This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Event-Based Motion Deblurring Using Task-Oriented 3D Gaussian Event Representations

Shengdong Xue[1,*] Haoxiang Ma[1,*] Hao Chen[2] Zhen Yang[1] Yongjian Deng[3,†]
[1]College of Computer Science, Beijing University of Technology
[2]Key Lab of Computer Network and Information Integration, Southeast University
[3]College of Artificial Intelligence, Nankai University
{shengdong@emails., mahaoxiang@emails., yangzhen@}bjut.edu.cn,
haochen303@seu.edu.cn, yjdengty@foxmail.com

Abstract

Event-based motion deblurring has attracted increasing attention, as the high temporal resolution of event cameras provides motion cues unavailable to conventional RGB sensors, thereby enabling more effective deblurring. In real-world scenes, motion blur is often complex and nonlinear, with different regions exhibiting diverse motion speeds and directions. However, most existing approaches rely on handcrafted event representations that overlook such spatiotemporal motion heterogeneity, resulting in suboptimal deblurring performance. To address this limitation, we propose a learnable 3D Gaussian event representation module that adaptively selects key spatiotemporal coordinates for deblurring based on the blurred image content and the event density distribution. The event stream is then aggregated with 3D Gaussian weighting kernels to extract local motion features that are sensitive to motion direction and speed. Furthermore, to fully exploit the motion information encoded in the event representation, we adopt a two-stage fusion strategy. In the first stage, local motion features are used to enhance fine detail restoration. In the second stage, a bidirectional attention fusion module leverages one-dimensional Gaussian-weighted event frames to correct global spatial misalignment, leading to more accurate structural alignment. Extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness of our method and show that it consistently outperforms state-of-the-art approaches.

1. Introduction

Image deblurring is a fundamental problem in computer vision, where motion blur—particularly severe around edges—is a major source of image degradation [18]. Event cameras [7, 41], as novel visual sensors with high temporal resolution and high dynamic range, capture rich motion information between conventional RGB frames with microsecond-level precision. Due to their imaging mechanism, events are predominantly triggered along scene edges, providing informative motion cues that can significantly fa-

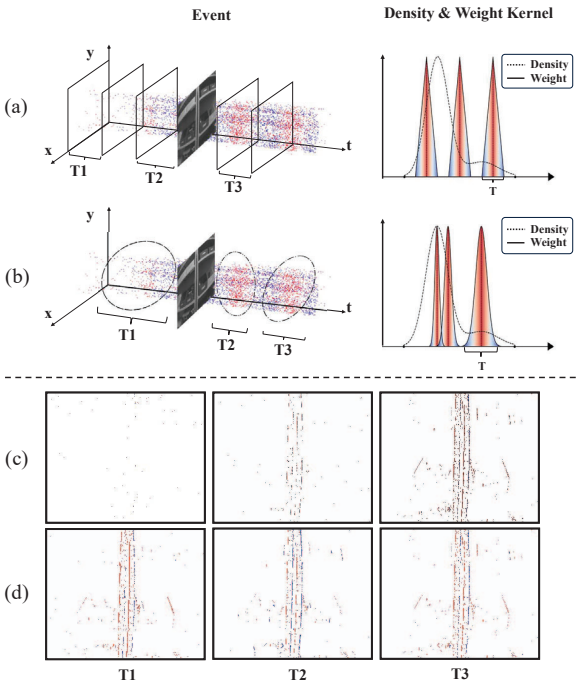

Figure 1. (a) The bilinear weight kernel. (b) 3D Gaussian weight kernel. (c) Event frame generated from a fixed kernel in sparse regions. (d) Event frame generated using our sample-adaptive Gaussian kernel. The density curve reflects the motion speed.

*: Equal contribution †: Corresponding author

cilitate deblurring. These properties make event cameras particularly well suited for assisting RGB image restoration.

However, the sparsity and irregular structure of event streams make their direct fusion with RGB images challenging, necessitating effective preprocessing to fully exploit the motion information they encode. In event-based motion deblurring, the dominant preprocessing strategy is to employ handcrafted fixed-weight kernel functions to aggregate sparse event points into event frames, enabling subsequent restoration networks to utilize event information [22, 26, 31, 39, 41, 42]. A representative approach is the Event Voxel Grid [42], which partitions events along the temporal axis into N bins and aggregates them into continuous event frames via fixed bilinear interpolation, as illustrated in Fig. 1 (a).

In real-world scenarios, however, event distributions are highly non-uniform, and motion patterns vary significantly in both speed and direction. For example, slow motion produces sparse events that require longer temporal integration, whereas fast motion generates dense events that demand shorter integration to avoid excessive edge blur. Existing event representations based on fixed-weight kernels lack sample adaptivity and cannot dynamically adjust to such spatiotemporal variations. Consequently, they often fail to assign appropriate weights to event-dense regions and produce low-quality representations in sparse areas. As shown in Fig. 1 (c), this leads to significant quality inconsistencies across regions, limiting the effective utilization of motion information for deblurring.

To address these limitations, we propose a learnable 3D Gaussian event representation that adaptively integrates local spatiotemporal regions, producing representations tailored to varying event densities and complex motion fields, as illustrated in Fig. 1 (b). Specifically, we first construct a spatiotemporal histogram of events and incorporate the blurred image as auxiliary guidance. A lightweight 3D convolutional neural network (3D-CNN) is then employed to extract degradation-aware spatiotemporal features. Inspired by SampleNet [15], we design a learnable 3D region sampler to identify key local motion regions. Based on these regions, multiple 3D Gaussian weighting kernels are applied to adaptively aggregate the event stream, effectively modeling nonlinear motion with diverse directions and speeds. In addition, we introduce one-dimensional Gaussian-weighted event frames to capture global motion cues, enabling joint modeling of local details and global structures.

To fully exploit the proposed representation, we further develop a two-stage fusion strategy. In the first stage, an attention mechanism integrates fine-grained motion cues from local event features with image textures to enhance detail restoration. In the second stage, we introduce a bidirectional attention fusion module that leverages global event features to correct spatial misalignment and improve structural consistency.

Our contributions are summarized as follows:

- We propose a learnable 3D Gaussian event representation that adapts to spatiotemporal event distributions and diverse local motion patterns, producing more effective representations for motion deblurring.
- We design a two-stage fusion strategy that combines local feature enhancement with a bidirectional attention module for global spatial alignment.
- Extensive experiments on both synthetic and real-world datasets demonstrate that our method achieves superior performance over state-of-the-art approaches.

2. Related Work

2.1. Image Deblurring

Traditional image deblurring was often formulated as a blur kernel estimation problem. With the rise of deep learning, CNN-based approaches have achieved significant improvements. U-Net-based restoration networks directly map blurred images to sharp outputs. For instance, [18] proposed a multi-scale deblurring network combining residual blocks and high-frame-rate images to synthesize blurred frames. [3] introduced a lightweight module without activation functions, simplifying attention to a gated dot-product. Transformer-based methods [2, 4, 14, 27, 32, 37] further enhance attention mechanisms, improving efficiency and performance. However, these methods rely solely on RGB inputs and lack motion cues during exposure, limiting their effectiveness in extreme motion scenarios.

2.2. Event-based Deblurring

Event cameras, with microsecond-level temporal resolution, have been increasingly used to assist image deblurring [9, 21, 25, 33, 40]. Directional Event Filtering (DEF) [12] extracts sharp edges, while [30] combines deblurring and super-resolution. EFNet [22] integrates transformer-based channel attention to fuse event and RGB features. Multi-scale spatiotemporal methods [36, 39] allow flexible temporal and spatial aggregation, and recurrent modules [26] capture motion at multiple scales. Other approaches employ frequency-domain fusion [13], video frame interpolation [23, 24], or spatiotemporal fusion [35]. [43] decouples edge and motion cues for improved results. Despite their progress in modality fusion, these methods mostly rely on handcrafted event representations with limited adaptivity, restricting the exploitation of embedded motion cues.

2.3. Event Representation

A common strategy is to aggregate events along the temporal dimension into frames [1, 5, 17, 28, 29, 31, 42]. The VoxelGrid [42] assigns weights based on temporal distance to encode timestamps and capture motion cues. EST [8]

introduced the first end-to-end learnable event spike tensor via MLP-based weighting. SCER [22] captures motion at multiple temporal strides, and EGER [39] adapts accumulation stride to simulate varying blur. DA [26] encodes richer motion but incurs high computational cost due to per-event processing. These methods share a common limitation: they cannot fully leverage diverse spatiotemporal motion cues in real-world scenes. In contrast, our approach adaptively models complex local motion fields, generating event frames better suited for motion deblurring.

3. Preliminary

3.1. 3D Gaussian Weight Kernel

In this work, we represent the event stream as a set of discrete events $\{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, where $(x_i, y_i, t_i) \in [0, 255]^3$ denotes the normalized and rescaled spatiotemporal coordinates of the i-th event, and $p_i \in \{0, 1\}$ denotes the event polarity ($p_i = 1$ for positive events and $p_i = 0$ for negative events), which we process separately. We define K event-integration kernels based on the 3D Gaussian distribution. Each kernel is parameterized by a mean vector $\boldsymbol{\mu}$, which specifies its center of focus, and a covariance matrix $\boldsymbol{\Sigma}$, which determines its attention range in 3D space. Specifically, the diagonal elements characterize the attention range along each dimension, while the off-diagonal elements capture the coupling relationships among different dimensions in the local nonlinear motion field.

$$\boldsymbol{\mu}_k = (x_k,\, y_k,\, t_k), \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{xx} & \rho_{xy} & \rho_{xt} \\ \rho_{xy} & \sigma_{yy} & \rho_{yt} \\ \rho_{xt} & \rho_{yt} & \sigma_{tt} \end{bmatrix} \tag{1}$$

Given an event (x_i, y_i, t_i), its corresponding weight under the k-th 3D Gaussian kernel is computed as

$$w_i^k = \exp\Big(-\tfrac{1}{2}\, \boldsymbol{\Delta}_i^\top\, \boldsymbol{\Sigma}_k^{-1}\, \boldsymbol{\Delta}_i\Big), \quad \boldsymbol{\Delta}_i = \begin{bmatrix} x_i - x_k \\ y_i - y_k \\ t_i - t_k \end{bmatrix}. \tag{2}$$

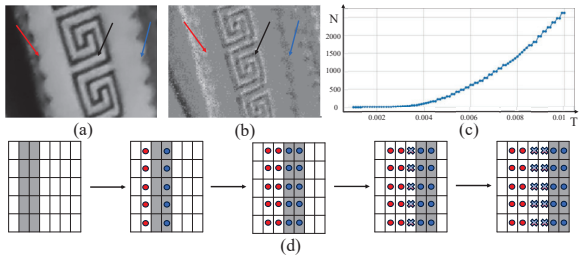

Figure 2. Visualization of polarity annihilation: (a) GT edges. (b) Event-frame edges with polarity accumulation. (c) Increasing number of annihilated pixels as the temporal range T increases. (d) A moving-edge example illustrating polarity annihilation.

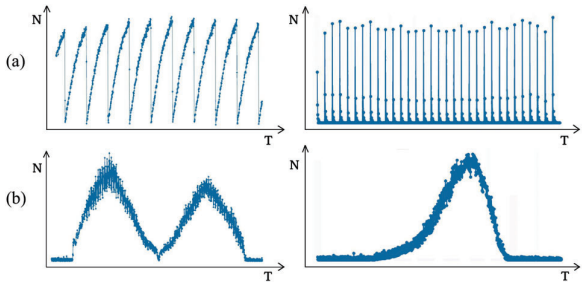

Figure 3. (a) A sample from the synthetic GoPro[18] dataset (with events generated by ESIM[20] and V2E[11]). This discrete distribution shape arises from the event simulation process, which generates data based on high-frame-rate interpolation. (b) An example from the real-world REBlur[22] dataset. The event-density curve along the t-axis reflects variations in camera motion speed.

We then aggregate the weighted events onto the 2D spatial grid (u, v) to form a 2D *event frame*:

$$E_k(u, v) = \sum_{i=1}^{N} w_i^k\, \delta(x_i - u)\, \delta(y_i - v), \tag{3}$$

where $u, v \in \{0, 1, \ldots, 255\}$. Thus, all events at pixel (u, v) contribute to the corresponding location according to their computed weights. Positive and negative polarities each produce K event frames, resulting in a total of $2K$ frames. We stack them along the channel dimension to obtain the final event-frame tensor of shape $[B, 2K, H, W]$:

$$\mathcal{E} = \big[E_1^+, \ldots, E_K^+,\; E_1^-, \ldots, E_K^-\big]. \tag{4}$$

3.2. Event Frame Characteristics

3.2.1. Polarity Annihilation

This phenomenon occurs when events with opposite polarities cancel each other during accumulation into an event frame. This happens because some moving edges generate only positive or only negative events, whereas others generate both, leading to ghosting artifacts and degraded edge quality. To mitigate this effect, positive and negative events are processed separately. Fig. 2 illustrates this phenomenon with a toy example of a moving edge on a white background, which causes polarity annihilation.

3.2.2. Event Temporal Distribution

As shown in Fig. 3, the event-density distributions of different samples across datasets exhibit significant variation. An adaptive sampling strategy should therefore allocate more sampling points to dense regions and fewer to sparse regions. Moreover, for regions with low event density, which usually correspond to slow motion, increasing the temporal accumulation range T can enhance the edge information in the event frame. In contrast, for regions with high event

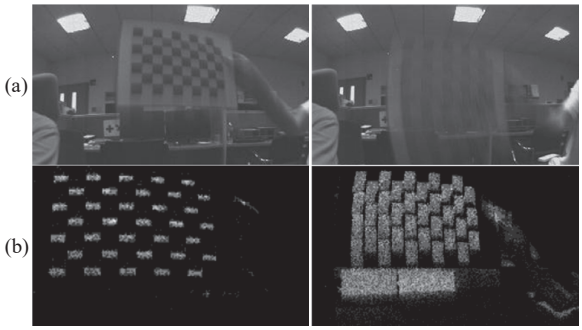

Figure 4. (a) Blurred images from the REBlur[22] dataset. (b) Event frames aggregated with different temporal ranges T, which affect the thickness of the event-frame edges.

density, which usually correspond to fast motion, reducing T can prevent overly thick edges. Fig. 4 further shows that different scenarios exhibit different motion speeds and thus different blur scales. Therefore, it is important to use a sample-adaptive temporal range T for event accumulation, rather than manually setting T to a fixed value.

4. Method

4.1. Architecture Overview

Fig. 5 illustrates the overall framework of our network. Given an event stream and a blurred image as prior guidance, our model first constructs a voxelized spatiotemporal representation with explicit 3D positional embeddings. A blur-aware guidance branch then modulates the event features via a lightweight blur-map generator. The fused representation is subsequently encoded by a stack of depthwise separable 3D convolutions to extract high-level spatiotemporal features, from which a multi-branch MLP sampler predicts a set of learnable 3D Gaussian kernels. These kernels serve as adaptive local attention weights for aggregating events into event frames. To fully exploit the complementary properties of these event frames, we design a two-stage fusion network based on EFNet[22]. In the first stage, multi-scale fusion is used to carefully integrate diverse local motion features generated by the 3D Gaussian kernels. In the second stage, a global event-position feature generated by a 1D Gaussian kernel (along the temporal dimension only) is used for global feature refinement.

4.1.1. Learnable 3D-Gaussian Representation Module

To generate sample-adaptive 3D Gaussian kernels, the event stream is first converted into a 3D count histogram of size (D, H, W). The polarity-separated voxel counts are computed as

$$V(x, y, t) = \sum \delta_{x,x_i}\, \delta_{y,y_i}\, \delta_{t,t_i}, \tag{5}$$

where $(x, y, t) \in \{0, \ldots, W-1\} \times \{0, \ldots, H-1\} \times \{0, \ldots, D-1\}$, and δ denotes the Kronecker delta function. We then construct a log-compressed magnitude channel and

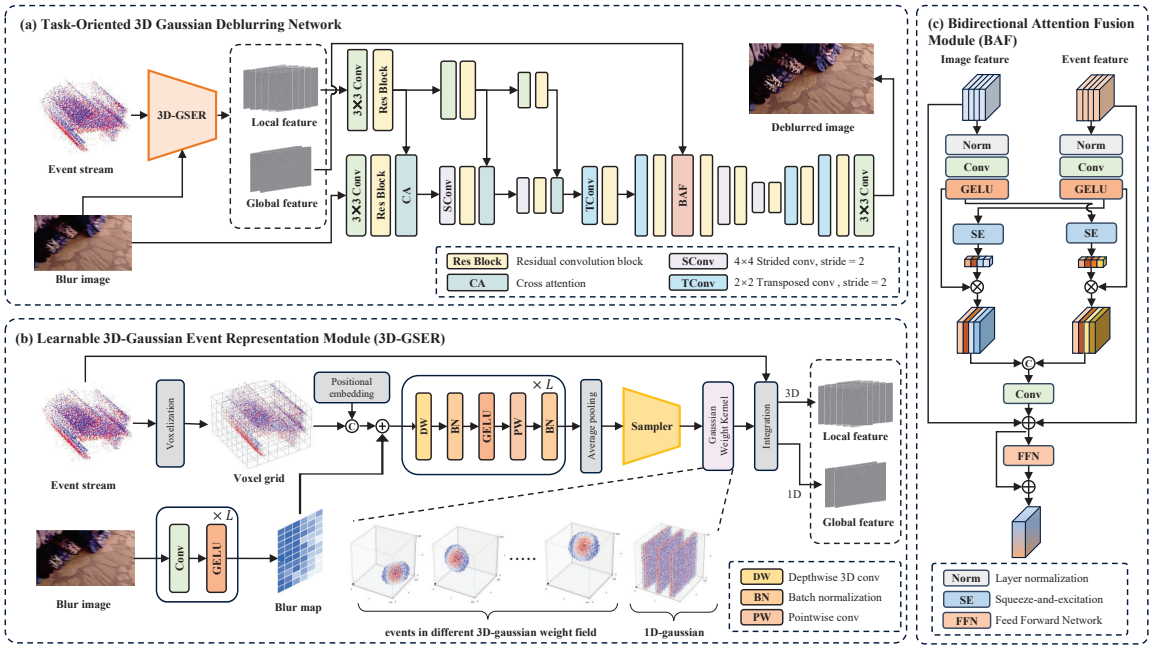

Figure 5. (a) Our Task-Oriented 3D Gaussian Deblurring Network. (b) The learnable 3D-Gaussian event representation module. (c) The Bidirectional Attention Fusion Module for global fusion, built upon SENet[10] and EFNet[22].

concatenate continuous 3D positional embeddings mapped to the range $[-1, 1]$. The resulting encoder input is

$$\tilde{V}(x, y, t) = \text{concat}\left(\log\left(V(x, y, t)\right),\ E(x, y, t)\right), \quad (6)$$

where the positional embedding $E(x, y, t)$ is defined as

$$E(x, y, t) = \phi(x, y, t) = 2\left[\frac{t}{D}, \frac{x}{W}, \frac{y}{H}\right] - 1. \quad (7)$$

To make the module aware of blurred regions, given the blurred image $I_b \in \mathbb{R}^{H\times W\times 3}$, we employ a lightweight convolutional block to generate a blur score map S_b, which highlights severely blurred regions. After broadcasting S_b to the voxel domain, we use a learnable scalar α to modulate the event features:

$$\tilde{V}_{\text{guided}}(x, y, t) = \tilde{V}(x, y, t) + \alpha\, S_b(x, y), \quad (8)$$

where the blur score map S_b is computed as

$$S_b = \sigma\left(\text{Conv}(I_b)\right) \in [0, 1]^{H\times W}. \quad (9)$$

To extract hierarchical spatiotemporal features, we stack L blocks of depthwise separable 3D convolutions. For an input tensor $X \in \mathbb{R}^{C\times D\times H\times W}$, each block applies the following sequence of operations:

$$Y = \sigma\left(\text{BN}\left(\text{PW}\left(\sigma\left(\text{BN}\left(\text{DW}(X)\right)\right)\right)\right)\right), \quad (10)$$

where DW denotes a depthwise 3D convolution, PW denotes a pointwise convolution, BN denotes batch normalization, and σ represents the GELU activation function. A global spatiotemporal average-pooling operation yields a global feature $F_{\text{global}} \in \mathbb{R}^{C_L}$. Inspired by the point-cloud method SampleNet[15], which uses multiple MLPs to learn the coordinates of 3D points, we employ multiple MLP heads sampler to predict, for each of the K kernels, a mean vector and a covariance matrix.

$$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k = \text{Sampler}\left(F_{\text{global}}\right). \quad (11)$$

4.1.2. Bidirectional Attention Fusion Module

The sample-adaptive 3D Gaussian kernels focus only on local spatiotemporal regions of the event stream and therefore primarily capture local motion fields. As a result, the event frames generated by different kernels may become spatially misaligned due to their different coordinates along the temporal axis. To address this issue, we further use a 1D Gaussian kernel (along the t-axis only) to generate an event frame that provides global edge-position cues for alignment. To fully exploit the complementary information provided by the global feature of the 1D Gaussian event frame, we propose an effective Bidirectional Attention Fusion (BAF) module for more coherent global fusion.

Given an image feature $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ and an event feature $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$, both features are first processed by normalization layers, followed by 1×1 convolutions to adjust their channel dimensions. The features are then passed through a GELU activation and fed into a Squeeze-and-Excitation (SE)[10] block, which applies global average pooling to compute channel-wise attention weights $\mathbf{A}_I$ for image features and $\mathbf{A}_E$ for event features:

$$\mathbf{A}_I = \text{SE}(\mathbf{I}), \quad \mathbf{A}_E = \text{SE}(\mathbf{E}). \quad (12)$$

The attention-modulated image and event features are then obtained via element-wise multiplication:

$$\mathbf{F}_I = \mathbf{I} \odot \mathbf{A}_I, \quad \mathbf{F}_E = \mathbf{E} \odot \mathbf{A}_E. \quad (13)$$

These modulated features are concatenated along the channel dimension, and a 1×1 convolution is applied to reduce the feature dimensionality. Finally, a Feed-Forward Network (FFN) processes the fused features, followed by a residual addition operation.

5. Experiment

5.1. Datasets and Experimental Settings

Dataset. We use three datasets containing synthetic, semi-synthetic, and real-world frames and events to evaluate the proposed method.

GoPro: The GoPro dataset[18] consists of paired blurred and corresponding ground-truth sharp images, both with a resolution of 1280×720. The blurred images are synthesized by averaging a series of connected sharp frames and then applying gamma correction. We use the GoPro dataset together with simulated events generated by the event simulator ESIM[20].

HS-ERGB: The HS-ERGB dataset[28] consists of sharp videos and real-world events. We use the normal-blur version released by [39], in which blurry frames are synthesized by averaging 49 interpolated images.

REBlur: The REBlur dataset[22] contains real-world event sequences paired with blurry images and reliable ground-truth sharp images. It covers 12 types of linear and non-linear motion. Following the original setting, we use 486 blurry-sharp image pairs with associated events for training and 983 pairs for testing.

Training Settings. We train our model on a single RTX 3090 GPU using PyTorch, directly on the GoPro-ESIM dataset without any pretraining. The input images are cropped into patches of size 256×256, and the corresponding event streams are synchronously segmented for training. The batch size is set to 4. We use the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and set the initial learning rate to 2×10^{-4}. The learning rate is scheduled with cosine annealing, where $T_{\max} = 400$K iterations. Data augmentation includes random rotations and flips. For the HS-ERGB

Table 1. Quantitative results on the GoPro, HS-ERGB, and REBlur test sets. FLOPs are estimated at a resolution of 224×224.

Method	Publication	Image	Event	GoPro		HS-ERGB		REBlur		Params(M)	FLOPs(G)
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
EDI[19]	CVPR2019	✓	✓	29.06	0.943	23.93	0.704	36.52	0.964	0.5	–
EVDI[38]	CVPR2022	✓	✓	30.40	0.906	25.13	0.707	–	–	0.4	–
HINet[2]	CVPR2021	✓	×	32.71	0.959	27.32	0.807	35.58	0.965	88.7	241.5
Restormer[37]	CVPR2022	✓	×	32.92	0.961	27.55	0.808	35.50	0.959	26.1	216.1
NAFNet[3]	ECCV2022	✓	×	33.71	0.967	27.64	0.811	36.15	0.969	67.8	96.8
UFPNet[6]	CVPR2023	✓	×	34.06	0.968	27.64	0.809	36.11	0.968	80.3	361.2
HINet+[2]	CVPR2021	✓	✓	34.63	0.968	27.66	0.808	37.92	0.976	88.7	241.5
UFPNet+[6]	CVPR2023	✓	✓	35.22	0.972	27.68	0.809	37.97	0.976	80.3	361.2
EFNet[22]	ECCV2022	✓	✓	35.46	0.972	26.68	0.800	38.12	0.975	8.5	153.9
EIFNet[34]	MM2023	✓	✓	35.99	0.979	26.74	0.797	37.16	0.972	10.8	145.1
MAENet[26]	ECCV2024	✓	✓	36.07	0.976	27.93	0.812	38.47	0.978	13.9	149.7
SepNet[43]	ICCV2025	✓	✓	36.70	0.977	–	–	38.53	0.977	–	–
Ours	–	✓	✓	36.86	0.977	28.55	0.813	38.68	0.977	16.7	172.6

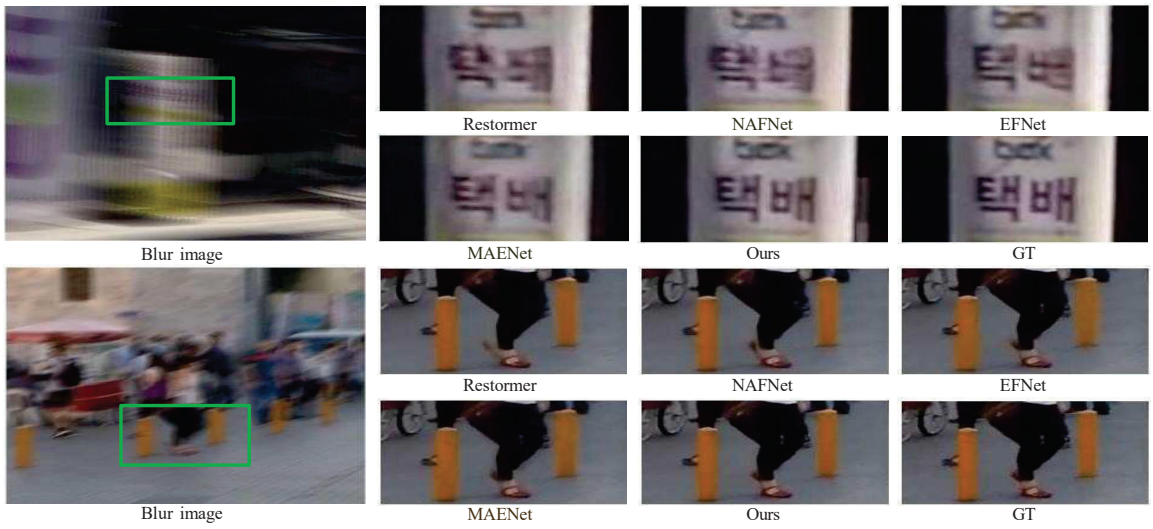

Figure 6. Visual comparison on the GoPro dataset. Our method restores fine textures and structural details more faithfully.

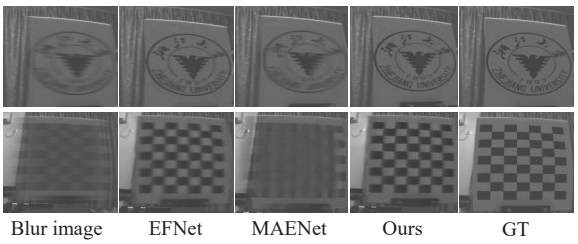

Figure 7. Visual comparison on the REBlur dataset. Our method restores fine textures and structural details more faithfully.

and REBlur datasets, we fine-tune the model pretrained on GoPro using their respective training sets. The fine-tuning stage runs for 4K iterations with a learning rate of 2×10^{-5}, while all other hyperparameters remain the same as those used in the initial training stage.

5.2. Comparison with State-of-the-Art Methods

Table 1 compares our method with recent state-of-the-art image-based and event-based deblurring approaches on GoPro [18], HS-ERGB [28], and REBlur [22]. Our model achieves the best PSNR on all three benchmarks, surpassing previous methods by 0.16 dB, 0.62 dB, and 0.15 dB on the synthetic (GoPro), semi-synthetic (HS-ERGB), and

real-world (REBlur) datasets, respectively. The qualitative results further show that our method recovers finer textures with fewer spatial distortions, benefiting from the adaptive aggregation of event information and the proposed two-stage fusion network.

GoPro. On GoPro, our two-stage deblurring network establishes a new state of the art. The learnable 3D Gaussian kernels, together with the attention-based fusion strategy, exploit event motion information more effectively than previous handcrafted representations. As illustrated in Fig. 6, image-only methods tend to generate overly smooth results and fail to recover high-frequency edge structures. EFNet [22] and MAENet [26] restore sharper edges, but still introduce noticeable noise and ringing artifacts around them. This is because their manually designed event encodings are not sufficiently flexible to model complex and spatially varying motion, and therefore cannot fully exploit local event cues.

Table 2. Ablation study of the effectiveness of each module.

Method	Blur Map	BAF	3D-GSER	PSNR	
				GoPro	REBlur
Baseline	×	×	×	36.13	38.01
A	×	×	✓	36.51	38.37
B	✓	×	✓	36.61	38.41
C	×	✓	✓	36.76	38.53
D	✓	✓	✓	**36.86**	**38.68**

REBlur. Representative examples from the REBlur test set are shown in Fig. 7. EFNet [22] produces relatively sharp results, yet residual motion blur remains because the SCER representation is constructed with a fixed temporal range T and cannot adapt to different motion speeds. MAENet [26] alleviates part of the blur, but tends to introduce artifacts and ghosting, especially under fast motion, due to the limited adaptability of its event representation. In contrast, our approach recovers both clean textures and sharp edges, highlighting the effectiveness of the proposed two-stage fusion strategy and the bidirectional attention fusion block. Benefiting from the adaptive local attention induced by the 3D Gaussian kernels, the network robustly restores heavily blurred regions while preserving the content and sharpness of already clear areas.

Table 3. Ablation study of different event representations on the GoPro dataset.

Method	Type	PSNR	SSIM
SBT[31]	handcrafted	35.98	0.9718
Voxel Grid[42]	handcrafted	36.13	0.9719
SCER[22]	handcrafted	35.95	0.9711
DA[26]	handcrafted	36.09	0.9713
EST[8]	learning-based	35.86	0.9704
LETC[16]	learning-based	35.84	0.9710
3D-GSER	learning-based	**36.51**	**0.9751**

5.3. Ablation Study and Discussion

We conduct several ablation experiments on the GoPro dataset to verify the effectiveness of the proposed method.

Effectiveness of Blur Map and BAF. To demonstrate the effectiveness of the blur map and BAF, we conduct the ablation study shown in Table 2. We use the Voxel Grid [42] event representation as the baseline. Variant A adopts our event representation, while variants B and C further incorporate the blur-map prior and BAF, respectively. Variant D corresponds to the complete model. Compared with A, B improves the PSNR by 0.10 dB, indicating that the blur-map prior provides effective guidance for degraded regions. With the introduction of the global fusion module in C, the PSNR further increases to 36.76 dB, demonstrating that integrating global features can enhance structural consistency. Variant D achieves the best performance by combining the blur prior with both local and global event features, which validates the complementarity of these modules and the effectiveness of the overall design.

Figure 8. Visual comparison in the ablation study of different modules.

Figure 9. Visual comparison of different event representations.

We further validate the effectiveness of each module through the visualizations in Fig. 8. Variant A exhibits ghosting artifacts in local regions, mainly due to temporal misalignment caused by the 3D Gaussian kernels and the lack of task-oriented guidance from the blur prior. Although variant B improves detail restoration in blurred regions, structural edges still show noticeable shifts. By introducing BAF for global feature fusion, variant C effectively corrects edge positions and suppresses ghosting artifacts. Finally, variant D achieves the best performance in both detail sharpness and structural consistency.

Effectiveness of Event Representation. To demonstrate the superiority of our event representation for motion deblurring, we compare it with several alternative representations, as shown in Table 3. For a fair comparison, all methods use the same number of bins (or kernels). The results reveal clear limitations of the alternative representations. SBT [31] and Voxel Grid [42] uniformly divide the entire exposure time, which is inadequate for modeling spatially varying blur. Although SCER [22] uses multiple temporal ranges to accumulate events, its handcrafted integration ranges ($T/6$, $T/3$, and $T/2$) may not be suitable for all blur scenarios with different motion speeds. We also compare with learning-based event representations, including EST [8] and LETC [16]. While these methods can learn weighting kernels for event integration, the kernel positions along the temporal axis remain uniformly distributed and therefore cannot adapt to the temporal distribution of each sample. In contrast, our representation not only adapts to the spatiotemporal distribution of the event stream by sampling multiple 3D coordinates, but also dynamically adjusts the attention range through the adaptive 3D Gaussian covariance matrix. As a result, it achieves a 0.38 dB improvement over the best alternative representation.

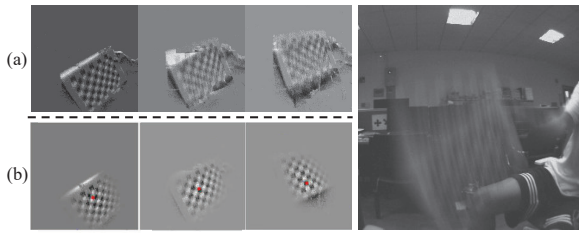

Figure 10. Visualization of SCER and 3D-GSER.

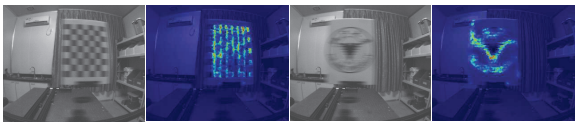

Figure 11. Visualization of the blur map.

Visualization. Fig. 10 shows an REBlur sample with extreme motion. SCER uses a fixed temporal window, which leads to thick edges and polarity cancellation, whereas 3D-GSER adapts to the motion speed more effectively, as indicated by the red dots denoting kernel centers. Fig. 11 further confirms that the blur map reliably highlights degraded regions, thereby providing effective guidance for spatiotemporal feature extraction.

6. Conclusion

In this paper, we present a task-oriented event representation framework for event-based motion deblurring. At the core of our method is a 3D Gaussian event representation module, which adaptively identifies informative spatiotemporal regions by jointly leveraging event density and blur-aware priors. This module enables dynamic aggregation of event streams through learnable 3D Gaussian kernels, thereby effectively capturing localized motion patterns under varying motion directions and speeds. To fully exploit the enriched motion cues, we further design a two-stage fusion strategy. In the first stage, local motion features are used to restore fine-grained details, while in the second stage, the bidirectional attention fusion module incorporates globally aligned event features to correct structural inconsistencies. Extensive experiments on synthetic, semi-synthetic, and real-world datasets demonstrate that our method achieves superior deblurring performance and generalizes well across diverse motion and blur scenarios.

Acknowledgments. This work is jointly supported by Beijing Natural Science Foundation (4252026) and the National Natural Science Foundation of China (62573369).

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[2] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 182–192, 2021. 2, 6

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 2, 6

[4] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2

[5] Yongjian Deng, Hao Chen, and Youfu Li. Mvf-net: A multi-view fusion network for event-based object classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8275–8284, 2021. 2

[6] Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. 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*, pages 18105–18114, 2023. 6

[7] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

[8] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5633–5643, 2019. 2, 7, 8

[9] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021. 2

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4, 5

[11] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 3

[12] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3320–3329, 2020. 2

[13] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024. 2

[14] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2

[15] Itai Lang, Asaf Manor, and Shai Avidan. Samplenet: Differentiable point cloud sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7578–7588, 2020. 2, 5

[16] Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 7, 8

[17] Yuhan Liu, Yongjian Deng, Hao Chen, and Zhen Yang. Video frame interpolation via direct synthesis with the event-based reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8477–8487, 2024. 2

[18] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 1, 2, 3, 5, 6

[19] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. 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*, pages 6820–6829, 2019. 6

[20] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018. 3, 5

[21] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4531–4540, 2021. 2

[22] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*, pages 412–428. Springer, 2022. 2, 3, 4, 5, 6, 7, 8

[23] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 2

[24] Lei Sun, Daniel Gehrig, Christos Sakaridis, Mathias Gehrig, Jingyun Liang, Peng Sun, Zhijie Xu, Kaiwei Wang, Luc Van Gool, and Davide Scaramuzza. A unified framework for event-based frame interpolation with ad-hoc deblurring in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[25] Lei Sun, Andrea Alfarano, Peiqi Duan, Shaolin Su, Kaiwei Wang, Boxin Shi, Radu Timofte, Danda Pani Paudel, Luc Van Gool, Qinglin Liu, et al. Ntire 2025 challenge on event-based image deblurring: Methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1324–1341, 2025. 2

[26] Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha. Motion aware event representation-driven image deblurring. In *European Conference on Computer Vision*, pages 418–435. Springer, 2024. 2, 3, 6, 7

[27] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European conference on computer vision*, pages 146–162. Springer, 2022. 2

[28] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 2, 5, 6

[29] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 2

[30] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*, pages 155–171. Springer, 2020. 2

[31] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. 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*, pages 10081–10090, 2019. 2, 7, 8

[32] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 2

[33] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 2

[34] Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi. Event-based motion deblurring with modality-aware decomposition and recomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8327–8335, 2023. 6

[35] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, and Guangming Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6531–6539, 2024. 2

[36] Wei Yu, Jianing Li, Shengping Zhang, and Xiangyang Ji. Learning scale-aware spatio-temporal implicit representation for event-based motion deblurring. In *Forty-first International Conference on Machine Learning*, 2024. 2

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 2, 6

[38] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 6

[39] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Generalizing event-based motion deblurring in real-world scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10734–10744, 2023. 2, 3, 5

[40] Fengan Zhao, Qianang Zhou, and Junlin Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In *European Conference on Computer Vision*, pages 190–205. Springer, 2024. 2

[41] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023. 1, 2

[42] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019. 2, 7, 8

[43] Yufei Zhu, Hao Chen, Yongjian Deng, and Wei You. Separation for better integration: Disentangling edge and motion in event-based deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14732–14742, 2025. 2, 6