from F_b and F_e.

Modality-specific features extraction (the orange path in Figure 2(b)). We first estimate specific features clues of image and event with separate cross-attention, which multiplies the Query from F_d and the Key from F_b and F_e:

$$\begin{aligned} A^b_{sp} &= Softmax\left(Q_d K_b^T\right), \\ A^e_{sp} &= Softmax\left(Q_d K_e^T\right), \end{aligned} \tag{2}$$

where the attention maps A^b_{sp} and A^e_{sp} contain the specific features clues of image and event, respectively. Then, we multiply the global attention maps A^b_{sp} and A^e_{sp} with Value V_b and V_e respectively to obtain respective modality-specific features of the image and event, and it is depicted in the following equations:

$$F^{sp}_b = A^b_{sp} V_b, F^{sp}_e = A^e_{sp} V_e, \tag{3}$$

where F^{sp}_b and F^{sp}_e are the modality-specific features of the image and event, respectively. In the end, sum the F^{sp}_b and F^{sp}_e to get the remixed total modality-specific features F_{sp}:

$$F_{sp} = \left(F^{sp}_b + F^{sp}_e\right) + \mathrm{FFN}\left(F^{sp}_b + F^{sp}_e\right), \tag{4}$$

where FFN denotes the two fully-connected layers with a non-linearity activation function GELU.

Modality-shared features extraction (the green path in Figure 2(b)). Similar to modality-specific features extraction, shared features clues of image and event are first computed by multiplying the Query from F_c and the Key from F_b and F_e:

$$\begin{aligned} A^b_{sh} &= Softmax\left(Q_c K_b^T\right), \\ A^e_{sh} &= Softmax\left(Q_c K_e^T\right), \end{aligned} \tag{5}$$

where the attention maps A^b_{sh} and A^e_{sh} contain the shared features clues of image and event, respectively. Then, different from specific feature extraction using respective attention, the modality-shared features are extracted using mutual attention, which is formulated as:

$$A^m_{sh} = A^b_{sh} A^e_{sh}, \tag{6}$$

where the mutual attention maps A^m_{sh} contain the common global clues of shared features for both image and event. Next, we multiply A^m_{sh} to Value V_b and V_e to acquire respective modality-shared features F^{sh}_b and F^{sh}_e of the image and event. Finally, sum the F^{sh}_b and F^{sh}_e to get the remixed total modality-shared features F_{sh}. This procedure can be formulated as:

$$\begin{aligned} F^{sh}_b &= A^m_{sh} V_b, F^{sh}_e = A^m_{sh} V_e, \\ F_{sh} &= \left(F^{sh}_b + F^{sh}_e\right) + \mathrm{FFN}\left(F^{sh}_b + F^{sh}_e\right). \end{aligned} \tag{7}$$

3.2.2 *Modality-aware Recomposition (MAR) Module.*

After obtaining the modality-specific features F_{sp} and modality-shared features F_{sh}, we build the MAR module to recompose them a bi-directional supplement manner, which transfers supplement information from shared-modality to specific-modality and vice versa by conducting long-range interaction between shared and specific features. Figure 2(c) shows the details of MAR.

To be specific, given the F_{sp} and F_{sh}, we first transform F_{sp} into Query Q_{sp}, Key K_{sp}, Value V_{sp}, and F_{sh} into Query Q_{sh}, Key K_{sh}, Value V_{sh}. We first estimate the information exchange attention through a bidirectional cross-attention between vectorized features from the F_{sp} and F_{sh} via:

$$\begin{aligned} A^{sp} &= Softmax\left(Q_{sh}K_{sp}^{T}\right), \\ A^{sh} &= Softmax\left(Q_{sp}K_{sh}^{T}\right), \end{aligned} \tag{8}$$

where A^{sp} denotes the supplement information clues from F_{sp} for F_{sh}, and A^{sh} denotes the supplement information clues from F_{sh} for F_{sp}.

Then, sufficient information supplement is implemented by multiplying the A^{sp} and A^{sh} with Key K_{sp} and K_{sh} respectively:

$$\begin{aligned} \overline{F_{sp}} &= \left(F_{sp} + A^{sh}V_{sh}\right) + \mathrm{FFN}\left(F_{sp} + A^{sh}V_{sh}\right), \\ \overline{F_{sh}} &= \left(F_{sh} + A^{sp}V_{sp}\right) + \mathrm{FFN}\left(F_{sh} + A^{sp}V_{sp}\right), \end{aligned} \tag{9}$$

where the $\overline{F_{sp}}$ and $\overline{F_{sh}}$ denote the features that have been complemented by each other after bi-directional information supplement.

Moreover, the features $\overline{F_{sp}}$ and $\overline{F_{sh}}$ are attentively merged through three fusion strategies, including element-wise summation $F^{+} = \overline{F_{sp}} + \overline{F_{sh}}$, element-wise product $F^{\times} = \overline{F_{sp}} \times \overline{F_{sh}}$ and element-wise maximization $F^{m} = \max\left(\overline{F_{sp}}, \overline{F_{sh}}\right)$. Finally, a convolutional layer is added to learn the contribution weight of the concatenated three fusions for the final fusion F_f:

$$F_f = Conv\left(Cat\left(F^{+}, F^{\times}, F^{m}\right)\right). \tag{10}$$

3.3 Loss Function

In this paper, we use the Charbonnier loss [4] to train our network in an end-to-end fashion:

$$L_{\text{char}} = \frac{1}{CHW}\sqrt{\|I - G\|^2 + \varepsilon^2}, \tag{11}$$

where I and G is deblurred out and ground truth, respectively, C, H, W are dimensions of image, and constant ε is empirically set to 10^{-3} as in [52] for all the experiments.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. Our EIFNet is evaluated on two datasets: 1) *GoPro*. GoPro dataset [29] is widely adopted for image-only and event-based deblurring, which contains synthetic blurring images and sharp clear ground-truth images, as well as synthetic events generated by simulation algorithm ESIM [33]. The blurry image is offered by averaging nearby (the number varies from 7 to 13) images. We follow the suggested training-testing split, 22 videos in GoPro dataset [29] are utilized for training and 11 for testing. 2) *REB*. For evaluation in real-world events, the REB dataset is a real event dataset captured by us with the DAVIS346 event camera. The REB dataset captures both real-world events and clear ground-truth images with slow camera motion in relatively stationary scenes or with a stationary camera in slow-motion scenes 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 dataset. 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 network is implemented using Pytorch. For training, we use the Adam optimizer [17] with standard settings, batch size is 8, patch size is 256× 256 and learning rate is 2×10^{-4} 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 training ends after 200k iterations for GoPro dataset and 100k iterations for REB dataset. The Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) are adopted as the evaluation metrics.

4.2 Comparison with State-of-the-Art Methods

We compare our EIFNet with state-of-the-art image-only deblurring methods, including DMCNN [29], MTRNN [31], DMPHN [53], DSD [14], MPRNet [52], MIMO-UNet++ [8], HINNet [7], MAXIM [40], Restormer [51], U-former [44], MPRNet-Local [9], NAFNet [6], and event-based deblurring methods, including RED [47], eSL-Net [43], D2Nets [35], LEMD [16], DS-Deblur [50], MADANET+ [49], ERD-Net [5], EFNet [37], on GoPro and REB.

***GoPro* dataset:** Table 1 reports the quantitative results on synthetic *GoPro* dataset. Compared to the best existing image-only deblurring methods, our method achieves outstanding performance improvements (2.3 dB improvement in PSNR), demonstrating the advantages of event-assisted deblurring than purely relying on image-only. Despite utilizing an extra modality, some event-based methods such as D2Nets [35], LEMD [16], eSL-Net [43] and DS-Deblur [50] do not improve significantly upon image-only methods, indicating that they do not effectively conduct cross-modal complementary fusion. Our EIFNet achieves the best performance against other event-based deblurring methods (by a margin of 0.53dB), showing the superiority of our decomposition-and-recomposition based fusion. We show in Figure 3 a visual comparison between our method and several state-of-the-art methods. The proposed method recovers the sharpest details, while the results restored by other methods still suffer from motion blur, losing sharp edge information.

***REB* dataset:** For evaluation in real-world events, we report quantitative results on *REB* dataset in Table 2. Note that for a fair comparison, we retrain several image-only methods and event-guided methods using the publicly available code provided by the authors. It is clear from Table 2 that our method significantly outperforms all other state-of-the-art competitors. Besides, one of the main drawbacks of state-of-the-art methods is the lack of generalization to real blur. Figure 4 shows the qualitative comparisons on the synthetic blur set in *REB* dataset and Figure 5 shows qualitative comparisons on the real blur set in *REB* dataset. The image-only methods do not perform well on these severe cases of real-world motion blur and event-based methods are more robust to such adverse conditions with the aid of events. Remarkably, compared to existing event-based methods, our method achieves the most visually plausible deblurring results with sharper textures while others

Table 1: Quantitative comparison with state-of-the-art methods on *GoPro* dataset. * denotes event-based methods.

Method	RED* [47]	DMCNN [29]	eSL-Net* [43]	MTRNN [31]	DMPHN [53]	DSD [14]	D2Nets* [35]
PSNR	28.98	29.08	30.23	31.15	31.20	31.58	31.76
SSIM	0.8499	0.9135	0.8703	0.9450	0.9453	0.9478	0.9430
Method	LEMD* [16]	MPRNet [52]	MIMO-UNet++ [8]	HINNet [7]	MAXIM [40]	Restormer [51]	U-former [44]
PSNR	31.79	32.66	32.68	32.71	32.86	32.92	33.06
SSIM	0.9490	0.9590	0.9590	0.9590	0.9610	0.9610	0.9670
Method	DS-Deblur* [50]	MPRNet-Local [9]	NAFNet [6]	MADANET+* [49]	ERDNet* [5]	EFNet* [37]	**EIFNet***
PSNR	33.13	33.31	33.69	33.84	34.25	35.46	**35.99**
SSIM	0.9465	0.9640	0.9670	0.9640	0.9534	0.9720	**0.9785**

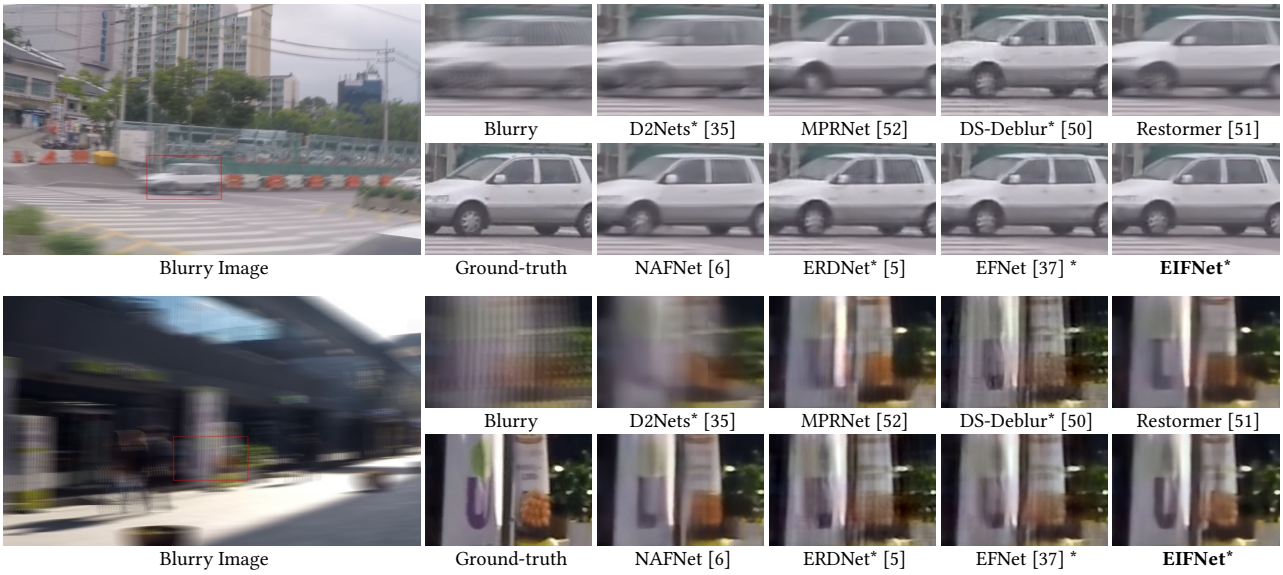


Figure 3: Visual comparisons on the *GoPro* datatset. * denotes event-based methods. Best viewed on a screen and zoomed in.

Table 2: Quantitative comparison with state-of-the-art methods on *REB* dataset. * denotes event-based methods.

Method	DMCNN [29]	DMPHN [53]	MPRNet [52]	MIMO-UNet++ [8]	MPRNet-local [9]	Restormer [51]	U-former [44]
PSNR	29.46	30.02	31.72	31.85	31.96	32.21	32.33
SSIM	0.9216	0.9273	0.9447	0.9500	0.9470	0.9505	0.9527
Method	D2Nets* [35]	NAFNet [6]	DS-Deblur* [50]	ERDNet* [5]	eSL-Net* [43]	EFNet* [37]	**EIFNet***
PSNR	32.47	32.75	32.84	34.02	34.55	34.91	**35.26**
SSIM	0.9585	0.9570	0.9583	0.9663	0.9710	0.9720	**0.9737**

produce results with more artifacts and cannot remove severe blur effectively.

4.3 Complexity Comparison

We calculate the parameters and average runtime for complexity analysis. All experiments are conducted on NVIDIA GeForce GTX 1080 with image size of $1280 \times 720 \times 3$. Results of average runtime and parameters are presented in Table 3. It is obvious that our method has comparable parameters and runtime with consideration of acceptable calculation consumption to achieve promising deblurring performance.

4.4 Ablation Study

To evaluate the effectiveness of the key components (MAD and MAR) in our model, we conduct ablation studies on *GoPro* dataset and *REB* dataset. Before that, a *Baseline* version is set to include only the feature extractor and reconstruction modules, which simply concatenates the intermediate features F_b and F_e along the channel-level. First row of Table 4 shows the performance of *Baseline*.

Effectiveness of MAD module. To demonstrate the overall effectiveness of MAD module, we append it to *Baseline* to decompose the modality-specific features F_{sp} and modality-shared features F_{sh} from F_b and F_e, but the decomposed features F_{sp} and

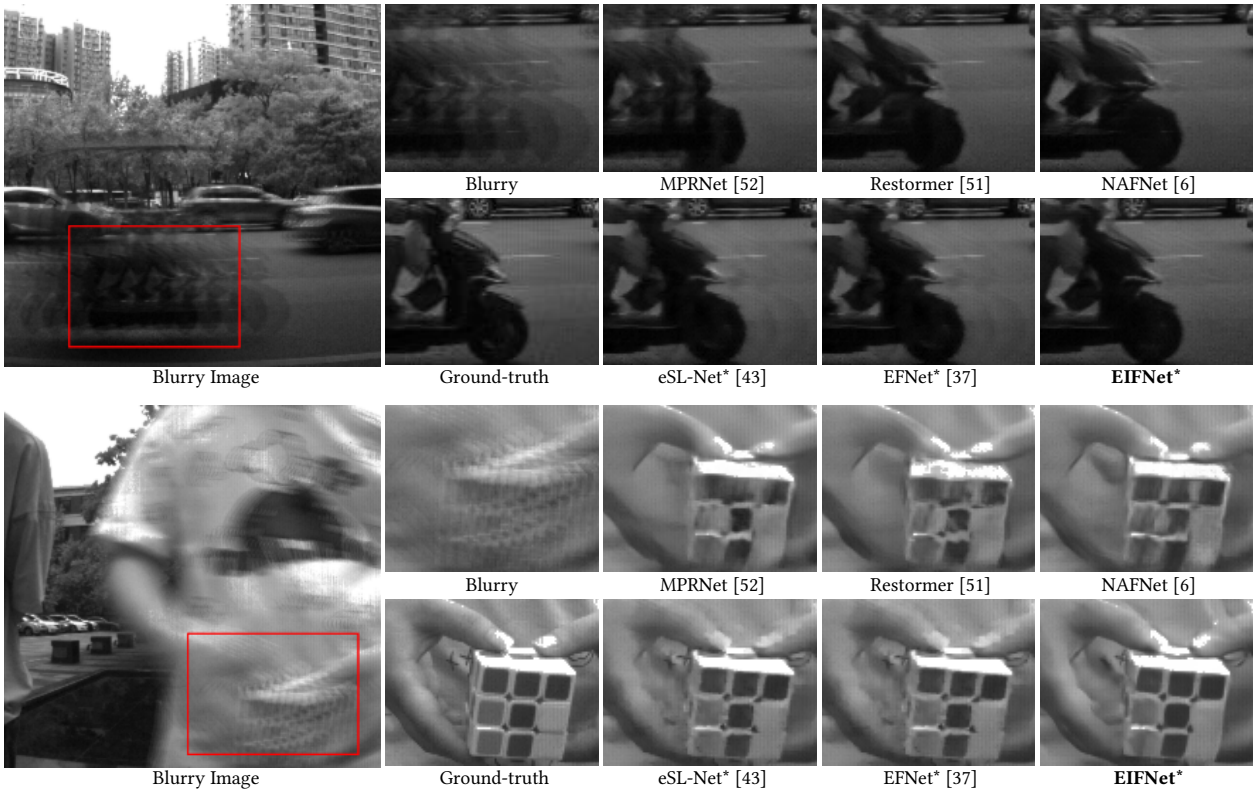


Figure 4: Visual comparison on the synthetic blur set in *REB* dataset. * denotes event-based methods. Best viewed on a screen and zoomed in.

Table 3: Complexity comparison with other methods. * denotes event-based methods.

Method	eSL-Net* [43]	MTRNN [31]	DSD [14]	D2Nets* [35]	MPRNet [52]	MIMO-UNet++ [8]
Params (M)	0.19	2.64	6.64	32.63	20.10	16.10
Runtime (s)	0.015	0.915	1.311	1.340	0.117	0.025
PSNR(dB)	30.23	31.15	31.58	31.76	32.66	32.68
Method	HINNet [7]	Restormer [51]	DS-Deblur* [50]	NAFNet [6]	ERDNet* [5]	**EIFNet***
Params (M)	88.79	26.09	15.60	67.78	18.08	10.82
Runtime (s)	0.508	1.1546	0.292	0.003	0.020	0.025
PSNR(dB)	32.71	32.92	33.13	33.69	34.25	35.99

F_{sh} are fused by the concatenation. There is a great performance gap in the first two rows of Table 4, which shows that different processing of different features is helpful and MAD can be competent for the decomposition task. To further verify the validity of MAD, we compared the performance of modality-specific/shared features (F_{sp}/F_{sh}), common/differential-mode (F_c/F_d) and original image/event features (F_b/F_e). All three types of features are simply concatenated for reconstruction, and the comparison results are shown in Table 5. Table 5 shows the effectiveness of modality-shared/specific features extracted by MAD compared to the direct use of common-differential mode and original image/event features.

Effectiveness of MAR module. To demonstrate the overall effectiveness of MAR module, we append it to *Baseline* to fuse the intermediate features F_b and F_e, and the results are shown in the first and third rows in Table 4. Apparently, MAR can improve deblurring performance significantly. Accordingly, MAR can be considered as an effective and universal method of cross-modal fusion. Further, we validate the effectiveness of bi-directional supplement manner in MAR, i.e., fusing modality-shared to modality-specific features and fusing modality-specific to modality-shared features. Table 6 shows the results of different type of fusion, proving the validity of bi-directional supplement.

Finally, most importantly, when both MAD and MAR are embedded in our method, we achieve even higher deblurring performance than inserting only one module.

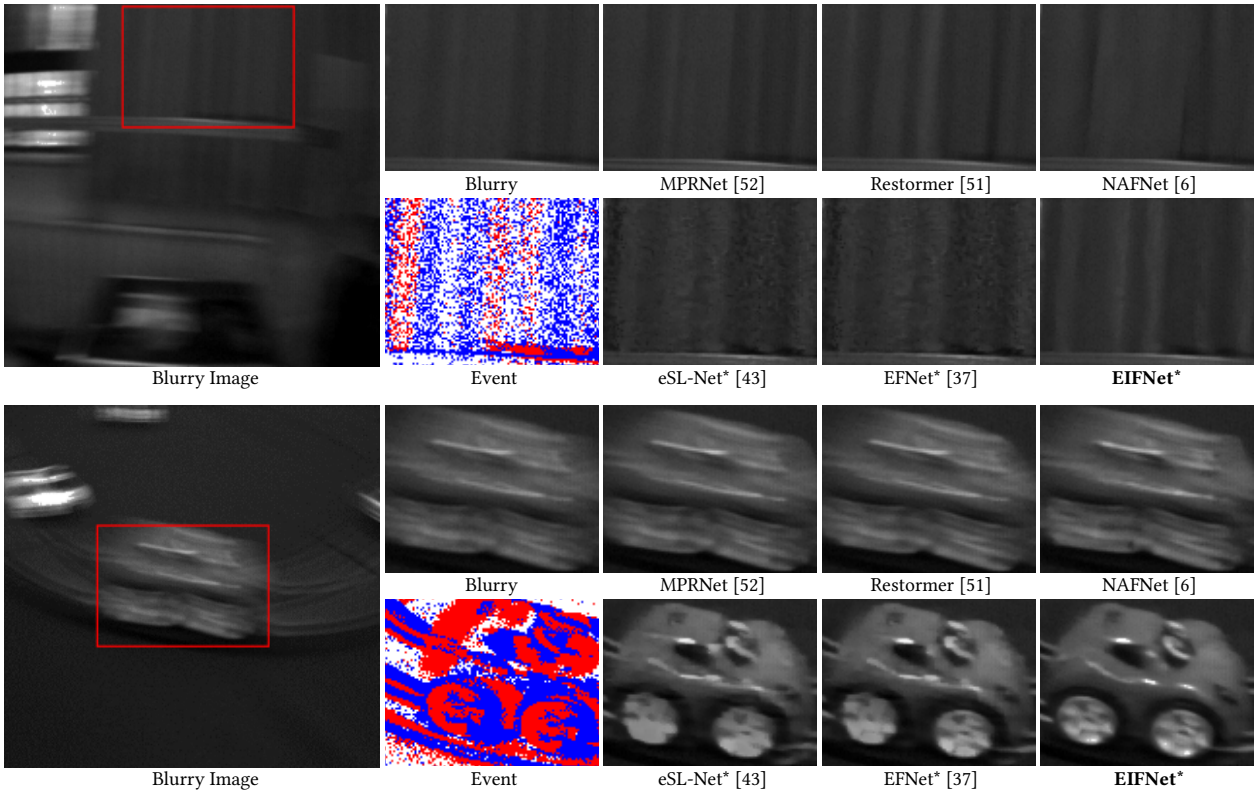


Figure 5: Visual comparison on the real blur set in *REB* dataset. * denotes event-based methods. Best viewed on a screen and zoomed in.

Table 4: Ablation study on individual components of EIFNet.

MAD	MAR	Gropo		REB	
		PSNR	SSIM	PSNR	SSIM
✗	✗	33.16	0.9571	33.04	0.9609
✓	✗	34.30	0.9686	34.10	0.9677
✗	✓	35.01	0.9734	34.42	0.9701
✓	✓	**35.99**	**0.9785**	**35.26**	**0.9737**

Table 5: Detailed ablation study of MAD.

Feature Type	Gropo		REB	
	PSNR	SSIM	PSNR	SSIM
F_b,F_e	33.16	0.9571	33.04	0.9609
F_c,F_d	33.72	0.9652	33.64	0.9650
F_{sh},F_{sp}	**34.30**	**0.9686**	**34.10**	**0.9677**

Table 6: Detailed ablation study of MAR.

Supplement Type	Gropo		REB	
	PSNR	SSIM	PSNR	SSIM
specific to shared	35.51	0.9761	34.94	0.9718
shared to specific	35.61	0.9765	34.92	0.9717
bi-directional	**35.99**	**0.9785**	**35.26**	**0.9737**

5 CONCLUSION

In this work, we propose a novel event-image fusion network (EIFNet) based on modality-aware decomposition and recomposition for motion deblurring, which explores the fusion of events and images with joint shared-modality and specific-modality attentions. We first design a modality-aware decomposition (MAD) module to divide the modality-shared and modality-specific features in parallel by exploring the global correlation of common-mode, differential-mode and two modalities with dual cross-attention. And then, the divided modality-shared and modality-specific features are merged with a modality-aware recomposition (MAR) module, in which information supplement is conducted in a bi-directional manner, i.e., shared-induced specific complement and specific-induced shared complement. Extensive evaluations show that our method achieves state-of-the-art performance on both synthetic and real-world datasets.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Yuval Bahat, Netalee Efrat, and Michal Irani. 2017. Non-uniform blind deblurring by reblurring. In *ICCV*.

[2] Leah Bar, Benjamin Berkels, Martin Rumpf, and Guillermo Sapiro. 2007. A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *ICCV*.

[3] Chengzhi Cao, Xueyang Fu, Yurui Zhu, Gege Shi, and Zheng-Jun Zha. 2022. Event-driven Video Deblurring via Spatio-Temporal Relation-Aware Network. In *IJCAI*.

[4] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*.

[5] Haoyu Chen, Minggui Teng, Boxin Shi, Yizhou Wang, and Tiejun Huang. 2022. A Residual Learning Approach to Deblur and Generate High Frame Rate Video With an Event Camera. *IEEE TMM* (2022).

[6] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. 2022. Simple baselines for image restoration. In *ECCV*.

[7] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. 2021. HINet: Half instance normalization network for image restoration. In *CVPR*.

[8] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. 2021. Rethinking Coarse-to-Fine Approach in Single Image Deblurring. In *ICCV*.

[9] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. 2022. Improving Image Restoration by Revisiting Global Information Aggregation. In *ECCV*.

[10] Shengyang Dai and Ying Wu. 2008. Motion from blur. In *CVPR*.

[11] Xin Deng and Pier Luigi Dragotti. 2020. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE TPAMI* (2020).

[12] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. 2006. Removing camera shake from a single photograph. In *ACM SIGGRAPH*.

[13] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. 2020. Event-based vision: A survey. *IEEE TPAMI* (2020).

[14] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. 2019. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*.

[15] Tae Hyun Kim and Kyoung Mu Lee. 2015. Generalized video deblurring for dynamic scenes. In *CVPR*.

[16] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. 2020. Learning event-based motion deblurring. In *CVPR*.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[18] Jan Kotera, Filip Šroubek, and Peyman Milanfar. 2013. Blind deconvolution using alternating maximum a posteriori estimation with heavy-tailed priors. In *CAIP*.

[19] Dilip Krishnan, Terence Tay, and Rob Fergus. 2011. Blind deconvolution using a normalized sparsity measure. In *CVPR*.

[20] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*.

[21] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*.

[22] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. 2009. Understanding and evaluating blind deconvolution algorithms. In *CVPR*.

[23] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. 2022. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288* (2022).

[24] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *ICCV*.

[25] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. 2022. Recurrent Video Restoration Transformer with Guided Deformable Attention. In *NeurIPS*.

[26] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. 2022. Flow-guided sparse transformer for video deblurring. In *ICML*.

[27] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. 2020. Learning event-driven video deblurring and interpolation. In *ECCV*.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

[29] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*.

[30] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. 2019. Bringing a blurry frame alive at high frame-rate with an event camera. In *CVPR*.

[31] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. 2020. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*.

[32] Kuldeep Purohit and AN Rajagopalan. 2020. Region-adaptive dense network for efficient motion deblurring. In *AAAI*.

[33] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. 2018. ESIM: an open event camera simulator. In *CoRL*.

[34] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. 2018. Continuous-time intensity estimation using event cameras. In *ACCV*.

[35] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. 2021. Bringing Events Into Video Deblurring With Non-Consecutively Blurry Frames. In *ICCV*.

[36] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. 2020. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*.

[37] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. 2022. Event-Based Fusion for Motion Deblurring with Cross-modal Attention. In *ECCV*.

[38] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. 2018. Scale-recurrent network for deep image deblurring. In *CVPR*.

[39] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin. 2022. BANet: A Blur-aware Attention Network for Dynamic Scene Deblurring. *IEEE TIP* (2022).

[40] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. Maxim: Multi-axis mlp for image processing. In *CVPR*.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[42] Shengdao Wan, Shu Tang, Xianzhong Xie, Jia Gu, Rong Huang, Bin Ma, and Lei Luo. 2020. Deep convolutional-neural-network-based channel attention for single image dynamic scene blind deblurring. *IEEE TCSVT* (2020).

[43] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. 2020. Event enhanced high-quality image recovery. In *ECCV*.

[44] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2022. Uformer: A general u-shaped transformer for image restoration. In *CVPR*.

[45] Jonas Wulff and Michael Julian Black. 2014. Modeling blurred video with layers. In *ECCV*.

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*.

[47] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. 2021. Motion Deblurring with Real Events. In *ICCV*.

[48] Qian Xu and Yuntao Qian. 2022. Bidirectional Transformer for Video Deblurring. *IEEE TCSVT* (2022).

[49] Dan Yang and Mehmet Yamac. 2022. Motion Aware Double Attention Network for Dynamic Scene Deblurring. In *CVPRW*.

[50] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, Weisheng Dong, and Guangming Shi. 2022. Learning for Motion Deblurring with Hybrid Frames and Events. In *ACM MM*.

[51] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*.

[52] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 2021. Multi-stage progressive image restoration. In *CVPR*.

[53] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. 2019. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*.

[54] Xiang Zhang and Lei Yu. 2022. Unifying Motion Deblurring and Frame Interpolation with Events. In *CVPR*.

[55] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*.

[56] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2020. Residual dense network for image restoration. *IEEE TPAMI* (2020).