Learning Dual Modality Interactions for Event-based Motion Deblurring

Zeyu Xiao, Zhuoyuan Li, Yang Zhao, Yu Liu, Zhao Zhang, and Wei Jia

***Abstract*—Event cameras hold great potential for motion deblurring because they capture motion information with microsecond precision, offering robustness to motion blur. However, the limited interaction between RGB frames and event streams presents a significant challenge, preventing the full utilization of the event cameras' unique advantages. To address this, we propose Dual frame-event Interaction and introduce a multi-scale Network structure, DuInt-Net. DuInt-Net aims to tackle two key challenges: (1) enhancing the representational and interaction capabilities between RGB frames and event streams, and (2) adaptively selecting richer visual features for improved motion deblurring. We introduce an event-frame joint interaction module that consists of three branches: a base branch, a global awareness attention branch, and a local enhancement attention branch. The base branch processes essential pixel-level features that retain the original structural information. The global branch integrates event data to improve large-scale motion understanding, while the local branch uses large-kernel convolutions to refine fine-grained details in RGB frames. For superior reconstruction performance, we also propose the event-guided multi-scale fusion attention module, which effectively combines local visual information and global frame-event relationships. Extensive experiments demonstrate that DuInt-Net achieves superior performance, both quantitatively and qualitatively, showcasing its superior motion deblurring capabilities.**

***Index Terms*—Image deblurring, Event camera, Cross-modal learning**

I. Introduction

Motion blur, typically caused by camera shake or object motion during exposure, severely degrades visual quality. This degradation poses significant challenges for various computer vision tasks, including image reconstruction [1]–[3], visual tracking [4]–[6], image segmentation [7], video surveillance [8], [9], and video compression [10]–[14]. The core challenge in motion deblurring is to restore sharp images while preserving essential edge structures and fine details, which are often lost during the blurring process. This makes it a highly complex task that requires sophisticated techniques to differentiate and recover the fine nuances in blurred images.

Corresponding author: Wei Jia.

Zeyu Xiao is with the department of Electrical and Computer Engineering, National University of Singapore, 117583, Singapore (e-mail: zeyuxiao1997@163.com).

Zhuoyuan Li is with the MOE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei 230027, China (email: zhuoyuanli@mail.ustc.edu.cn).

Yang Zhao, Zhao Zhang and Wei Jia are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: yzhao@hfut.edu.cn, cszzhang@gmail.com, jiawei@hfut.edu.cn)

Yu Liu is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: yuliu@hfut.edu.cn)

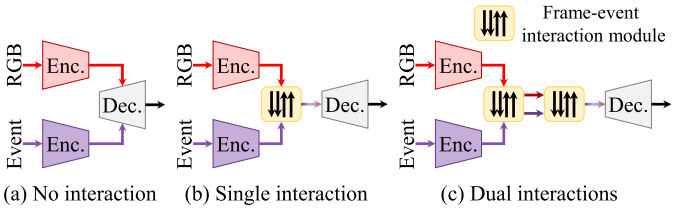

Fig. 1. (a) Non-interaction methods rely on basic fusion strategies that fail to incorporate meaningful event-frame interactions in the spatio-temporal domain. (b) Single-interaction methods integrate RGB frames and event streams in a limited fashion, not fully leveraging the microsecond precision and motion blur robustness provided by event cameras. (c) We propose the *dual frame-event interactions* approach to enhance both the representational and interaction capabilities between RGB and event modalities, adaptively fusing richer visual features for more effective deblurring.

Early methods for motion deblurring, such as the Wiener filter [15] and the Richardson-Lucy algorithm [16], were based on Bayesian frameworks and used iterative approaches to restore sharp images. While effective for certain blur types, these methods were computationally expensive and limited in their ability to handle complex blur patterns. In subsequent years, researchers focused on developing more sophisticated image priors [17]–[20] or introducing more complex data terms [21] to enhance deblurring performance under a wider range of conditions. With the advent of deep learning, particularly convolutional neural networks (CNNs) [22], [23], vision Transformers [24]–[26], and MLP-based architectures [27], significant progress has been made in motion deblurring. These models have demonstrated impressive results across various scenarios, particularly in challenging conditions such as large motion blur or noisy inputs. More recently, Mamba-based architectures [28], [29] have been proposed to enhance performance further, leveraging the strengths of both convolutional and Transformer-based designs. Despite recent advancements, motion deblurring remains a highly ill-posed problem, primarily due to the challenge of accurately estimating spatially varying blur kernels from limited and noisy observations. This issue becomes more pronounced under severe blur, where essential motion information is lost during exposure, complicating restoration. While recent deep learning approaches have shown promise, more robust models capable of handling complex blur patterns are still needed to ensure high-quality restoration in even the most challenging scenarios.

Event cameras, which asynchronously detect pixel-level brightness changes with microsecond precision, excel at capturing high-temporal-resolution motion, making them highly suitable for motion deblurring tasks [30]–[38]. Despite recent

advances, the limited interaction between RGB frames and event streams remains a critical bottleneck, hindering the full utilization of the event cameras' superior precision. Current methods primarily rely on basic fusion strategies to combine the spatial complementarity of RGB frames and event data. However, they fail to address modality redundancy and often overlook the potential for deeper, more meaningful event-frame interactions (see Figure 1). This limitation ultimately restricts the performance of these approaches, leaving room for significant improvements in integrating the two modalities.

In this paper, we introduce a novel approach for motion deblurring by leveraging Dual frame-event Interactions within a multi-scale Network architecture called DuInt-Net. This model addresses two fundamental challenges when integrating event cameras into motion deblurring tasks: (1) how to enhance the representational and interaction capabilities between RGB frames and event data, and (2) how to adaptively select and fuse richer visual features from both sources to achieve superior deblurring performance (see Figure 1). To enhance the interaction between RGB frames and event data, we propose the event-frame joint interaction (EFJI) module. This module integrates event data into the RGB feature extraction pipeline through three distinct branches: global, local, and base. The global branch leverages the event camera's high temporal resolution to capture overall motion patterns and object localization, providing comprehensive motion context. The local branch refines fine-grained spatial details using large-kernel convolutions to enhance the RGB frames with precise event cues. Meanwhile, the base branch encodes RGB features independently, ensuring a robust foundational representation that anchors the fusion process. This three-branch organization ensures that the event and frame data interact efficiently, each branch contributing distinct insights into motion understanding. We further propose the event-guided multi-scale fusion attention (EMFA) module, which enhances reconstruction performance by extracting rich visual information from both RGB and event modalities. The EMFA module combines the strengths of both sources through a gated fusion mechanism. It captures multi-scale local features with convolutional blocks and models global relationships between frames and events. This integration of global and local features at multiple scales ensures that the system accurately preserves fine details while reconstructing broader motion contexts. By incorporating the EFJI and EMFA modules into a Unet-based framework, DuInt-Net achieves state-of-the-art performance in motion deblurring. Extensive experiments demonstrate that our method outperforms existing solutions, delivering robust and high-quality restoration even in challenging conditions.

In summary, this paper presents three key contributions: (1) The EFJI module: We propose a novel module that effectively integrates event data and RGB frames using global and local attention mechanisms, enhancing motion understanding and enabling finer detail refinement. (2) The EMFA module: We introduce a multi-scale fusion attention mechanism that strengthens cross-scale interactions by combining complementary features from event and RGB modalities, resulting in a more robust feature representation. (3) Advanced results: We demonstrate that DuInt-Net achieves superior performance on multiple benchmark datasets, showcasing the effectiveness of our approach compared to existing methods.

II. Related Work

A. Motion Deblurring

Motion deblurring is a challenging, ill-posed problem that aims to reconstruct sharp images from blurry observations. Early methods rely on handcrafted priors and empirical constraints [39]–[43]. With the rise of deep learning, data-driven approaches become dominant. Initial CNN-based methods estimate blur kernels and refine them with prior-based restoration techniques [44]–[47]. However, inaccuracies in kernel estimation often introduce artifacts due to treating blur kernels and clear images separately. Subsequent methods focus on direct clear image estimation. Nah *et al.* [48] propose a multi-scale CNN that leverages coarser-scale outputs to guide finer estimates, while Zhang *et al.* [49] introduce a hierarchical multi-patch network with cross-stage feature concatenation. Cho *et al.* [50] develop a multi-input, multi-output U-Net for improved efficiency. Recently, Transformers gain attention for modeling global contexts in deblurring. Zamir *et al.* [51] introduce channel-wise self-attention, and Tsai *et al.* [26] use intra- and inter-strip tokens for feature reweighting. In video deblurring, spatio-temporal correlations are exploited via recurrent networks and CNNs [52], [53], while optical flow [54] and deformable convolutions [55] further enhance modeling. Transformers also demonstrate potential for capturing long-range dependencies [56].

B. Event-based Motion Deblurring

Event cameras, inspired by the human visual system, capture rapid scene dynamics with high temporal resolution and low latency, making them invaluable for motion deblurring and precise image restoration. Early works employ physical models to describe the relationship between sharp and blurry images [30], [57], but performance is hindered by sensor noise. More recent methods have shifted to learning-based approaches [38], [58], [59], with some using events as auxiliary inputs in a unidirectional manner, integrating event features at single levels [36], [60]. In contrast, others employ cross-modal attention modules for multi-level fusion [35], [61]. Bidirectional methods treat events and frames equally, facilitating single- [31], [62] or multi-level interactions [34]. Additionally, real-world challenges like unknown exposure times have been addressed to enhance practical applications [63]. Recently, Shen *et al.* [64] propose a two-stage framework that first restores degraded real-world event streams and then uses the restored events to guide image deblurring. In this paper, we propose a novel framework for event-based motion deblurring that addresses the challenge of limited interaction between RGB frames and event streams by assuming clean event inputs and focusing on effective event-RGB joint interaction and fusion. Our method leverages cross-modal attention mechanisms and multi-level fusion to fully exploit event cameras' microsecond precision and motion blur robustness, improving the overall deblurring performance.

C. Event-based Video Processing

One of the key advantages of event cameras is their ability to capture motion information during exposure, serving as crucial motion cues for deblurring [65]–[67]. This enables them to effectively address motion blur, outperforming traditional RGB cameras, especially in fast-moving scenes. In video frame interpolation, methods like TimeLens [68] leverage event data to enhance accuracy and temporal resolution. Recent works focus on sophisticated interaction modules to facilitate seamless event-RGB fusion, leading to substantial performance improvements [69]–[73]. Event cameras also mitigate rolling shutter artifacts by providing real-time motion feedback [74]–[77], improving video quality in dynamic environments. They also demonstrate strong potential in various tasks, including depth estimation [78]–[80], high-dynamic-range imaging [81]–[84], deraining [61], low-light enhancement [85], [86], and video super-resolution [87]–[89], underscoring their versatility in advancing visual applications.

III. Method

A. Overview

Given a motion-blurred image I^B and an event stream $E_T \triangleq \{(x_i, y_i, p_i, t_i)\}_{t_i \in T}$, which captures all events triggered during the exposure time T, the goal is to recover a sharp, clear frame $\hat{I}$. The event stream E_T consists of asynchronous, high-temporal-resolution events, where each event $\{(x_i, y_i, p_i, t_i)\}$ represents a change in intensity at pixel location (x_i, y_i) with polarity $p = \pm 1$, indicating whether the intensity increased or decreased, at time t_i. These events provide valuable temporal information, which is often lost in traditional frame-based imaging systems.

To achieve this, we propose DuInt-Net, a framework that fuses spatial information from I^B with temporal information from E_T through dual cross-modal interactions. The sharp frame reconstruction can be formulated as

$$\hat{I} = \text{DuInt-Net}(I^B, E_T). \tag{1}$$

As illustrated in Figure 2, we first extract multi-scale features from I^B and E_T using separate encoders: F_I^l for the image and F_E^l for the event stream, where $l \in 1, 2, 3$ indicates the level. To account for the domain gap, we employ distinct encoders. For the RGB encoder $\text{Encoder}_\text{I}(\cdot)$, we use residual blocks and residual atrous spatial pyramid pooling (ResASPP) blocks [90], [91] to capture large receptive fields and multi-scale features, essential for deblurring [54]. For events, we use stacked convolution layers in $\text{Encoder}_\text{E}(\cdot)$ to handle noise and sparsity as

$$F_I^l = \text{Encoder}_\text{I}(I^B), \tag{2}$$
$$F_E^l = \text{Encoder}_\text{E}(E_T). \tag{3}$$

The structures of both feature encoders are shown in Figure 3.

The extracted features are then processed through the EFJI modules, which enhance the interaction between event and frame data, resulting in refined features $\overline{F}_I^l$ and $\overline{F}_E^l$. These features are further refined by the EMFA modules, which fuse the spatial and temporal information, yielding the final fused

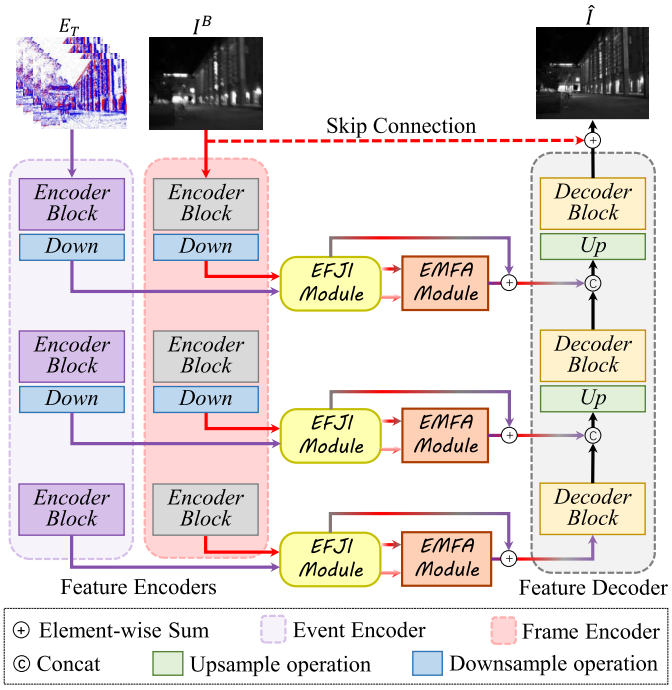

Fig. 2. Overview of the proposed DuInt-Net. DuInt-Net adopts a multi-scale hierarchical design that jointly leverages the EFJI (event-frame joint interaction) and EMFA (event-guided multi-scale fusion attention) modules. It comprises dual-stream encoders for event and frame inputs, enabling rich cross-modal interactions at multiple feature levels. The fused features are progressively decoded to produce the final deblurred output, with the skip connection operation preserving spatial details.

representations $\overline{F}^l$. To effectively restore the sharp image, the output features from the EMFA modules are combined with the original frame features $\overline{F}_I^l$ and fed into the decoder. This step facilitates the deblurring process.

A residual connection between the blurred image I^B and the decoder output reduces the learning difficulty, ensuring that the network can effectively focus on refining the details rather than learning the entire image from scratch. The final reconstruction is given by

$$\hat{I} = \text{Decoder}(\overline{F}_I^l + \overline{F}^l) + I^B, \tag{4}$$

where $\text{Decoder}(\cdot)$ denotes the decoder, which consists of both residual blocks and ResASPP blocks.

The encoder-decoder architecture is illustrated in Figure 3, where the encoder captures high-level spatial features and the decoder refines these features to produce the final deblurred image. The differences in encoder structures reflect the distinct characteristics of RGB frames and event data. We design separate encoders tailored to their specific roles in the deblurring process. The frame encoder adopts a hierarchical convolutional structure with downsampling residual blocks and ResASPP modules to effectively capture multi-scale spatial features and expand receptive fields, which is crucial for reconstructing fine textures, structural edges, and global context in blurred RGB frames. In contrast, the event encoder employs a simpler design with stacked convolutional layers, omitting ResASPP or downsampling residual blocks, to efficiently process sparse, asynchronous event data and extract high-temporal-resolution

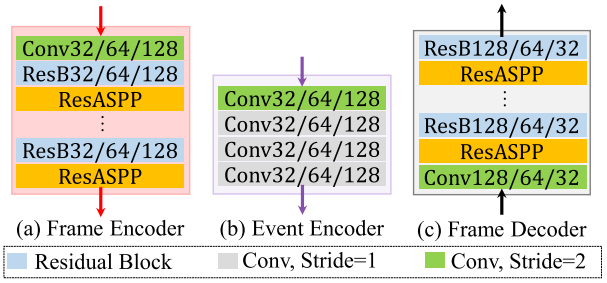

Fig. 3. Detailed architectures of the proposed feature encoders and decoder. (a) The frame encoder integrates convolutional layers, residual blocks with downsampling, and ResASPP modules for enhanced spatial encoding. (b) The event encoder employs a series of convolutional layers to extract multi-level event representations. (c) The frame decoder mirrors the encoder structure with upsampling residual blocks and ResASPP modules to reconstruct high-quality frames progressively.

motion cues with minimal computational overhead. Since events inherently provide localized motion information, extensive spatial context aggregation is less critical. Overall, this design enables the frame encoder to focus on spatial structures and the event encoder to capture temporal dynamics, allowing each to leverage its modality's strengths efficiently and robustly.

B. Event-Frame Joint Interaction

Integrating event and RGB data can significantly enhance motion deblurring by providing complementary information from both modalities. However, existing methods struggle to effectively fuse these sources due to limitations in representation and insufficient cross-modal interactions, leading to underutilization of the rich temporal and spatial cues from event cameras and RGB frames. Strengthening these interactions is crucial for capturing more detailed and precise representations, particularly in motion deblurring, where both fine details and motion information are essential. To address this, we propose the EFJI module, which improves representation and interaction capabilities by combining global and local features from both modalities. The global branch captures motion across the scene, while the local branch refines fine details, ensuring precise feature fusion for better deblurring.

As shown in Figure 4, the EFJI module consists of three branches: the global, local, and base branches, which facilitate the interaction between RGB and event data. The global branch integrates event information to enhance implicit object localization and motion estimation from a broader perspective for deblurring, while the local branch employs large kernel convolutions to capture fine-grained event details, thereby refining the RGB feature representations. The global branch integrates event information to enhance implicit object localization and motion estimation from a broader perspective for deblurring, while the local branch employs large kernel convolutions to capture fine-grained event details, thereby refining the RGB feature representations. The base branch focuses solely on encoding the RGB features, ensuring a strong foundational representation before interacting with event features. After feeding F_I^l and F_E^l to the three branches, we obtain the respective outputs: F_G^l (global), F_L^l (local), and

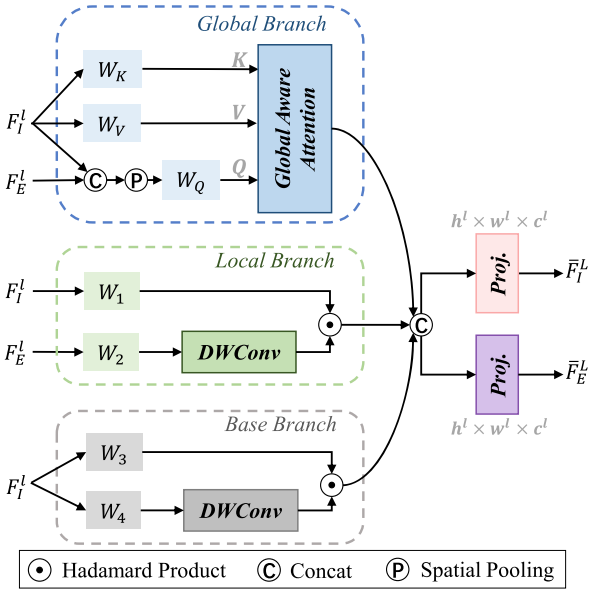

Fig. 4. Detailed structure of the proposed EFJI module. EFJI module enhances cross-modal interaction between event and RGB modalities, featuring three distinct branches: the global, local, and base branches.

F_B^l (base). These features are then fused via concatenation and linear projection to obtain the RGB features $\overline{F}_I^l$ and event features $\overline{F}_E^l$. This process can be denoted as

$$F_G^l = \mathrm{GB}(F_I^l, F_E^l), \tag{5}$$
$$F_L^l = \mathrm{LB}(F_I^l, F_E^l), \tag{6}$$
$$F_B^l = \mathrm{BB}(F_I^l), \tag{7}$$
$$\overline{F}_I^l, \overline{F}_E^l = \mathrm{Projection}([F_G^l, F_L^l, F_B^l]), \tag{8}$$

where $\mathrm{GB}(\cdot,\cdot)$, $\mathrm{LB}(\cdot,\cdot)$, $\mathrm{BB}(\cdot)$, $\mathrm{Projection}(\cdot)$, and $[\cdot,\cdot]$ denote the global branch, the local branch, the base branch, the linear projection operation and the concatenate operation. Specifically, concatenation stacks RGB and event feature maps along the channel dimension to integrate both modalities, followed by a linear projection using a fully connected layer to reduce dimensionality and effectively mix features. Through this design, the RGB and event features interact efficiently, eliminating the need for complex interaction modules while ensuring effective feature fusion. This approach balances computational efficiency and performance, enabling better integrating complementary information from both modalities.

a) Global branch: Our global branch fuses event and RGB features to establish relationships across the entire image. Unlike traditional self-attention mechanisms [92], which suffer from quadratic computation growth as pixels or tokens increase, we down-sample the query Q in the global branch to a fixed size, reducing computational complexity. In our design, Q is formed by concatenating RGB and event features, while K and V are derived solely from RGB features. Given the RGB features F_I^l and event features F_E^l, the process can be formulated as

$$\begin{aligned} Q &= \mathrm{LN}(\mathrm{P}_k([F_I^l, F_E^l])), \\ K &= \mathrm{LN}(F_I^l), \\ V &= \mathrm{LN}(F_I^l), \end{aligned} \tag{9}$$

where $\mathrm{P}_k(\cdot)$ performs adaptive average pooling to size $k \times k$, and $\mathrm{LN}(\cdot)$ is the linear transformation. Based on the resulting $Q \in \mathbb{R}^{k\times k\times c^G}$, $K \in \mathbb{R}^{h\times w\times c^G}$, and $V \in \mathbb{R}^{h\times w\times c^G}$, the global branch can be expressed as

$$F_G^l = \mathrm{Up}(V \cdot \mathrm{Softmax}(Q^\top K/\sqrt{C^d})), \tag{10}$$

where $\mathrm{Up}(\cdot)$ is a bilinear upsampling operation that converts the spatial size from $k \times k$ to $h \times w$. We set $k = 7$ in our experiments.

b) Local branch: We design the local branch to capture finer details, complementing the global branch. Instead of using addition or concatenation like previous methods, we apply depth-wise convolution with a large kernel on event features. The output serves as attention weights to reweight the RGB features via the Hadamard product. This is effective since adjacent pixels with similar event patterns often share a similar motion pattern, allowing event information to be embedded into blurred RGB features for improved deblurring. The process for the local branch is defined as

$$F_L^l = \mathrm{DWConv}_k(\mathrm{LN}(F_E^l)) \odot \mathrm{LN}(F_I^l), \tag{11}$$

where DWConv_k is a depth-wise convolution with kernel size $k \times k$ and $\odot$ is the Hadamard product.

c) Base branch: To retain the diverse appearance cues inherent in RGB images, we additionally introduce a base branch that transforms the RGB features F_I^l into F_B^l, maintaining the same spatial resolution as F_G^l and F_L^l for consistent fusion. The calculation process of F_B^l can be defined as

$$F_B^l = \mathrm{DWConv}_k(\mathrm{LN}(F_I^l)) \odot \mathrm{LN}(F_I^l). \tag{12}$$

C. Event-guided Multi-scale Fusion Attention

RGB images provide rich spatial structures and color information but tend to suffer from significant degradation under fast motion, resulting in blur. In contrast, event data offers high temporal resolution and robustness to motion blur but lacks dense spatial and color cues. Effectively fusing these complementary modalities is critical for accurate motion deblurring. To this end, we propose the EMFA module, which aims to integrate motion-aware event features and spatially rich RGB features adaptively. Motivated by the observation that *different convolution kernel sizes capture content at varying receptive fields*, the EMFA module employs a set of asymmetric convolution branches (*e.g.*, 7×7, 11×11, 21×21) to model multi-scale contextual information. This enables the network to attend to both fine-grained details and broader structures, enhancing the quality of feature fusion. Furthermore, a cross-modal gated branch is introduced to selectively control event information flow, ensuring that *only informative cues are preserved during fusion*.

As shown in Figure 5, the EMFA module consists of four components: a 5×5 convolution layer for extracting initial local features $\overline{F}_I^{l'}$, a multi-scale convolution block to capture local features at different scales, a gated fusion mechanism for integrating global frame-event relationships, and a 1×1 convolution layer to model inter-branch interactions.

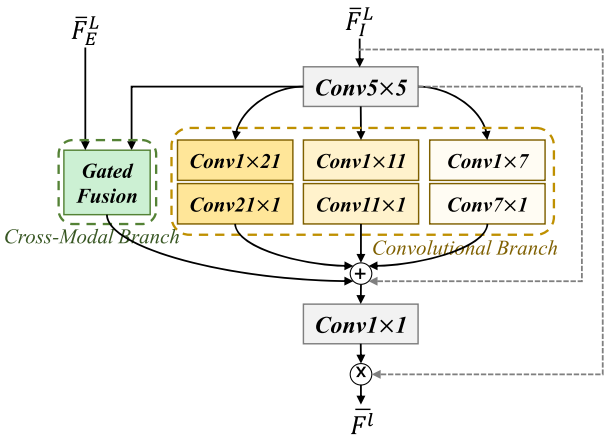

Fig. 5. Detailed structure of the proposed EMFA module. This module captures local visual details and global frame-event cues.

a) Multi-scale convolution activation: We employ three parallel convolutional branches with varying receptive fields to effectively capture diverse local visual cues. Each branch consists of sequential strip-shaped convolutions along vertical and horizontal axes to model spatially rich features with strong inductive bias. The resulting multi-scale activation map Att^{conv} is computed as

$$Att^{conv} = \sum_{t=1}^{3} \mathrm{Conv}_R^t(\mathrm{Conv}_C^t(\overline{F}_I^{l'})), \tag{13}$$

where t indexes the convolution branch, Conv_R^t indicates a $1 \times k_t$ convolution function for horizontal linear features, and Conv_C^t indicates a $k_t \times 1$ convolution function for vertical linear features. The strip-like convolution kernels are designed to efficiently capture fine-grained local structures, enabling detailed spatial representation with reduced computational overhead.

b) Gated fusion operation: We propose a gated fusion mechanism to effectively integrate cross-modal features from both RGB frames and event data. This module employs a series of convolutional layers to dynamically modulate the fusion process, enabling the selective emphasis of modality-specific informative cues. The gating operation is defined as

$$Att^{gate} = f(\overline{F}_I^{l'}, \overline{F}_E^l), \tag{14}$$

where $f(\cdot)$ denotes a learnable fusion function implemented via multiple convolutional layers.

c) Integrated attention: To enable comprehensive cross-modal understanding, we integrate both local and global contextual cues by combining the outputs of the multi-scale convolutional branches and the gated cross-modal branch. A 1×1 convolution is applied to unify these representations and compute integrated attention weights, which are used to reweight the original RGB features $\overline{F}_I^l$. The final integrated cross-modal feature map $\overline{F}^l$ is obtained as

$$\overline{F}^l = \mathrm{Conv}_{1\times1}(Att^{\mathrm{conv}} + Att^{\mathrm{gate}} + \overline{F}_I^{l'}) \odot \overline{F}_I^l, \tag{15}$$

where $\odot$ denotes element-wise multiplication and $\mathrm{Conv}_{1\times1}(\cdot)$ represents a 1×1 convolution layer for branch-wise interaction modeling.

By embedding the proposed EFJI and EMFA modules into a multi-scale U-Net framework, DuInt-Net effectively captures cross-modal dependencies and hierarchical visual representations. This holistic design significantly enhances motion perception and fine-grained detail recovery, resulting in superior deblurring performance.

IV. Experiments

A. Experimental Settings

a) Selected datasets: The proposed DuInt-Net is evaluated on two types of datasets. (1) Synthetic datasets. We conduct experiments on GoPro [48] and DVD [93], two widely adopted benchmarks for both image-based and event-guided motion deblurring. These datasets provide synthetically blurred images, corresponding sharp ground-truth frames, and simulated event streams generated via ESIM [94]. We follow the standard training and testing splits used in prior works to ensure a fair comparison. (2) Real-world dataset. REBlur [35] is a challenging dataset comprising real blurry-sharp image pairs and corresponding event streams, collected under 12 motion types and 3 movement patterns across 36 sequences. It includes a total of 1,469 image pairs, with 486 for training and 983 for testing. To evaluate generalization, DuInt-Net is first trained on GoPro and directly tested on DVD. For real-world adaptation, we fine-tune the model on REBlur, which effectively narrows the domain gap between synthetic and real-world event data [35]. We also directly test DuInt-Net on the FEVD [66] dataset to evaluate its generalization performance in real-world scenarios. FEVD is a real-captured dataset providing 21 sequences with a resolution of 1024×768, including dynamic urban scenes with diverse motion modes such as ego-motion, object motion, and combinations of both. It contains challenging cases with extreme blur, offering a comprehensive benchmark for assessing real-world deblurring effectiveness.

b) Implementation details: DuInt-Net is implemented in PyTorch and trained on a single NVIDIA A800 80GB GPU. The model is trained from scratch using 256×256 cropped patches from the GoPro dataset. Data augmentation techniques include horizontal and vertical flipping, random noise injection, and the simulation of hot pixels in event voxels [95]. Given the ground-truth sharp frame I^{GT} and the predicted deblurred frame $\hat{I}$, we adopt the Charbonnier loss for optimization

$$\mathcal{L} = \sqrt{\|I^{GT} - \hat{I}\|^2 + \varepsilon^2}, \tag{16}$$

where ε is empirically set to 1×10^{-3}. We use the Adam optimizer with an initial learning rate of 4×10^{-4}, decayed via a cosine annealing schedule down to 1×10^{-7}. The network is trained for 300k iterations with a batch size of 64. For the REBlur dataset, we fine-tune the pre-trained model for 600 iterations using the same hardware setup, with a reduced learning rate of 1×10^{-5} while keeping all other settings unchanged. We evaluate all methods using two standard image quality metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [96], and learned perceptual image patch similarity (LPIPS) [97].

B. Quantitative and Qualitative Comparisons

We comprehensively evaluate our proposed DuInt-Net against advanced methods across two major categories: RGB-only deblurring and event-based deblurring. (1) RGB-only deblurring methods. These approaches rely solely on RGB frame information and include both image-based and video-based methods. We compare against MemDeblur [98], MMP-RNN [99], MPRNet [100], MIMO-UNet++ [50], Restormer [51], RNN-MBP [101], NAFNet [102], VRT [56], DFFN [25], and DSTN [54]. These methods primarily capture spatial and temporal information from frames but often struggle in fast-motion or low-light conditions where motion blur severely deteriorates visual quality. (2) Event-based deblurring methods. These methods incorporate event streams to recover sharp frames and are more robust to extreme motion blur. We include comparisons with RED [103], eSL-Net [104], D2Nets [60], DS-Deblur [31], ERDNet [36], EFNet [35], REFID [61], STCNet [34], TRMD [58], DA [105] and FAEVD [66]. While these methods leverage the high temporal resolution of event cameras, many still suffer from suboptimal fusion strategies and limited cross-modal interaction mechanisms.

a) Quantitative results: As illustrated in Table I, DuInt-Net establishes itself as the state-of-the-art method for motion deblurring by achieving the highest PSNR of 37.00 dB and SSIM of 0.9792 on the GoPro dataset. This performance significantly surpasses that of FAEVD (36.70 dB, 0.9780) and STCNet (36.45 dB, 0.9809), indicating that DuInt-Net not only delivers the best quantitative metrics but also excels in preserving the structural integrity of restored images. The superior PSNR and SSIM values suggest that DuInt-Net effectively reduces noise and artifacts while maintaining fine details and edges. The DVD dataset, known for its complex and varied motion blur scenarios, serves as a stringent test of a model's generalization ability. As presented in Table II, DuInt-Net again demonstrates its superiority with a PSNR of 34.25 dB and an SSIM of 0.9708. Compared to the second-best performer, STCNet, which achieves a PSNR of 33.94 dB and an SSIM of 0.9692, these results underscore DuInt-Net's robustness and adaptability across different types of motion blur. The DVD dataset includes a wide range of challenging scenarios, and DuInt-Net's consistently high performance indicates its ability to handle diverse and complex motion blur patterns. To assess real-world applicability, we evaluate DuInt-Net on the REBlur dataset, which consists of real blurred and sharp image pairs. As shown in Table III, DuInt-Net sets a new benchmark with 40.42 dB PSNR and 0.9815 SSIM. While STCNet reaches a slightly higher SSIM (0.9820), DuInt-Net achieves a clear advantage in PSNR and overall perceptual quality. As shown in Table V, we also compare the LPIPS performance of different deblurring methods on the GoPro dataset. DuInt-Net achieves an LPIPS score of 0.0232, which is significantly better than most conventional and learning-based methods such as NAFNet (0.1259) and eSL-Net (0.1598). Although TRMD achieves the lowest LPIPS score (0.0200), DuInt-Net remains highly competitive, demonstrating its effectiveness in preserving perceptual quality while maintaining strong quantitative performance. To further

TABLE I
COMPARISON OF MOTION DEBLURRING METHODS ON THE GOPRO DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	RED	eSL-Net	D2Nets	MemDeblur	MMP-RNN	MPRNet	MIMO-UNet++	Restormer	DS-Deblur	RNN-MBP	NAFNet
PSNR↑	28.98	30.23	31.76	31.76	32.64	32.66	32.68	32.92	33.13	33.32	33.69
SSIM↑	0.8499	0.8703	0.9430	0.9230	0.9359	0.9590	0.9590	0.9610	0.9465	0.9627	0.9670
Method	DFFN	ERDNet	VRT	DSTN	EFNet	REFID	STCNet	TRMD	DA	FAEVD	DuInt-Net
PSNR↑	34.21	34.25	34.81	35.05	35.46	35.91	36.45	36.68	36.07	36.70	**37.00**
SSIM↑	0.9692	0.9534	0.9724	0.9733	0.9720	0.9730	0.9809	0.9380	0.9760	0.9780	**0.9792**

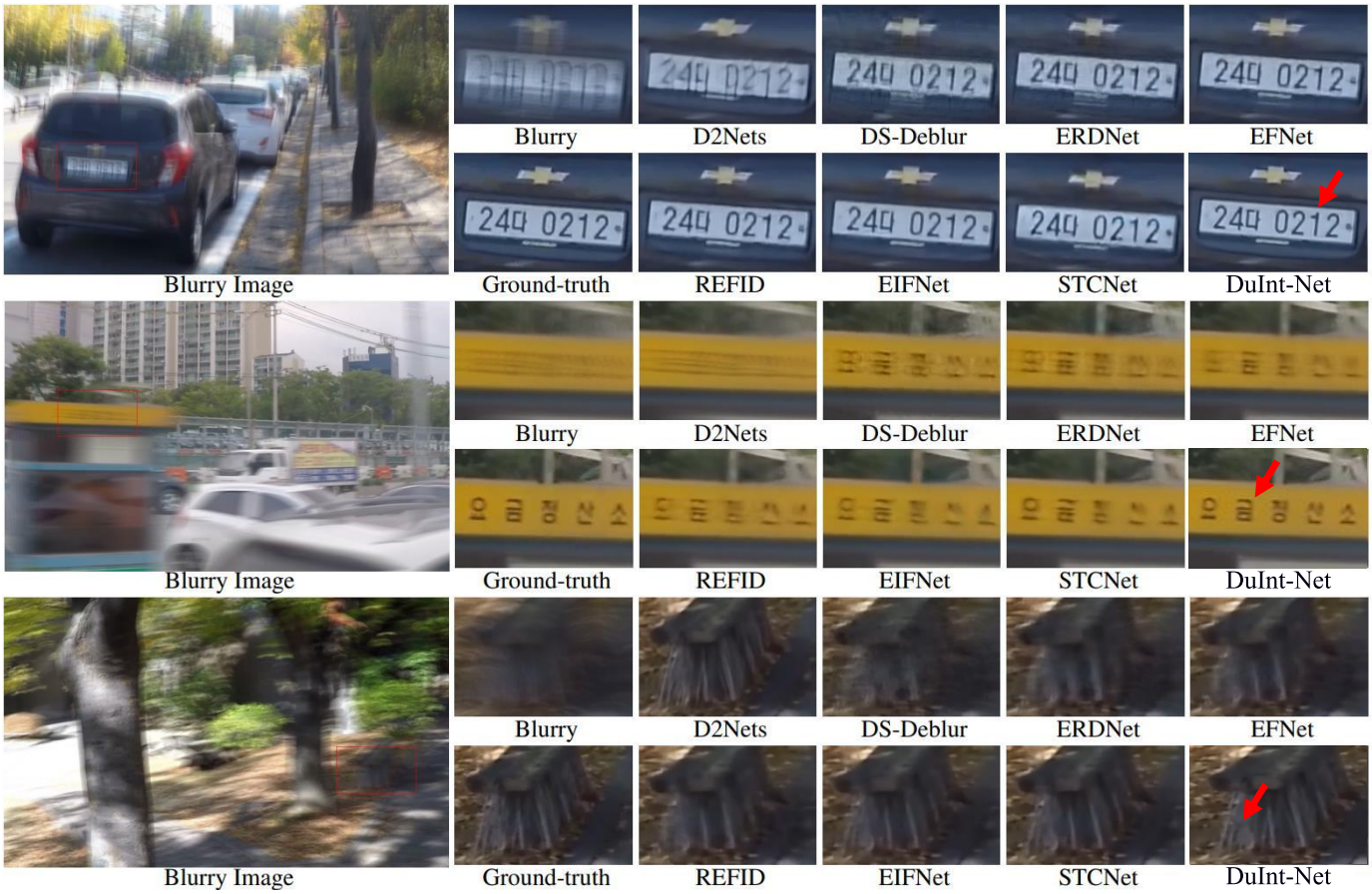

Fig. 6. Compared with advanced RGB-only and event-based motion deblurring methods, DuInt-Net more effectively restores fine textures and structural details. Red arrows indicate regions where DuInt-Net achieves noticeably better visual clarity and contrast.

TABLE II
COMPARISON OF MOTION DEBLURRING METHODS ON THE DVD DATASET. THE BEST RESULTS ARE MARKED IN **BOLD**, AND THE SECOND ONES ARE MARKED WITH UNDERLINES.

Method	D2Nets	MPRNNet	eSL-Net	DS-Deblur	NAFNet	ERDNNet	VRT	EFNet	REFID	STCNet	DuInt-Net
PSNR↑	26.64	27.80	27.50	31.63	27.94	32.29	31.94	32.85	33.15	33.94	**34.25**
SSIM↑	0.8819	0.9091	0.8914	0.9436	0.9126	0.9506	0.9602	0.9571	0.9611	0.9692	**0.9708**

reduce the perceptual gap, future work could integrate VGG-based perceptual losses or lightweight adversarial refinement modules, enabling DuInt-Net to generate texture details that better align with human perceptual preferences while retaining its strong structural reconstruction capability.

b) Computational cost results: We evaluate DuInt-Net's computational efficiency by measuring its parameter count and average inference time on 1280×720 resolution blurry images. The results, summarized in Table IV, reveal that DuInt-Net effectively balances model complexity and deblurring performance. Specifically, DuInt-Net contains only 14.12M param-

TABLE III
COMPARISON OF DIFFERENT DEBLURRING METHODS ON THE REBLUR DATASET. THE BEST RESULTS ARE MARKED IN **BOLD**, AND THE SECOND ONES ARE MARKED WITH UNDERLINES.

Method	D2Nets	eSL-Net	ERDNet	EFNet	REFID	STCNet	DuInt-Net
PSNR↑	35.10	35.50	37.98	38.12	38.34	38.98	**40.42**
SSIM↑	0.9621	0.9563	0.9506	0.9750	0.9752	**0.9820**	0.9815

TABLE IV
COMPLEXITY COMPARISON OF DIFFERENT DEBLURRING METHODS. WE COMPARE THE NUMBER OF PARAMETERS (M), THE AVERAGE RUNTIME (S), AND THE PSNR RESULTS ON THE GOPRO DATASET.

Method	eSL-Net	D2Nets	MemDeblur	MPRNet	MIMO-UNet++	Restormer	DS-Deblur	ERDNet	REFID	STCNet	DuInt-Net
#Params	**0.19**	32.63	6.10	20.10	16.10	26.09	15.60	18.08	15.9	16.25	14.12
Time	**0.015**	1.340	0.911	0.117	0.025	1.155	0.292	0.020	0.072	0.098	0.130
PSNR	30.23	31.76	31.76	32.66	32.68	32.92	33.13	34.25	35.91	36.45	**37.00**

Fig. 7. Visual comparison on the DVD dataset.

TABLE V
COMPARISON OF DIFFERENT DEBLURRING METHODS ON THE GOPRO DATASET. THE BEST LPIPS RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST ARE UNDERLINED.

Methods	NAFNet	eSL-Net	EFNet	TRMD	DuInt-Net
LPIPS↓	0.1259	0.1598	0.0382	0.0200	0.0232

eters, which is notably smaller than many strong baselines, such as D2Nets (32.63M), MPRNet (20.10M), and Restormer (26.09M). Despite its compactness, DuInt-Net consistently achieves the highest PSNR (37.00 dB) on the GoPro dataset, demonstrating that efficient architecture design can yield high-quality restoration without increasing model size. However, DuInt-Net's average runtime of 0.130 seconds is not the fastest among all methods. Lightweight models like eSL-Net (0.015s) and MIMO-UNet++ (0.025s) offer faster inference, though at the expense of significantly lower PSNR. This highlights a critical trade-off between accuracy and speed, where DuInt-Net favors precise restoration, especially under complex motion blur. The relatively slower inference speed of DuInt-Net, when compared to STCNet (0.098s), is primarily attributed to its parallel three-branch structure in the EFJI module. This design introduces synchronization bottlenecks, as the final output must wait for all branches (global, local, base) to finish the computation. Additionally, the network's multi-scale hierarchy, while beneficial for capturing diverse motion blur patterns, further increases computation due to repeated processing across scales. These observations emphasize the importance of architectural efficiency in practical deployment. While DuInt-Net demonstrates strong performance and compactness, its runtime overhead suggests future improvements may focus on optimizing cross-branch interactions and reducing multi-scale redundancy, potentially through early-exit mechanisms or adaptive inference.

c) Qualitative results: We present qualitative comparisons on the GoPro, DVD, REBlur, and FEVD datasets in Figures 6, 7, 8, and 9, respectively. We compare DuInt-Net against a range of state-of-the-art RGB-based and event-based motion deblurring methods. On the GoPro dataset (Figure 6), DuInt-Net effectively recovers fine structures severely degraded in competing methods' outputs. For example, in the first row, our method successfully reconstructs the characters on the license plate, which remain blurry or distorted in the results of other approaches. In the third row, DuInt-Net preserves the complex texture of the tree bark, capturing subtle edge details that others fail to retain. On the DVD dataset (Figure 7), DuInt-Net again demonstrates superior reconstruction accuracy. Compared with STCNet, which produces softened outlines of the clock's needle, DuInt-Net recovers sharper contours and more faithful geometry, reflecting its enhanced capacity to resolve high-frequency motion-induced degradation. For real-world motion blur in the REBlur dataset (Figure 8), our method shows clear advantages in structural clarity. In the first row, the checkered pattern is heavily blurred in the input and only partially recovered by ERDNet, with residual smearing along the edges. In contrast, DuInt-Net precisely restores the square boundaries and textures, producing an output that closely matches the ground-truth image in sharpness and structural integrity. As shown in Figure 9, DuInt-Net produces clearer structural details and sharper edges compared to other methods. Specifically, it restores the contours and boundaries of vehicles more distinctly. It preserves fine textures in building windows, such as frames and inner structures, which are often oversmoothed by approaches like Restormer and EFNet. These results confirm that DuInt-Net achieves stronger perceptual quality by leveraging dual frame-event interactions and multi-scale fusion, particularly in restoring high-frequency and semantically important details across diverse scenes.

C. Ablation Studies

In this section, we conduct experiments on the GoPro dataset to demonstrate the effectiveness of the DuInt-Net.

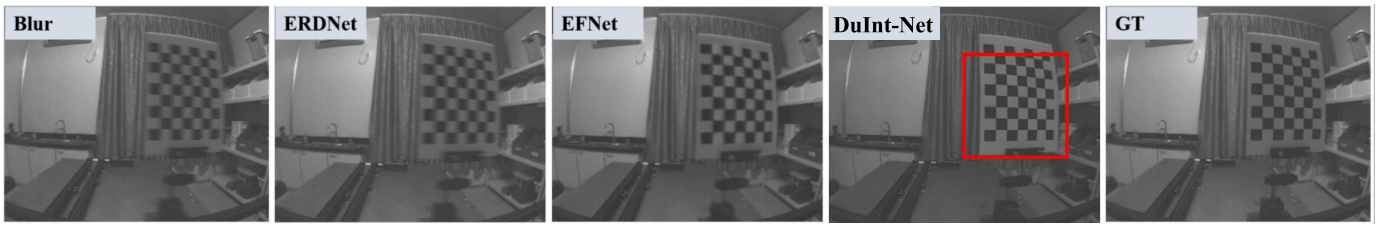

Fig. 8. Visual comparison on the REBlur dataset. Compared to advanced event-based motion deblurring methods, DuInt-Net restores fine texture better. The areas marked in red are particularly notable.

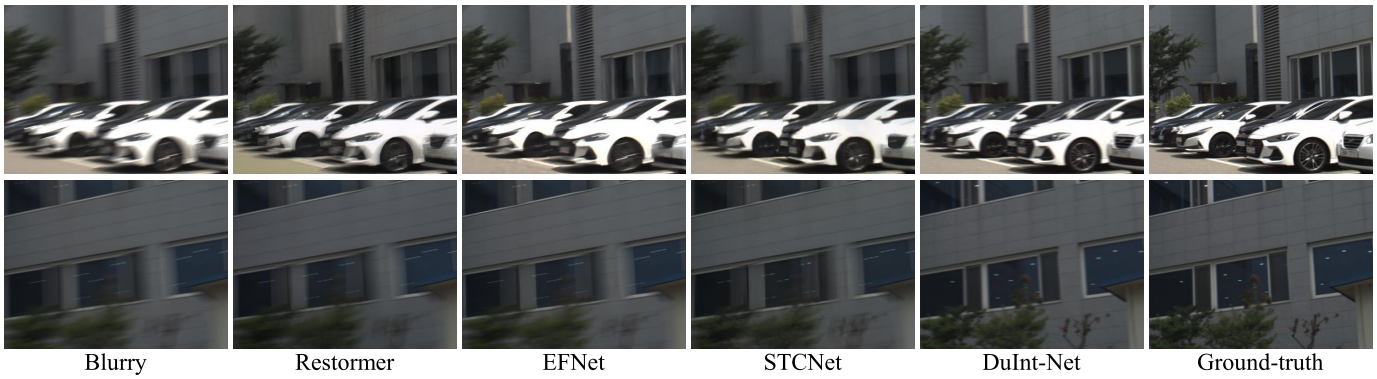

Fig. 9. Visual comparison on the REVD dataset. Compared to advanced motion deblurring methods, DuInt-Net restores fine textures more accurately, preserving details such as vehicle edges and window frames. Please zoom in for better visualization.

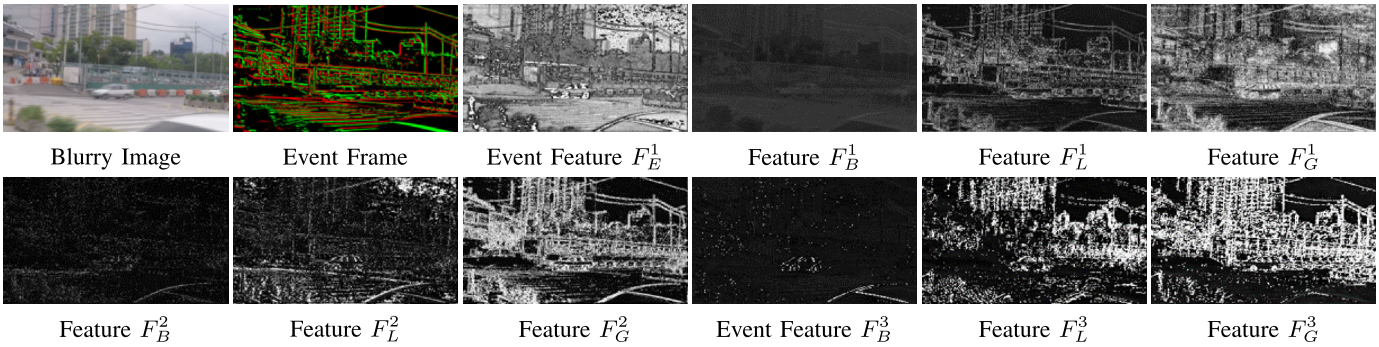

Fig. 10. Visualization of the feature maps from the EFJI module. Black represents zero values, and white indicates maximum values. The base branch extracts general frame features, the local branch emphasizes fine-grained details, and the global branch captures broader context with extensive pixel activations. The feature maps at all three different scales exhibit consistent displays across the branches.

TABLE VI
ANALYSIS ON THE EFJI MODULE AND THE EMFA MODULE.

Method	EFJI	EMFA	PSNR	SSIM
DuInt-Net-Baseline	✗	✗	36.15	0.9748
DuInt-Net-EFJI	✓	✗	36.55	0.9766
DuInt-Net-EMFA	✗	✓	36.49	0.9770
DuInt-Net	✓	✓	37.00	0.9792

TABLE VII
ANALYSIS ON THE EFJI MODULE AND ITS VARIANTS.

Method	GB	LB	BB	PSNR	SSIM
EFJI-baseline	✗	✗	✗	36.49	0.9770
EFJI-G	✓	✗	✗	36.60	0.9778
EFJI-L	✗	✓	✗	36.54	0.9774
EFJI-B	✗	✗	✓	36.58	0.9772
EFJI-GL	✓	✓	✗	36.62	0.9780
EFJI-GB	✓	✗	✓	36.68	0.9782
EFJI-BL	✗	✓	✓	36.71	0.9785
EFJI-GBL	✓	✓	✓	37.00	0.9792

a) Effectiveness of the core components in the DuInt-Net: The EFJI and EMFA modules form the core components of DuInt-Net. To isolate and assess their individual contributions, we introduce the following network variants: (1) DuInt-Net-Baseline: Both EFJI and EMFA modules are removed and replaced with residual blocks of equivalent parameter count. (2) DuInt-Net-EFJI: Only the EFJI module is retained, while the EMFA module is replaced. (3) DuInt-Net-EMFA: Only the EMFA module is retained, while the EFJI module is replaced. The results, presented in Table VI, demonstrate the effectiveness of both components. The baseline model

Fig. 11. Visual comparison on the EFJI module and its variants. From left to right: EFJI-G, EFJI-L, EFJI-B, and EFJI-GBL.

TABLE VIII
ANALYSIS ON THE EMFA MODULE AND ITS VARIANTS.

Method	Cross-modal	Convolutional			PSNR	SSIM
		7	11	21		
EMFA-baseline	✗	✗	✗	✗	36.55	0.9766
EMFA-Cr	✓	✗	✗	✗	36.63	0.9770
EMFA-Co	✗	✓	✓	✓	36.68	0.9773
EMFA-Cr7	✓	✓	✗	✗	36.70	0.9775
EMFA-Cr11	✓	✓	✓	✗	36.75	0.9778
EMFA	✓	✓	✓	✓	37.00	0.9792

TABLE IX
PERFORMANCE COMPARISON USING DIFFERENT CONVOLUTION KERNEL SIZE COMBINATIONS.

Kernel Sizes	1+3+5	7+11+21	7+21+35
PSNR / SSIM	36.59 / 0.9762	37.00 / 0.9792	37.03 / 0.9798
Kernel Sizes	7+31+55	11+21+31	31+51+71
PSNR / SSIM	36.88 / 0.9787	37.07 / 0.9802	36.74 / 0.9775

performs the worst, yielding a PSNR of 36.15 dB and an SSIM of 0.9748. Adding the EFJI module alone improves the PSNR by 0.40 dB and SSIM by 0.0018, while the EMFA module alone results in a 0.34 dB PSNR gain and 0.0022 SSIM increase over the baseline. When both modules are integrated into the full DuInt-Net, the model achieves the highest performance with a PSNR of 37.00 dB and SSIM of 0.9792. This represents a substantial improvement of 0.85 dB in PSNR and 0.0044 in SSIM over the baseline. These results clearly highlight the complementary benefits of the EFJI and EMFA modules, demonstrating that both are essential for maximizing deblurring quality.

b) A close look at the EFJI module: The EFJI module is specifically designed to enhance both interaction and representational capabilities between the RGB and event modalities. Table VII presents an ablation study evaluating various configurations of the EFJI module, including individual and combined contributions from the global branch (GB), local branch (LB), and base branch (BB). Removing all branches (EFJI-Baseline) yields the lowest performance, with a PSNR of 36.49 dB and SSIM of 0.9770. Introducing only the global branch (EFJI-G) leads to a PSNR gain of 0.11 dB over the baseline, indicating that global contextual modeling benefits motion understanding. EFJI-L and EFJI-B provide PSNR improvements of 0.05 dB and 0.09 dB, respectively, suggesting that both local detail enhancement and foundational RGB encoding also contribute positively. Notably, combining branches results in more substantial gains. For example, EFJI-GL (GB + LB) achieves 36.62 dB PSNR and 0.9780 SSIM, EFJI-GB (GB + BB) improves further to 36.68 dB and 0.9782, and EFJI-BL (BB + LB) performs even better, reaching 36.71 dB and 0.9785. The full configuration, EFJI-GBL, which integrates all three branches, delivers the highest performance, with 37.00 dB PSNR and 0.9792 SSIM. This demonstrates a cumulative effect: combining global, local, and base information leads to a well-rounded and discriminative feature representation.

We further visualize the feature maps extracted by the EFJI module in Figure 10 to gain deeper insight into its internal representation behavior. (1) The base branch (F_B^l) captures the coarse structural layout of the scene, providing a strong backbone of general RGB features. The local branch (F_L^l) emphasizes edge-aware, fine-grained textures, particularly in areas of high-frequency details such as object boundaries and textures. In contrast, the global branch (F_G^l) produces more spatially extensive activation patterns, indicating its effectiveness in modeling large contextual dependencies and enhancing global motion cues. (2) Across different spatial scales ($l = 1, 2, 3$), the visual patterns of each branch remain consistent. This suggests the EFJI module maintains coherent semantics across resolutions, which benefits the multi-scale decoder in reconstructing sharp structures from blurry inputs. (3) Moreover, the global branch consistently exhibits stronger activations in regions with complex motion patterns (*e.g.*, moving vehicles and scene boundaries), while the local branch precisely enhances subtle textural variations. These complementary activations highlight the advantage of using parallel branches for hierarchical representation and fusion.

Figure 11 visually compares the EFJI configurations. EFJI-G fails to recover fine textures, leading to overly smoothed results. EFJI-L struggles with long-range dependencies, resulting in poor global coherence. EFJI-B introduces artifacts due to the lack of cross-modal guidance. Only the full EFJI-GBL setup effectively reconstructs both global structures and fine-grained details, confirming the necessity of all three components for optimal performance.

c) A close look at the EMFA module: The EMFA module is designed to efficiently fuse information from two distinct modalities, leveraging our proposed attention mechanism for effective cross-modal interaction. To assess the effectiveness of the EMFA module and its components, we introduced several variants, as shown in Table VIII. The analysis of the EMFA module and its variants shows a clear trend of performance improvement through cross-modal and convolutional enhancements. The baseline EMFA variant scores a PSNR of 36.55 and an SSIM of 0.9766. Subsequent variants demonstrate incremental gains, with EMFA-Cr reaching a PSNR of 36.63, and EMFA-Co and EMFA-Cr7 achieving 36.68 and 36.70, respectively. Combining cross-modal and convolutional features in EMFA-Cr11 results in a PSNR of 36.75. Ultimately, the fully integrated EMFA variant achieves the highest performance, with a PSNR of 37.00 and an SSIM of 0.9792. This progression underscores the effectiveness of

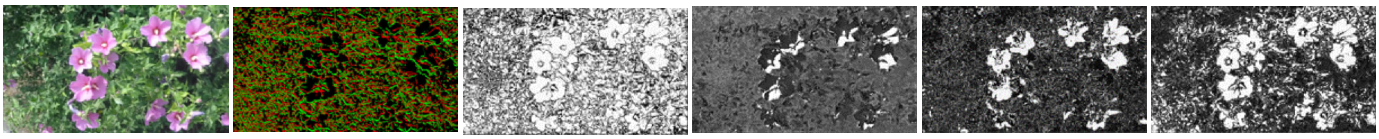

Fig. 12. Visualization of the feature maps from the EMFA module. Black represents zero values, and white indicates maximum values. Feature Att^{gate} focuses on capturing high-frequency details and edge textures. The convolution branches with different kernel sizes attend to various content ranges, with smaller kernels focusing on finer local details and larger kernels capturing broader contextual information.

TABLE X
ANALYSIS ON THE NUMBER OF LEVEL IN DUINT-NET.

Method	Level 1	Level 2	Level 3	Level 4	Level 5
PSNR	36.61	36.88	37.00	37.07	37.08
SSIM	0.9770	0.9785	0.9792	0.9795	0.9795

TABLE XI
COMPARISON OF DUINT-NET AND ITS DISTILLED VERSION.

Method	#Params	Time	PSNR
DuInt-Net	14.12	0.130	37.00
DuInt-Net (Distilled)	8.59	0.098	36.62

DuInt-Net in enhancing motion deblurring through advanced feature fusion techniques.

To investigate the impact of kernel size in our multi-scale convolutional design, we retrain DuInt-Net variants with different kernel combinations and report the results in Table IX. Large kernels clearly benefit motion deblurring, as configurations like $11 + 21 + 31$ and $7 + 21 + 35$ yield superior performance (37.07/0.9802 and 37.03/0.9798 in PSNR/SSIM, respectively). However, huge kernels (*e.g.*, $31 + 51 + 71$) degrade performance (36.74/0.9775), likely due to oversmoothing and diminished local detail modeling, along with increased computational overhead and potential overfitting on limited data. These observations underscore the need to balance global context and local texture representation. Future work could explore adaptive kernel strategies, such as deformable convolutions, to dynamically adjust receptive fields based on input structure, offering a more efficient trade-off between accuracy and complexity.

As illustrated in Figure 12, we visualize the feature maps generated by the EMFA module to better understand its behavior. We observe two notable insights: (1) The gated fusion operation Att^{gate} exhibits activation patterns that resemble the event modality, emphasizing high-frequency textures and edge structures. This confirms its effectiveness in selectively integrating event-driven cues that are crucial for recovering motion-induced details. (2) The convolution branches with different kernel sizes (7, 11, and 21) display complementary activation patterns. Specifically, smaller kernels (*e.g.*, $k = 7$) focus on fine local structures such as floral textures and edge contours, while larger kernels (*e.g.*, $k = 21$) respond to broader contextual information, capturing coarse semantic layouts. This diversity of spatial sensitivity allows the EMFA module to extract rich features across multiple receptive fields, facilitating robust and detail-preserving fusion for motion deblurring.

d) Effectiveness of the multi-scale structure: Table X investigates the impact of the number of hierarchical levels in DuInt-Net on deblurring performance. As levels increase from 1 to 3, PSNR and SSIM steadily improve—from 36.61 dB and 0.9770 (Level 1) to 37.00 dB and 0.9792 (Level 3). This indicates that deeper hierarchical structures allow for more effective feature aggregation and cross-modal interaction. Further increasing the level count to 4 and 5 yields marginal gains, with PSNR improving slightly to 37.07 and 37.08 dB and SSIM stabilizing at 0.9795. These results suggest diminishing returns beyond three levels, implying that Level 3 offers the best trade-off between performance and computational efficiency. Hence, DuInt-Net with three levels balances model complexity and accuracy.

e) Comparison between DuInt-Net and its distilled version: To construct the distilled student model, we simplify the frame-related branches of DuInt-Net by retaining only a single convolutional layer followed by a ResASPP block at each scale, preserving essential spatial context modeling while significantly reducing parameters and computation. The event encoder remains unchanged to maintain its crucial capability in capturing fine-grained motion dynamics for accurate deblurring under large motion or severe blur conditions. For distillation, we adopt a two-level strategy. At the feature level, we apply an ℓ_1 loss between the intermediate features of the teacher and student models (excluding event-specific blocks) to encourage representational alignment. At the output level, we minimize the Charbonnier loss between the restored frames of both models to enhance perceptual fidelity. As is shown in Table XI, this distilled variant achieves a favorable trade-off, significantly reducing the model size and inference time while maintaining competitive deblurring performance.

f) Impact of different event representations and feature extraction methods: We follow the widely adopted event representation method, consistent with [35], to ensure a fair comparison with prior works. To further investigate the impact of event representation, we also experiment with the method proposed in [106] and re-train DuInt-Net using this setting. The results show comparable performance to our default representation (PSNR 36.93 dB, SSIM 0.9788), suggesting that DuInt-Net is not highly sensitive to the specific event representation. Additionally, we explore the use of a spiking neural network (SNN) from [107] for event feature extraction.

Fig. 13. Failure cases. The first and third images are our results, while the second and fourth images represent the ground truth. Please zoom in for better visualization.

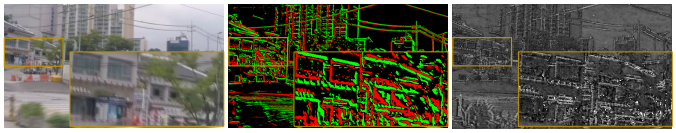

Fig. 14. Failure case analysis. From left to right: DuInt-Net reconstructed output, visualization of the input event stream, and the extracted event feature map. DuInt-Net struggles to recover extremely fine details such as small text regions or thin structures because event streams cannot record precise spatial details, and the corresponding event features lack sufficient representational capability for such intricate information.

The SNN-based model achieves a slightly improved performance, with a PSNR of 37.05 dB and an SSIM of 0.9800, compared to our original design. These results indicate the potential of learnable event encoders.

D. Discussions

While DuInt-Net achieves strong performance across benchmarks, several limitations remain. For challenging scenes, it struggles with fine-grained reconstruction (e.g., blurred text) and large object motion, as shown in Figure 13. Figure 14 further illustrates cases where DuInt-Net fails to recover extremely fine details such as small text or thin structures. This is because event cameras primarily encode intensity changes rather than detailed textures, making them unable to capture precise spatial details. When RGB inputs also lack these details due to severe blur, events do not provide sufficient complementary cues. Moreover, event cameras exhibit spatial sparsity, primarily responding to edges or significant brightness changes, resulting in insufficient activation in textureless areas. Even when triggered, events predominantly reflect motion edges rather than static fine textures, resulting in features that convey motion dynamics and edges but lack dense pixel-level detail. These limitations suggest that future improvements may require integrating additional priors or generative perceptual modules to hallucinate fine textures beyond what is available in RGB frames and event streams.

In addition, DuInt-Net can be further improved in several aspects. First, it assumes well-aligned and clean frame-event inputs, but in real-world scenarios, event misalignment or polarity noise can impair fusion quality, especially under fast motion or jitter. Incorporating noise-robust alignment modules or confidence-aware fusion strategies could mitigate this issue. Second, although large-kernel convolutions enhance global context modeling, they introduce considerable memory and computational overhead, resulting in slow inference speed. As shown in our complexity analysis, the triple-branch EFJI structure increases inference time, limiting its applicability in real-time or resource-constrained environments. Lightweight alternatives such as re-parameterized convolutions [108] or neural architecture search-based compression [109] may provide more efficient solutions. Third, DuInt-Net relies on single-frame RGB input, limiting its temporal modeling capabilities; in cases of dynamic lighting or large object displacements, single-frame information may be insufficient. Extending the framework to multi-frame inputs or learning temporal priors could improve robustness. Fourth, the model lacks explicit handling of illumination changes and occlusions, which can result in degraded performance in scenes with flickering lights or occluding objects. Future work may explore semantic priors or attention-based reasoning to address these challenges.

V. Conclusion

In this paper, we introduce DuInt-Net, a multi-scale neural network designed for motion deblurring by leveraging event cameras to enhance event-frame interaction and adaptively capture rich spatiotemporal features. Our proposed EFJI module facilitates the effective fusion of event and frame information, significantly improving motion understanding and fine-grained detail restoration. The EMFA module also integrates local and global contextual cues, enabling more robust and precise image reconstruction. Extensive experiments demonstrate that DuInt-Net achieves superior performance across various benchmark datasets.

References

[1] Z. Xiao, J. Bai, Z. Lu, and Z. Xiong, "A dive into sam prior in image restoration," *arXiv preprint arXiv:2305.13620*, 2023.
[2] R. Xu, Z. Xiao, J. Huang, Y. Zhang, and Z. Xiong, "Edpn: Enhanced deep pyramid network for blurry image restoration," in *CVPRW*, 2021.
[3] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *ICCV*, 2023.
[4] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *ICCV*, 2023.
[5] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, "Visevent: Reliable object tracking via collaboration of frame and event flows," *IEEE Transactions on Cybernetics*, 2023.
[6] X. Wang, S. Wang, C. Tang, L. Zhu, B. Jiang, Y. Tian, and J. Tang, "Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline," in *CVPR*, 2024.
[7] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.
[8] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103116, 2021.
[9] N. Haering, P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5, pp. 279–290, 2008.
[10] Z. Li, J. Liao, C. Tang, H. Zhang, Y. Li, Y. Bian, X. Sheng, X. Feng, Y. Li, C. Gao *et al.*, "Ustc-td: A test dataset and benchmark for image and video coding in 2020s," *IEEE Transactions on Multimedia*, 2025.
[11] Z. Li, Y. Li, C. Tang, L. Li, D. Liu, and F. Wu, "Uniformly accelerated motion model for inter prediction," in *VCIP*, 2024.
[12] Z. Li, Z. Yuan, L. Li, D. Liu, X. Tang, and F. Wu, "Object segmentation-assisted inter prediction for versatile video coding," *IEEE Transactions on Broadcasting*, 2024.
[13] Z. Li, J. Li, Y. Li, L. Li, D. Liu, and F. Wu, "In-loop filtering via trained look-up tables," in *VCIP*, 2024.
[14] C. Tang, Z. Li, Y. Bian, L. Li, and D. Liu, "Neural video compression with context modulation," in *CVPR*, 2025.
[15] N. Weiner and I. Extrapolation, "Smoothing of stationary time series: With engineering applications," 1949.

[16] W. H. Richardson, "Bayesian-based iterative method of image restoration," *JoSA*, vol. 62, no. 1, pp. 55–59, 1972.

[17] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," in *NeurIPS*, 2009.

[18] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *CVPR*, 2014.

[19] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *ECCV*, 2010.

[20] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *ICCV*, 2011.

[21] J. Dong, J. Pan, D. Sun, Z. Su, and M.-H. Yang, "Learning data terms for non-blind deblurring," in *ECCV*, 2018.

[22] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *CVPR*, 2019.

[23] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *CVPR*, 2018.

[24] M. Cao, Y. Fan, Y. Zhang, J. Wang, and Y. Yang, "Vdtr: Video deblurring with transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 1, pp. 160–171, 2022.

[25] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *CVPR*, 2023.

[26] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *ECCV*, 2022.

[27] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *CVPR*, 2022.

[28] H. Gao and D. Dang, "Aggregating local and global features via selective state spaces model for efficient image deblurring," *arXiv preprint arXiv:2403.20106*, 2024.

[29] L. Kong, J. Dong, M.-H. Yang, and J. Pan, "Efficient visual state space model for image deblurring," *arXiv preprint arXiv:2405.14343*, 2024.

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

[31] W. Yang, J. Wu, J. Ma, L. Li, W. Dong, and G. Shi, "Learning for motion deblurring with hybrid frames and events," in *ACM MM*, 2022.

[32] W. Yang, J. Wu, L. Li, W. Dong, and G. Shi, "Event-based motion deblurring with modality-aware decomposition and recomposition," in *ACM MM*, 2023.

[33] Z. Liu, J. Wu, G. Shi, W. Yang, W. Dong, and Q. Zhao, "Motion-oriented hybrid spiking neural networks for event-based motion deblurring," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[34] W. Yang, J. Wu, J. Ma, L. Li, and G. Shi, "Motion deblurring via spatial-temporal collaboration of frames and events," in *AAAI*, vol. 38, no. 7, 2024, pp. 6531–6539.

[35] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *ECCV*, 2022.

[36] H. Chen, M. Teng, B. Shi, Y. Wang, and T. Huang, "A residual learning approach to deblur and generate high frame rate video with an event camera," *IEEE Transactions on Multimedia*, vol. 25, pp. 5826–5839, 2022.

[37] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, "Learning event-based motion deblurring," in *CVPR*, 2020.

[38] S. Lin, J. Zhang, J. Pan, Z. Jiang, D. Zou, Y. Wang, J. Chen, and J. Ren, "Learning event-driven video deblurring and interpolation," in *ECCV*, 2020.

[39] M. Ben-Ezra and S. K. Nayar, "Motion deblurring using hybrid imaging," in *CVPR*, 2003.

[40] S. K. Nayar and M. Ben-Ezra, "Motion-based motion deblurring," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 6, pp. 689–698, 2004.

[41] S. Cho and S. Lee, "Fast motion deblurring," in *ACM SIGGRAPH Asia 2009 papers*, 2009, pp. 1–8.

[42] S. Cho, J. Wang, and S. Lee, "Video deblurring for hand-held cameras using patch-based synthesis," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1–9, 2012.

[43] T. Hyun Kim and K. Mu Lee, "Generalized video deblurring for dynamic scenes," in *CVPR*, 2015.

[44] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1439–1451, 2015.

[45] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *CVPR*, 2015.

[46] A. Chakrabarti, "A neural approach to blind motion deblurring," in *ECCV*, 2016.

[47] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *CVPR*, 2017.

[48] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017.

[49] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *CVPR*, 2019.

[50] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *CVPR*, 2021.

[51] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*, 2022.

[52] Z. Zhong, Y. Gao, Y. Zheng, and B. Zheng, "Efficient spatio-temporal recurrent neural network for video deblurring," in *ECCV*, 2020.

[53] S. Nah, S. Son, and K. M. Lee, "Recurrent neural networks with intra-frame iterations for video deblurring," in *CVPR*, 2019.

[54] J. Pan, B. Xu, J. Dong, J. Ge, and J. Tang, "Deep discriminative spatial and temporal network for efficient video deblurring," in *CVPR*, 2023.

[55] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *CVPRW*, 2019.

[56] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *IEEE Transactions on Image Processing*, 2024.

[57] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *ACCV*, 2018.

[58] K. Chen and L. Yu, "Motion deblur by learning residual from events," *IEEE Transactions on Multimedia*, vol. 26, pp. 6632–6647, 2024.

[59] K. Chen, S. Chen, J. Zhang, B. Zhang, Y. Zheng, T. Huang, and Z. Yu, "Spikereveal: Unlocking temporal sequences from real blurry inputs with spike streams," *NeurIPS*, 2024.

[60] W. Shang, D. Ren, D. Zou, J. S. Ren, P. Luo, and W. Zuo, "Bringing events into video deblurring with non-consecutively blurry frames," in *CVPR*, 2021.

[61] L. Sun, C. Sakaridis, J. Liang, P. Sun, J. Cao, K. Zhang, Q. Jiang, K. Wang, and L. Van Gool, "Event-based frame interpolation with ad-hoc deblurring," in *CVPR*, 2023.

[62] X. Zhang and L. Yu, "Unifying motion deblurring and frame interpolation with events," in *CVPR*, 2022.

[63] T. Kim, J. Lee, L. Wang, and K.-J. Yoon, "Event-guided deblurring of unknown exposure time videos," in *ECCV*, 2022.

[64] Y. Shen, S. Li, and K. Song, "Restoring real-world degraded events improves deblurring quality," in *ACM MM*, 2024.

[65] H. Cho, Y. Jeong, T. Kim, and K.-J. Yoon, "Non-coaxial event-guided motion deblurring with spatial alignment," in *ICCV*, 2023.

[66] T. Kim, H. Cho, and K.-J. Yoon, "Frequency-aware event-based video deblurring for real-world motion blur," in *CVPR*, 2024.

[67] ——, "Cross-modal temporal alignment for event-guided video deblurring," *arXiv preprint arXiv:2408.14930*, 2024.

[68] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens: Event-based video frame interpolation," in *CVPR*, 2021.

[69] G. Paikin, Y. Ater, R. Shaul, and E. Soloveichik, "Efi-net: Video frame interpolation from fusion of events and frames," in *CVPR*, 2021.

[70] Z. Xiao, W. Weng, Y. Zhang, and Z. Xiong, "Eva2: Event-assisted video frame interpolation via cross-modal alignment and aggregation," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 1145–1158, 2022.

[71] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. Scaramuzza, "Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion," in *CVPR*, 2022.

[72] T. Kim, Y. Chae, H.-K. Jang, and K.-J. Yoon, "Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields," in *CVPR*, 2023.

[73] H. Cho, T. Kim, Y. Jeong, and K.-J. Yoon, "Tta-evf: Test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation," in *CVPR*, 2024.

[74] X. Zhou, P. Duan, Y. Ma, and B. Shi, "Evunroll: Neuromorphic events based rolling shutter image correction," in *CVPR*, 2022.

[75] J. Erbach, S. Tulyakov, P. Vitoria, A. Bochicchio, and Y. Li, "Evshutter: Transforming events for unconstrained rolling shutter correction," in *CVPR*, 2023.

[76] Y. Lu, G. Liang, and L. Wang, "Self-supervised learning of event-guided video frame interpolation for rolling shutter frames," *arXiv preprint arXiv:2306.15507*, 2023.
[77] Y. Wang, X. Zhang, M. Lin, L. Yu, B. Shi, W. Yang, and G.-S. Xia, "Self-supervised scene dynamic recovery from rolling shutter images and events," *arXiv preprint arXiv:2304.06930*, 2023.
[78] H. Cho and K.-J. Yoon, "Event-image fusion stereo using cross-modality feature propagation," in *AAAI*, vol. 36, no. 1, 2022, pp. 454–462.
[79] H. Cho, J. Cho, and K.-J. Yoon, "Learning adaptive dense event stereo from the image domain," in *CVPR*, 2023.
[80] H. Cho, J.-Y. Kang, and K.-J. Yoon, "Temporal event stereo via joint learning with stereoscopic flow," in *ECCV*, 2024.
[81] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 1964–1980, 2019.
[82] Y. Zou, Y. Zheng, T. Takatani, and Y. Fu, "Learning to reconstruct high speed and high dynamic range videos from events," in *CVPR*, 2021.
[83] N. Messikommer, S. Georgoulis, D. Gehrig, S. Tulyakov, J. Erbach, A. Bochicchio, Y. Li, and D. Scaramuzza, "Multi-bracket high dynamic range imaging with event cameras," in *CVPR*, 2022.
[84] Y. Yang, J. Han, J. Liang, I. Sato, and B. Shi, "Learning event guided high dynamic range video reconstruction," in *CVPR*, 2023.
[85] T. Kim, J. Jeong, H. Cho, Y. Jeong, and K.-J. Yoon, "Towards real-world event-guided low-light video enhancement and deblurring," *arXiv preprint arXiv:2408.14916*, 2024.
[86] Y. Jiang, Y. Wang, S. Li, Y. Zhang, M. Zhao, and Y. Gao, "Event-based low-illumination image enhancement," *IEEE Transactions on Multimedia*, 2023.
[87] Z. Xiao, D. Kai, Y. Zhang, Z.-J. Zha, X. Sun, and Z. Xiong, "Event-adapted video super-resolution," in *ECCV*, 2024.
[88] Z. Xiao, D. Kai, Y. Zhang, X. Sun, and Z. Xiong, "Asymmetric event-guided video super-resolution," in *ACM MM*, 2024.
[89] D. Kai, Y. Zhang, J. Wang, Z. Xiao, Z. Xiong, and X. Sun, "Event-enhanced blurry video super-resolution," in *AAAI*, vol. 39, no. 4, 2025, pp. 4175–4183.
[90] Y. Wang, J. Yang, L. Wang, X. Ying, T. Wu, W. An, and Y. Guo, "Light field image super-resolution using deformable convolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1057–1071, 2020.
[91] Z. Xiao, Z. Li, and W. Jia, "Occlusion-embedded hybrid transformer for light field super-resolution," in *AAAI*, vol. 39, no. 8, 2025, pp. 8700–8708.
[92] A. Vaswani, "Attention is all you need," *NeurIPS*, 2017.
[93] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *CVPR*, 2017.
[94] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: an open event camera simulator," *CoRL*, 2018.
[95] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.
[96] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[97] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.
[98] B. Ji and A. Yao, "Multi-scale memory-based video deblurring," in *CVPR*, 2022.
[99] Y. Wang, Y. Lu, Y. Gao, L. Wang, Z. Zhong, Y. Zheng, and A. Yamashita, "Efficient video deblurring guided by motion magnitude," in *ECCV*, 2022.
[100] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *CVPR*, 2021.
[101] C. Zhu, H. Dong, J. Pan, B. Liang, Y. Huang, L. Fu, and F. Wang, "Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring," in *AAAI*, vol. 36, no. 3, 2022, pp. 3598–3607.
[102] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *ECCV*, 2022.
[103] F. Xu, L. Yu, B. Wang, W. Yang, G.-S. Xia, X. Jia, Z. Qiao, and J. Liu, "Motion deblurring with real events," in *ICCV*, 2021.
[104] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event enhanced high-quality image recovery," in *ECCV*, 2020.
[105] Z. Sun, X. Fu, L. Huang, A. Liu, and Z.-J. Zha, "Motion aware event representation-driven image deblurring," in *ECCV*, 2024.
[106] Q. Qu, X. Chen, Y. Y. Chung, and Y. Shen, "Evrepsl: Event-stream representation via self-supervised learning for event-based vision," *IEEE Transactions on Image Processing*, 2024.
[107] C. Cao, X. Fu, Y. Zhu, Z. Sun, and Z.-J. Zha, "Event-driven video restoration with spiking-convolutional architecture," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[108] M. Hu, J. Feng, J. Hua, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Online convolutional re-parameterization," in *CVPR*, 2022.
[109] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and S. Wang, "Dual-level cross-modality neural architecture search for guided image super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.