Motion-adaptive Transformer for Event-based Image Deblurring

Senyan Xu, Zhijing Sun, Mingchen Zhong, Chengzhi Cao, Yidi Liu, Xueyang Fu*, Yan Chen

University of Science and Technology of China
{syxu, sunzhijing, zmcsa24010071, chengzhicao, liuyidi2023}@mail.ustc.edu.cn, {xyfu, eecyan}@ustc.edu.cn

Abstract

Event cameras, which capture pixel-level brightness changes asynchronously, provide rich motion information that is often missed during traditional frame-based camera exposures, thereby offering fresh perspectives for motion deblurring. Although current approaches incorporate event intensity, they neglect essential spatial motion information. Unlike their CNN architectures, Transformers excel in modeling long-range dependencies but struggle with establishing relevant non-local connections in sparse events and fail to highlight significant interactions in dense images. To address these limitations, we introduce a Motion-Adaptive Transformer network (MAT) that utilizes spatial motion information to forge robust global connections. The core design is an Adaptive Motion Mask Predictor (AMMP) that identifies key motion regions, guiding the Motion-Sparse Attention (MSA) to eliminate irrelevant event tokens and enabling the Motion-Aware Attention (MAA) to focus on relevant ones, thereby enhancing long-range dependency modeling. Additionally, we elaborately design a Cross-Modal Intensity Gating mechanism that efficiently merges intensity data across modalities while minimizing parameter use. The learnable Expansion-Controlled Spatial Gating further optimizes the transmission of event features. Comprehensive testing confirms that our approach sets a new benchmark in image deblurring, surpassing previous methods by up to 0.60dB on the GoPro dataset, 1.04dB on the HS-ERGB dataset, and achieving an average improvement of 0.52dB across two real-world datasets.

Code — https://github.com/QUEAHREN/MAT

Introduction

Motion blur commonly arises from the relative motion between the camera and the scene during the exposure time of frame-based cameras. Image deblurring, a complex inverse problem, seeks to restore a sharp image from its blurred counterpart. Recent strides in deep learning have markedly enhanced traditional image deblurring techniques, as evidenced by various studies (Chen et al. 2021; Cho et al. 2021; Chen et al. 2022; Zamir et al. 2022; Li et al. 2022; Kong et al. 2023; Peng et al. 2024; Liu et al. 2024). Unlike conventional cameras, which fail to capture motion details during exposure, event cameras record per-pixel intensity changes asynchronously at microsecond precision. This high temporal resolution and rich motion data have proven invaluable in augmenting traditional deblurring methods, as demonstrated in several research efforts (Maqueda et al. 2018; Wang et al. 2019; Pan et al. 2019; Cao et al. 2022, 2023).

Recent studies on event-based deblurring, such as those by (Sun et al. 2022; Yang et al. 2023; Sun et al. 2025), treat events as a distinct modality and use dual-branch encoders to extract features from both events and images, followed by various cross-modal fusion techniques. While these methods mark significant progress, they still face challenges that limit their effectiveness in diverse and complex real-world situations. Typically, these approaches focus predominantly on the intensity information from events, neglecting the vital spatial information that pinpoints motion areas. This spatial motion information, which directly correlates with blur regions, has been exploited in certain conventional deblurring methods to improve deblurring quality (Zhang, Xie, and Yao 2024; Kim et al. 2024). Furthermore, the reliance on CNN architectures restricts their ability to utilize global features.

Unlike the convolution operation, Transformers are adept at capturing non-local information (Vaswani et al. 2017) and have been extensively utilized in various image restoration tasks (Xiao et al. 2022a, 2024; Chen et al. 2023; Zhou et al. 2024), including deblurring (Zamir et al. 2022; Wang et al. 2022; Kong et al. 2023). While standard Transformer architectures excel at modeling global relationships between queries and keys, aiding in feature aggregation in image deblurring, the high-density visual information presents a challenge where the attention mechanism struggles to concentrate on the most relevant tokens without spatial guidance. Additionally, when dealing with events, we note that due to their inherent sparsity, many tokens from the key are irrelevant to those from the query, especially in scenarios with limited relative movement. In these cases, the standard attention mechanism compromises the quality of latent representations during feature aggregation due to its inability to discern relevant interactions. These limitations highlight the necessity for more tailored attention strategies that can manage the unique aspects of sparse events and dense images.

In this paper, we introduce a new Motion-Adaptive Transformer (MAT) tailored for event-based image deblurring. At the core of our architecture lies the Adaptive Motion Mask

*Corresponding author.

Predictor (AMMP), which extracts masks that capture relative movement from event features, thereby guiding spatial attention mechanisms. Specifically, during the encoding stage for events, we implement an Expansion-Controlled Spatial Gating to regulate information flow and enhance gradient propagation through AMMP during backpropagation, making AMMP trainable. Subsequently, Motion-Sparse Attention leverages the mask from AMMP to filter out irrelevant tokens. For images, we employ a straightforward yet efficacious Cross-Modal Intensity Gating to amalgamate intensity data. Motion-Aware Attention then uses the mask to bolster interactions among tokens linked to motion dynamics.

The contributions of this work are summarized as follows:

- We present MAT, a motion-adaptive Transformer framework for event-based image deblurring that utilizes both intensity and spatial information from events to establish robust non-local dependencies.
- We develop an adaptive motion mask predictor that discerns motion-related information, providing essential spatial guidance for both motion-sparse and motion-aware attention mechanisms tailored to events and images.
- We introduce simple yet effective cross-modal intensity gating and expansion-controlled spatial gating, which together manage the flow of event feature transmission and render AMMP trainable.

Our experimental results demonstrate that our method achieves state-of-the-art performance across various benchmarks. It surpasses previous studies (Sun et al. 2025) with improvements of up to 0.60dB on the GoPro dataset and 1.04dB on the HS-ERGB dataset, achieving an average enhancement of 0.52dB across two real-world datasets.

Related Work

Event-based Image Deblurring

Event cameras can asynchronously record per-pixel intensity changes with low latency, providing crucial motion information in some image restoration tasks (Ge, Fu, and Zha 2022; Ge et al. 2024; Xu et al. 2024), particularly in deblurring (Wang et al. 2020; Xu et al. 2021; Cao et al. 2022, 2023; Zhang and Yu 2022; Kim et al. 2022; Kim, Cho, and Yoon 2025). Early methods such as (Pan et al. 2019; Zhang and Yu 2022) utilize the Double Integral model to establish a correlation between a sharp and a blurry image. In recent research (Sun et al. 2022; Yang et al. 2023; Sun et al. 2025), events have typically been considered a new modal to explore effective fusion and alignment methods. EFNet (Sun et al. 2022) proposes a symmetric cumulative event representation (SCER) and first introduces cross attention to better fuse intensity information of events and an image. EIFNet (Yang et al. 2023) employs modality-aware decomposition and recomposition to utilize shared and specific features among each modal. MAENet (Sun et al. 2025) proposes a deviation accumulation representation (DA) and aligns event and image features for improved fusion. However, these approaches overlook the crucial spatial information that indicates motion regions, and the foundational design of these approaches, which is based on CNN architectures, inherently limits their ability to aggregate global information.

Vision Transformers

Transformers (Vaswani et al. 2017; Dosovitskiy et al. 2020; Liu et al. 2021) have been widely applied to image restoration (Liang et al. 2021; Xiao et al. 2022b,a, 2024; Chen et al. 2023) and achieve better performance than CNN-based methods due to its ability to model long-range dependencies. For the field of image deblurring, Uformer (Wang et al. 2022) calculates self-attention in a locally window-based way to reduce complexity. Restormer (Zamir et al. 2022) proposes an efficient way of computing the transposed attention to obtain global information aggregation. FFT-Former (Kong et al. 2023) develops a frequency domain-based self-attention to estimate the scaled dot-product attention by an element-wise product operation. However, these methods compromise the quality of latent representations during feature aggregation due to their inability to discriminate relevant interactions among tokens, thereby limiting their performance.

Methodology

Overall Framework

Our proposed Motion-Adaptive Transformer (MAT) is designed to handle event-based deblurring by dynamically adapting to the motion information of event streams. As shown in Figure 1 (a), MAT consists of Adaptive Motion Mask Predictor (AMMP), Motion-Sparse Event Block (MSEB), and Motion-Aware Image Block (MAIB). Specifically, MSEB combines Expansion-Controlled Spatial Gating (ECSG) with Motion-Sparse Attention (MSA), and MAIB integrates Cross-Modal Intensity Gating (CMIG) with Motion-Aware Attention (MAA).

The encoding stage has two parallel branches with identical hierarchical designs. Each layer in this stage is equipped with an AMMP, which guides MSA and MAA at the same layer by predicting motion masks $\mathbf{M}_i$. We initialize the mask $\mathbf{M}_0$ to an all-one matrix and progressively update it, adapting dynamically to capture motion-relevant regions. In detail, at the same layer of the encoding stage, there are N_e MSEBs and N_f MAIBs, which respectively take event feature $\mathbf{F}_i^{\text{event}}$ and image feature $\mathbf{F}_j^{\text{image}}$ as inputs, where $i \in [0, N_e-1], j \in [0, N_f-1]$. The feature propagation can be formulated as:

$$\begin{aligned} \mathbf{M}_{i+1}, \mathbf{S}_{i+1} &= \text{AMMP}(\mathbf{F}_i^{\text{event}}, \mathbf{M}_i), \\ \mathbf{F}_{i+1}^{\text{event}} &= \text{MSEB}_{i+1}(\mathbf{F}_i^{\text{event}}, \mathbf{M}_{i+1}, \mathbf{S}_{i+1}), \\ \mathbf{F}_{j+1}^{\text{image}} &= \text{MAIB}_{j+1}(\mathbf{F}_j^{\text{image}}, \mathbf{F}_{N_e}^{\text{event}}, \mathbf{M}_{N_e}), \end{aligned} \tag{1}$$

where the $\mathbf{S}_i$ is the spatial score map estimated by AMMP. See Supplementary Materials for a detailed visualization of mask evolution. In the decoding stage, each layer consists of Transformer blocks proposed by (Zamir et al. 2022) and an upsampling layer. We detail each module in the following section.

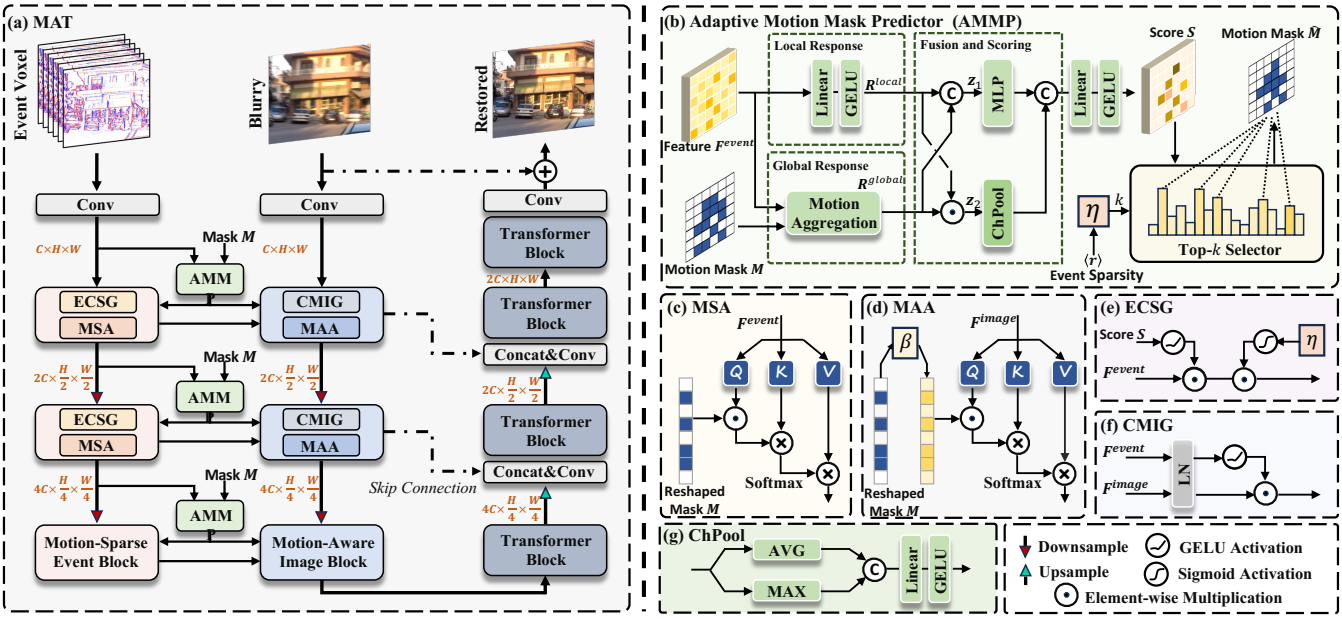

Figure 1: The overall architecture of the proposed motion-adaptive Transformer (MAT) for event-based deblurring, which mainly contains adaptive motion mask predictor (AMMP), expansion-controlled spatial gating (ECSG), motion-sparse attention (MSA), cross-modal intensity gating (CMIG), and motion-aware attention (MAA). LN refers to the layer normalization. Note that the layer normalization and feed-forward layer have been omitted for simplicity.

Adaptive Motion Mask Predictor

As illustrated in Figure 1 (b), Adaptive Motion Mask Predictor (AMMP) is designed to obtain a mask that indicates motion regions from event features and a spatial score map.

Initially, we maintain a binary motion mask $\mathbf{M} \in \{0,1\}^N$ to indicate the motion attribute of each token, where $N = H \times W$ is the number of tokens. AMMP takes the current mask $\mathbf{M}$ and the reshaped event feature $\mathbf{F}_e \in \mathbb{R}^{N \times C}$ as input. Firstly, $\mathbf{F}_e$ is sent to the feature aggregation unit. Local response $\mathbf{R}^{local}$ is computed using a linear layer followed by a GELU layer. And global motion response $\mathbf{R}^{global}$ is computed by a Motion Aggregation function. The aggregation process can be formulated as:

$$\begin{aligned} \mathbf{R}^{local} &= \phi(W_L \cdot \mathbf{F}_e), \\ \mathbf{R}^{global} &= \text{MotionAgg}(\mathbf{F}_e, \mathbf{M}) = \frac{\Sigma_{i=1}^{N} \mathbf{M}_i \mathbf{F}_{e,i}}{\Sigma_{i=1}^{N} \mathbf{M}_i}, \end{aligned} \tag{2}$$

where $\phi(\cdot)$ is a GELU layer (Hendrycks and Gimpel 2016), W_L represents the weight of the linear layer and MotionAgg$(\cdot,\cdot)$ is implemented as an average pooling according to $\mathbf{M}$, aggregating the global motion information. Then, the local and global responses are fused mutually, capturing both detailed and broad aspects of the scene dynamics:

$$\begin{aligned} \mathbf{z}_1 &= \text{Cat}([\mathbf{R}^{local}, \mathcal{E}(\mathbf{R}^{global})]), \\ \mathbf{z}_2 &= \mathbf{R}^{local} \odot \mathcal{E}(\mathbf{R}^{global}), \end{aligned} \tag{3}$$

where Cat$(\cdot)$ is the concatenation operation, $\mathcal{E}(\cdot)$ represents the dimension expansion operation and $\odot$ denotes element-wise multiplication. To derive a spatial motion score $\mathbf{S}$, we apply a series of linear, non-linear, and pooling operations on these combined features:

$$\mathbf{S} = \phi(W_S \cdot \text{Cat}([\text{MLP}(\mathbf{z}_1), \phi(W_z \cdot \text{ChPool}(\mathbf{z}_2))])), \tag{4}$$

where W_S and W_z represent the weight of the linear layer, andChPool$(\cdot)$ denotes a pooling operation that concatenates the results of max pooling and average pooling along the channel dimension. This pooling operation emphasizes the salient features while maintaining the overall average level of the information. We then generate a new motion mask $\hat{\mathbf{M}}$ by scattering the indices sampled based on the top k importance scores from $\mathbf{S}$:

$$\hat{\mathbf{M}} = \mathcal{S}(\mathbf{0}, \mathcal{T}_k(\mathbf{S}), 1), \tag{5}$$

where $\mathcal{T}_k(\cdot)$ is the Top-k indexing operator and $\mathcal{S}(\cdot)$ denotes the scattering operation of placing ones at the positions specified by the indices $\mathcal{T}_k(\mathbf{S})$ in an all-zero matrix $\mathbf{0}$. Naturally, the selection of k is crucial to the whole framework. We find that the selection of the proportion of tokens is closely related to the sparsity rate of the original event streams. We can calculate non-zero ratios $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_B)$ of B voxel bins of events to obtain sparsity rate $\mathbf{r}$. Given that numerous convolution operations within the network aggregate local features, the characteristics at each pixel position tend to expand toward the surrounding regions. To manage this, k can be calculated by:

$$k = \frac{\langle \mathbf{r} \rangle}{\eta + \epsilon} \cdot N, \tag{6}$$

where $\eta \in [0,1]$ is a learnable expansion factor, ϵ is set to 1×10^{-2} to ensure stable training, $N = H \times W$ is the number of tokens and $\langle \mathbf{r} \rangle$ represents the average sparsity rate.

Overall, AMMP obtains the spatial motion score $\mathbf{S}$ by aggregating global motion and local responses. Then, introducing the Expansion Factor η allows AMMP to adaptively adjust the value of k based on the extent of feature aggregation and expansion. However, since $\mathcal{T}_k(\cdot)$ is a non-differentiable operation, we designed ECSG, which takes $\mathbf{S}$ and η as inputs, to make AMMP learnable.

Motion-Sparse and Motion-Aware Attention

We first revisit the self-attention mechanism in standard Transformers, which has become a fundamental operation in most Transformer-based models. Given a query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) projected from the same feature, the dot-product attention is generally formulated as:

$$\begin{aligned} \text{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \mathcal{A}(\mathbf{Q},\mathbf{K})\cdot\mathbf{V}, \\ &= \text{Softmax}(\mathbf{Q}\cdot\mathbf{K}^\top/\lambda)\cdot\mathbf{V}, \end{aligned} \tag{7}$$

where $\mathbf{Q},\mathbf{K},\mathbf{V}\in\mathbb{R}^{N\times d}$ and $\lambda=\sqrt{d}$ is a scaling factor. The conventional self-attention paradigm relies on a densely fully connected attention map, which involves computation for all query-key pairs (Chen et al. 2023; Zhou et al. 2024). This inability to discriminate between relevant and irrelevant interactions results in sub-optimal feature aggregation.

To address this, AMMP extracts spatial information from events, providing a motion mask $\mathbf{M}$. This mask enables the attention mechanism to focus on modeling correlations among relevant tokens, mitigating the interference caused by irrelevant tokens. Considering the specific characteristics of events and images, as shown in Figure 1 (c) and Figure 1 (d), we devise two distinct attention computation methods.

Motion-sparse attention. Specifically, given a feature tensor $\mathbf{X}\in\mathbb{R}^{C\times H\times W}$, we perform layer normalization $\text{LN}(\cdot)$ on it and generates $\mathbf{Q},\mathbf{K}$ and $\mathbf{V}$ by applying local projection operation W_p with 1×1 convolutions and 3×3 depth-wise convolutions before performing reshape operation $\mathcal{R}(\cdot)$, $\mathbf{Q}=\mathcal{R}(W_p^Q\cdot\text{LN}(\mathbf{X})),\mathbf{K}=\mathcal{R}(W_p^K\cdot\text{LN}(\mathbf{X}))$ and $\mathbf{V}=\mathcal{R}(W_p^V\cdot\text{LN}(\mathbf{X}))$. However, the time and memory complexity of the standard key-query dot-product attention increases quadratically with the spatial resolution of the input. While some window-based methods(Wang et al. 2022; Kong et al. 2023) can reduce this complexity, they also restrict the scope of the global motion mask. To overcome this, we mitigate the complexity by computing transposed attention ($\mathbf{K}^\top\cdot\mathbf{Q}\in\mathbb{R}^{C\times C}$), leveraging the strong connection between the Gram matrix and the covariance matrix as inspired by (Zamir et al. 2022; Ali et al. 2021). Then, motion-sparse attention (MSA) can be formulated:

$$\mathcal{A}_{ms}(\mathbf{K},\mathbf{Q})=\text{Softmax}(\mathbf{K}^\top\cdot(\mathbf{Q}\odot\mathbf{M})/\alpha), \tag{8}$$

where $\mathbf{M}$ represents the motion mask and α is a learnable scaling factor. MSA restricts the scope of attention interaction by simply masking irrelevant tokens.

Motion-aware attention. Similarly to MSA, we use the motion mask to enable motion-aware attention (MAA), which dynamically adjusts the focus of the attention mechanism based on the relative movement within the scene. Specifically, MAA can be formulated as:

$$\mathcal{A}_{ma}(\mathbf{K},\mathbf{Q})=\text{Softmax}(\mathbf{K}^\top\cdot(\mathbf{Q}\odot(\beta\mathbf{M}+1))/\alpha), \tag{9}$$

where β is a learnable motion modulation factor that scales the influence of the mask $\mathbf{M}$ on the attention weights. This strategy enables the attention to selectively enhance or suppress based on the presence of motion, thereby making the model more sensitive to dynamic changes in the image.

Expansion-Controlled Spatial Gating

The score $\mathbf{S}$ aggregated by AMMP integrates global motion and local response information. As shown in Figure 1 (e), we introduce a gating mechanism, using GELU for non-linear activation of the score $\mathbf{S}$, to control the information flow of event feature transmission. The expansion factor η spatially indicates the dilation extent of local convolution operations in the network. A sigmoid layer transforms the expansion factor into a mixed information transmission rate, which controls the transmission rate of the gating mechanism. Given a event tensor $\mathbf{F}_e\in\mathbb{R}^{C\times H\times W}$, ECSG is formulated as:

$$\mathbf{F}_e'=\sigma(\eta)\cdot\phi(\mathbf{S})\odot\mathbf{F}_e+\mathbf{F}_e, \tag{10}$$

where $\sigma(\cdot)$ represents sigmoid layer. ECSG balances and controls the flow of information through MSTB, leveraging global and local motion cues. Additionally, ECSG facilitates gradient flow through AMMP during backpropagation, bypassing the non-differentiable top-k indexing operation, thereby making the AMMP learnable.

Cross-Modal Intensity Gating

Fusing intensity information from event streams and frames is crucial for event-based deblurring. Previous approaches have attempted to establish long-range dependencies between modalities from a cross-modal fusion perspective. However, this method is inefficient in our network due to the inherent differences in temporal and spatial characteristics between events and images, which can lead to misalignment of information integration.

Similar to ECSG, as shown in Figure 1 (f), we utilize GELU to modulate event features. It controls the transmission of intensity information within image features based on the intensity information present in the event features. Specifically, given an input image tensor $\mathbf{F}_i$ and event feature tensor $\mathbf{F}_e$, CMIG is formulated as follows:

$$\mathbf{F}_i'=\phi(\text{LN}(\mathbf{F}_e))\odot\text{LN}(\mathbf{F}_i)+\mathbf{F}_i. \tag{11}$$

CMIG ensures the intensity information from event features dynamically influences the image features, enhancing the deblurring process by effectively leveraging the complementary strengths of both data modalities.

Experiments and Analysis

Datasets

(1) **GoPro** (Nah, Hyun Kim, and Mu Lee 2017) is a widely recognized benchmark for motion deblurring. The blurry images are generated by averaging adjacent sharp frames. Additionally, we employ the raw event dataset from (Sun et al. 2022), where events are synthesized using the ESIM simulator (Rebecq, Gehrig, and Scaramuzza 2018). (2) **HS-ERGB** (Tulyakov et al. 2021) consists of sharp videos and

Methods		Venue	Input	GoPro		HS-ERGB		REBlur		#Pra.	FLOPs
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)	(G)
Prior-based	EDI	CVPR'19	F+E	29.06	0.943	23.93	0.704	36.52	0.964	0.5	-
	EVDI	CVPR'22	F+E	30.40	0.906	25.13	0.707	-	-	0.4	-
CNN-based	HINet	CVPR'21	F	32.71	0.959	27.32	0.807	35.58	0.965	88.7	241.5
	MSDI-Net	ECCV'22	F	33.28	0.964	27.46	0.809	36.14	0.968	135.4	522.5
	NAFNet	ECCV'22	F	33.71	0.967	27.64	0.811	36.15	0.969	67.8	96.8
	UFPNet	CVPR'23	F	34.06	0.968	27.64	0.809	36.11	0.968	80.3	361.2
	UFPNet$*$	CVPR'23	F+E	35.22	0.972	27.68	0.809	37.97	0.976	80.3	361.2
	EFNet	ECCV'22	F+E	35.46	0.972	26.68	0.800	38.12	0.975	8.5	155.6
	EIFNet	MM'23	F+E	35.99	**0.979**	26.74	0.797	37.16	0.972	10.8	145.1
	MAENet	ECCV'24	F+E	<u>36.07</u>	0.976	27.93	0.812	<u>38.47</u>	<u>0.978</u>	13.9	148.3
Transformer -based	Restormer	CVPR'22	F	32.92	0.961	27.55	0.808	35.50	0.959	26.1	216.1
	FFTFormer	CVPR'23	F	34.21	0.969	<u>28.11</u>	<u>0.813</u>	36.25	0.968	16.6	201.5
	Restormer$*$	CVPR'22	F+E	35.96	0.975	27.19	0.800	38.31	0.975	26.1	216.1
	FFTFormer$*$	CVPR'23	F+E	33.08	0.958	27.61	0.808	38.39	0.975	16.6	201.5
	Ours	-	F+E	**36.67**	<u>0.978</u>	**28.97**	**0.816**	**38.69**	**0.978**	11.7	186.9

Table 1: The quantitative results on GoPro, HS-ERGB, and REBlur test datasets. UFPNet$*$, Restormer$*$, and FFTFormer$*$ are event-enhanced versions of UFPNet, Restormer, and FFTFormer by concatenating events to their input. FLOPs are estimated with the resolution of 224×224. The best and the second results are boldfaced and underlined, respectively.

real-world events; we use (Zhang et al. 2023) released normal blur version, which synthesizes blurry images by averaging 49 interpolating images. (3) **REBlur** (Sun et al. 2022) collects sequences of real-world events corresponding with real blurry images and sharp images. (4) **REVD** (Kim, Cho, and Yoon 2024) provides 21 sequences of real-world blur-sharp image pairs and event streams. The resolution of the image and event is 1024×768.

Implementation Details

We implement our proposed network via the PyTorch 1.8 platform on NVIDIA RTX 3090 GPU. AdamW (Loshchilov and Hutter 2017) optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is adopted to optimize our network. The learnable expansion factor η and motion modulation factor β are initialized to 0.5 and 1.0, respectively. We train our network with patch size 128×128 and batch size 20 using the charbonnier loss (Charbonnier et al. 1994). The initial learning rate is 5×10^{-4} and changes with Cosine Annealing scheme to 1×10^{-6}, including 320K iterations in total. Note that we set the bin size of event voxels as 6.

Comparison with State-of-the-Art Methods

We conduct experiments on synthesized and real-world datasets. Table 1 and Table 2 present the quantitative results of our proposed method alongside other state-of-the-art (SOTA) techniques. We compare our network against recent event-based methods, traditional image-only deblurring networks, and their enhanced versions with event input.

Synthesized Datasets: Table 1 shows our performance on the GoPro and the HS-ERGB datasets. We can observe that most event-based deblurring methods outperform image-only methods, including event-enhanced versions of UFPNet (Fang et al. 2023), Restormer (Zamir et al. 2022). Additionally, we can see that even with a simple concatenation of events to the input, Restormer$*$ can achieve performance comparable to carefully designed event-based CNN architectures, which indicates that the local receptive field limits the performance of event-based CNN methods. Compared to the SOTA event-based method MAENet (Sun et al. 2025), we achieve improvements of 0.60dB and 1.04dB in terms of PSNR on the GoPro and the HS-ERGB datasets, respectively, while using fewer parameters. Compared to other Transformer-based methods, our method achieves an average improvement of 0.71dB and 0.86dB on the GoPro and the HS-ERGB datasets with fewer FLOPs, respectively. Furthermore, we compare the qualitative visual quality in Fig-

Methods	Venue	Input	REVD		FLOPs
			PSNR	SSIM	(T)
eSL-Net	ECCV'20	V+E	26.99	0.787	0.25
REDNet	ICCV'21		31.90	0.921	19.07
UEVD	ECCV'22		31.97	0.921	39.03
FEVD	CVPR'24		<u>32.99</u>	**0.933**	30.27
Restormer$*$	CVPR'22	F+E	32.11	0.921	16.92
FFTFormer$*$	CVPR'23		30.85	0.901	15.78
EFNet	ECCV'22		31.00	0.907	12.95
EIFNet	MM'23		31.28	0.911	11.39
MAENet	ECCV'24		32.23	0.923	11.57
Ours	-		**33.05**	<u>0.932</u>	14.63

Table 2: The quantitative results on REVD test datasets. Note that the FLOPs for all methods are estimated with an input of 5 frames and a resolution of 1024×768.

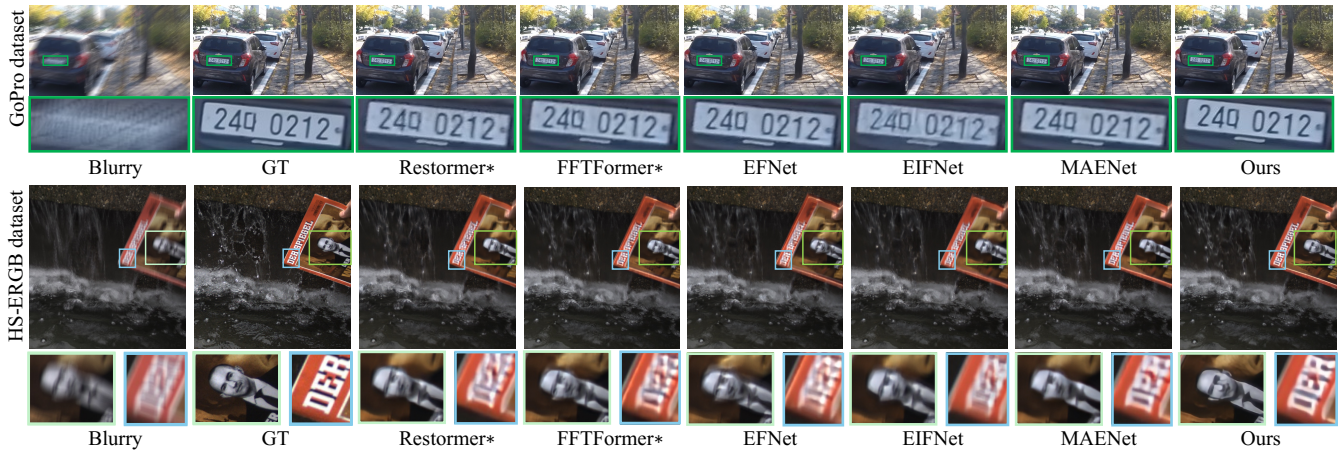

Figure 2: Qualitative comparisons on the GoPro dataset and the HS-ERGB dataset. The notation is the same as in Table 1.

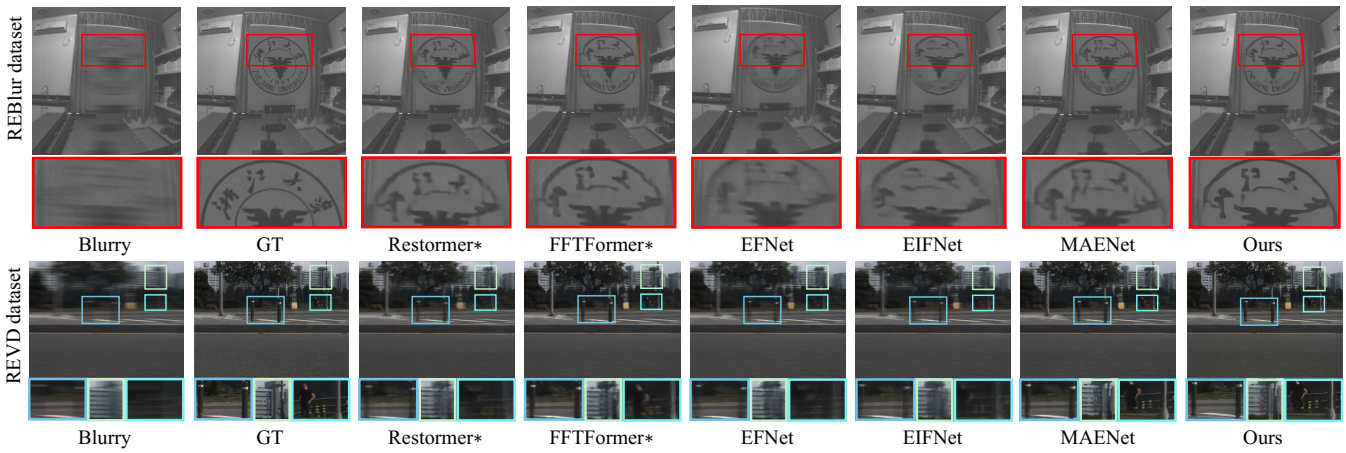

Figure 3: Qualitative comparisons on the REBlur dataset and the REVD dataset. The notation is the same as in Table 1.

ure 2, demonstrating that our proposed approach achieves superior recovery of sharp texture details.

Real-world Datasets: We compare our method on the REBlur and the REVD datasets with real-events and real-blur pairs. Among them, the REVD dataset features higher resolution and a greater diversity of scenes, whereas the REBlur dataset includes denser linear and non-linear motion. Table 1 shows our quantitative results on REBlur, where our method significantly outperforms all other SOTA competitors, improving 0.22dB over MAENet. Table 2 presents the quantitative results of event-based methods on the REVD dataset. Here, 'V+E' indicates video deblurring methods that process five consecutive frames and 16 bins of events, while 'F+E' refers to image deblurring methods that process one image and 6 bins of events. We adopted some methods' metrics reported in (Kim, Cho, and Yoon 2024). Note that more temporal information can be utilized by inputting more bins of events voxels. Compared to image deblurring methods, we can observe that we significantly surpass the previous state-of-the-art MAENet by 0.78dB in terms of PSNR and 0.009 in terms of SSIM. Moreover, compared to video methods that utilize much additional temporal information, we still exceed the latest SOTA FEVD (Kim, Cho, and Yoon 2024) performance by 0.06dB with fewer than 15.64 TFLOPs. Moreover, we compare the qualitative visual quality in Figure 3. Our method achieves the most visually plausible deblurring results with sharper textures in highly challenging real-world scenarios. Our results are more realistic and cleaner than other methods, which produce more artifacts and struggle to effectively remove severe blur.

Ablation Studies

We conduct two ablation studies on the GoPro dataset to analyze the contribution of different components of our network (Table 3) and varied event representations (Table 4). We change the experiment setting to a batch size of 4 and 180K iterations. Note that the baseline method uses full tokens to compute standard attention without mask predictors and employs simple addition for feature fusion.

Effects of AMMP. Our AMMP predicts the motion mask to guide our designed attention. For comparison, the baseline method (labeled as 'Top-k' in tables and figures) generates

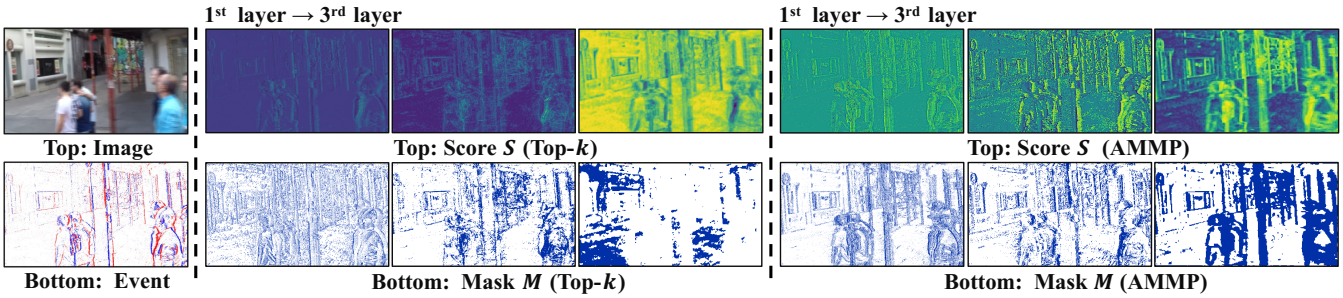

Figure 4: Visualizations of score heatmaps and motion masks. The middle column uses simple average pooling and Top-k to generate the mask, whereas the right column shows AMMP's predictions. For each method, we visualize the score maps and motion masks from the first, second, and third layers, from left to right, as the network progresses.

Mask Predictor	Attention MSA	Attention MAA	Fusion Module	PSNR	#Pra. (M)
n/a	-	-	Add	34.19	11.35
	-	-	CMIG	34.47	11.60
TopK	✓	-	CMIG	34.74	11.60
	-	✓		34.55	
	✓	✓		34.82	
AMMP	✓	✓	Concat.	34.31	12.13
	✓	✓	CA	34.84	15.66
AMMP	✓	-	CMIG	34.85	11.78
	-	✓		34.64	
	✓	✓		35.23	

Table 3: Ablation study of each module effects. '-' denotes utilizing standard transposed attention without mask guidance (i.e., using all tokens for attention computation).

the score by performing average pooling on event features along the channel dimensions, then applies the Top-k operation to predict the mask. Compared to simply applying Top-k, AMMP improves PSNR by 0.41dB with only more than 0.18M parameters when using both MSA and MAA. Besides, when using either MSA or MAA individually, AMMP also improves performance compared to baseline. Note that to ensure AMMP is learnable, we default to using ECSG when employing AMMP. As shown in Figure 4, we visualize the score map estimated by AMMP and the baseline. It is clear that AMMP generates more accurate score maps that indicate motion cues, which means that AMMP can adaptively estimate the spatial information in events, providing more precise guidance for subsequent attention modules.

Effects of MSA and MAA. The mask predictor provides guidance to MSA and MAA. Therefore, the baseline for comparison is the standard attention without mask guidance. When combined as attention strategies, MSA and MAA with AMMP or TopK as the mask predictor enhance performance by 0.75dB and 0.35dB in PSNR over the baseline, respectively. Moreover, even when utilizing MSA or MAA individually, there is still a noticeable improvement over the baseline. MSA/MAA doesn't introduce any additional parameters.

Method	SBT PSNR	SBT SSIM	SCER PSNR	SCER SSIM	DA PSNR	DA SSIM
EFNet	35.12	0.970	35.46	0.972	35.08	0.970
EIFNet	<u>35.99</u>	**0.979**	33.35	0.960	33.21	0.960
MAENet	35.63	0.973	<u>35.81</u>	<u>0.975</u>	<u>36.07</u>	<u>0.976</u>
Ours	**36.62**	<u>0.977</u>	**36.67**	**0.978**	**36.46**	**0.977**

Table 4: Comparison across various event representations.

Effects of CMIG. Our CMIG fusion block effectively fuses event and image intensity, improving PSNR by 0.39dB or more compared to simple fusion modules such as addition (Add), concatenating (Concat.), and cross attention (CA). Additionally, CMIG saves 0.35M parameters or more compared to other fusion modules.

Robustness across Different Event Representations As shown in Table 4, we conduct experiments on three event-based image deblurring methods using several common event voxel representations as input. Our proposed approach presents more robust across different event representations, achieving improvements of 0.63dB, 0.86dB, and 0.39dB in terms of PSNR on SBT (Wang et al. 2019), SCER (Sun et al. 2022), and DA (Sun et al. 2025), respectively.

Conclusion

We propose a new motion-adaptive Transformer (MAT) for event-based deblurring by utilizing both intensity and spatial information from events to establish robust non-local dependencies. MAT incorporates an Adaptive Motion Mask Predictor and Motion-Sparse and Motion-Aware Attention mechanisms to model long-range dependencies focusing on motion-relevant tokens. Additionally, it features cross-modal intensity gating and expansion-controlled spatial gating to integrate and regulate information flow efficiently. Our extensive testing shows that MAT sets a new state-of-the-art in event-based image deblurring, significantly outperforming existing methods across various datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62422609 and 62276243.

References

Ali, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; et al. 2021. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34: 20014–20027.

Cao, C.; Fu, X.; Zhu, Y.; Shi, G.; and Zha, Z.-J. 2022. Event-driven Video Deblurring via Spatio-Temporal Relation-Aware Network. In *IJCAI*, 799–805.

Cao, C.; Fu, X.; Zhu, Y.; Sun, Z.; and Zha, Z.-J. 2023. Event-Driven Video Restoration With Spiking-Convolutional Architecture. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, volume 2, 168–172. IEEE.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *Proceedings of the European conference on computer vision (ECCV)*, 17–33. Springer.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 182–192.

Chen, X.; Li, H.; Li, M.; and Pan, J. 2023. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5896–5905.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 4641–4650.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, Z.; Wu, F.; Dong, W.; Li, X.; Wu, J.; and Shi, G. 2023. Self-supervised Non-uniform Kernel Estimation with Flow-based Motion Prior for Blind Image Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18105–18114.

Ge, C.; Fu, X.; He, P.; Wang, K.; Cao, C.; and Zha, Z.-J. 2024. Neuromorphic Event Signal-Driven Network for Video De-raining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1878–1886.

Ge, C.; Fu, X.; and Zha, Z.-J. 2022. Learning Dual Convolutional Dictionaries for Image De-raining. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6636–6644. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Kim, I.; Choi, J. S.; Seo, G.; Kwon, K.; Shin, J.; and Lee, H.-E. 2024. Real-World Efficient Blind Motion Deblurring via Blur Pixel Discretization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25879–25888.

Kim, T.; Cho, H.; and Yoon, K.-J. 2024. Frequency-aware Event-based Video Deblurring for Real-World Motion Blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24966–24976.

Kim, T.; Cho, H.; and Yoon, K.-J. 2025. CMTA: Cross-Modal Temporal Alignment for Event-guided Video Deblurring. In *Proceedings of the European conference on computer vision (ECCV)*, 1–19. Springer.

Kim, T.; Lee, J.; Wang, L.; and Yoon, K.-J. 2022. Event-guided deblurring of unknown exposure time videos. In *Proceedings of the European conference on computer vision (ECCV)*, 519–538. Springer.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5886–5895.

Li, D.; Zhang, Y.; Cheung, K. C.; Wang, X.; Qin, H.; and Li, H. 2022. Learning degradation representations for image deblurring. In *Proceedings of the European conference on computer vision (ECCV)*, 736–753. Springer.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, C.; Wang, X.; Xu, X.; Tian, R.; Li, S.; Qian, X.; and Yang, M.-H. 2024. Motion-adaptive Separable Collaborative Filters for Blind Motion Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25595–25605.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maqueda, A. I.; Loquercio, A.; Gallego, G.; García, N.; and Scaramuzza, D. 2018. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5419–5427.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

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 *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 6820–6829.

Peng, L.; Cao, Y.; Sun, Y.; and Wang, Y. 2024. Lightweight Adaptive Feature De-drifting for Compressed Image Classification. *IEEE Transactions on Multimedia*.

Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. ESIM: an open event camera simulator. In *Conference on robot learning*, 969–982. PMLR.

Sun, L.; Sakaridis, C.; Liang, J.; Jiang, Q.; Yang, K.; Sun, P.; Ye, Y.; Wang, K.; and Gool, L. V. 2022. Event-based fusion for motion deblurring with cross-modal attention. In *Proceedings of the European conference on computer vision (ECCV)*, 412–428. Springer.

Sun, Z.; Fu, X.; Huang, L.; Liu, A.; and Zha, Z.-J. 2025. Motion Aware Event Representation-driven Image Deblurring. In *Proceedings of the European conference on computer vision (ECCV)*, 418–435. Springer.

Tulyakov, S.; Gehrig, D.; Georgoulis, S.; Erbach, J.; Gehrig, M.; Li, Y.; and Scaramuzza, D. 2021. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16155–16164.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020. Event enhanced high-quality image recovery. In *Proceedings of the European conference on computer vision (ECCV)*, 155–171. Springer.

Wang, L.; Ho, Y.-S.; Yoon, K.-J.; et al. 2019. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10081–10090.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A General U-Shaped Transformer for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

Xiao, J.; Fu, X.; Liu, A.; Wu, F.; and Zha, Z.-J. 2022a. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12978–12995.

Xiao, J.; Fu, X.; Wu, F.; and Zha, Z.-J. 2022b. Stochastic window transformer for image restoration. *Advances in Neural Information Processing Systems*, 35: 9315–9329.

Xiao, J.; Fu, X.; Zhu, Y.; Li, D.; Huang, J.; Zhu, K.; and Zha, Z.-J. 2024. HomoFormer: Homogenized Transformer for Image Shadow Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25617–25626.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2583–2592.

Xu, S.; Sun, Z.; Zhu, J.; Zhu, Y.; Fu, X.; and Zha, Z.-J. 2024. DemosaicFormer: Coarse-to-Fine Demosaicing Network for HybridEVS Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1126–1135.

Yang, W.; Wu, J.; Li, L.; Dong, W.; and Shi, G. 2023. Event-based Motion Deblurring with Modality-Aware Decomposition and Recomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8327–8335.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, H.; Xie, H.; and Yao, H. 2024. Blur-aware Spatio-temporal Sparse Transformer for Video Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2673–2681.

Zhang, X.; and Yu, L. 2022. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17765–17774.

Zhang, X.; Yu, L.; Yang, W.; Liu, J.; and Xia, G.-S. 2023. Generalizing Event-Based Motion Deblurring in Real-World Scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10734–10744.

Zhou, S.; Chen, D.; Pan, J.; Shi, J.; and Yang, J. 2024. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2952–2963.