

# CompEvent: Complex-valued Event-RGB Fusion for Low-light Video Enhancement and Deblurring

Mingchen Zhong<sup>1</sup>, Xin Lu<sup>1</sup>, Dong Li<sup>1,†</sup>  
Senyan Xu<sup>1</sup>, Ruixuan Jiang<sup>1</sup>, Xueyang Fu<sup>1</sup>, Baocai Yin<sup>2,†</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>iFlytek Research, iFlytek Co., Ltd.

{zmcsa24010071, luxion, dongli6, syxu, sa25010021}@mail.ustc.edu.cn, xyfu@ustc.edu.cn, bcyin@iflytek.com

## Abstract

Low-light video deblurring poses significant challenges in applications like nighttime surveillance and autonomous driving due to dim lighting and long exposures. While event cameras offer potential solutions with superior low-light sensitivity and high temporal resolution, existing fusion methods typically employ staged strategies, limiting their effectiveness against combined low-light and motion blur degradations. To overcome this, we propose CompEvent, a complex neural network framework enabling holistic full-process fusion of event data and RGB frames for enhanced joint restoration. CompEvent features two core components: 1) Complex Temporal Alignment GRU, which utilizes complex-valued convolutions and processes video and event streams iteratively via GRU to achieve temporal alignment and continuous fusion; and 2) Complex Space-Frequency Learning module, which performs unified complex-valued signal processing in both spatial and frequency domains, facilitating deep fusion through spatial structures and system-level characteristics. By leveraging the holistic representation capability of complex-valued neural networks, CompEvent achieves full-process spatiotemporal fusion, maximizes complementary learning between modalities, and significantly strengthens low-light video deblurring capability. Extensive experiments demonstrate that CompEvent outperforms SOTA methods in addressing this challenging task.

**Code** — <https://github.com/YuXie1/CompEvent>

## Introduction

In applications such as nighttime surveillance and autonomous driving, video capture in low-light environments inevitably requires extended exposure times, often suffering from the dual degradations of insufficient brightness and motion blur, leading to a sharp decline in video quality (Kim et al. 2024). These two degradations are tightly coupled: the long exposure required to increase brightness actually exacerbates the blur of moving objects and obliterates significant edge and texture details. This makes the joint task of video enhancement and deblurring a highly ill-posed problem.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>†</sup> Corresponding author.

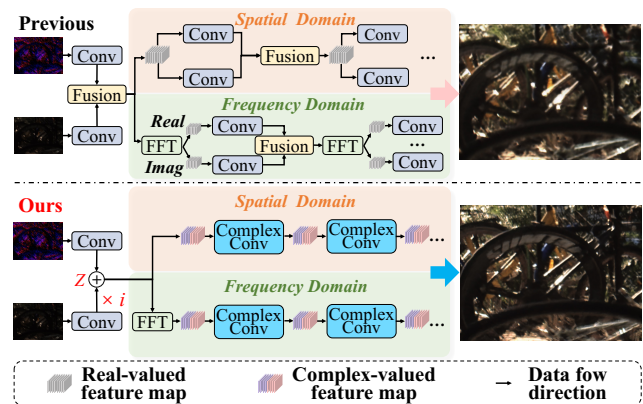


Figure 1: Comparison of our method with previous methods. Previous methods perform fusion in a staged manner and split complex features into real-valued components before convolution in the frequency domain. Our method uses complex representations for both modalities, enabling full-process fusion by interacting features during processing. Moreover, our method directly applies complex convolutions to frequency features without separating them.

Traditional video restoration methods primarily rely on information within the frame itself and face a fundamental information bottleneck when dealing with this compound degradation. Early model-driven approaches (Horn and Schunck 1981; Lucas and Kanade 1981) and more recent deep learning-driven approaches, including convolutional neural networks (CNNs) (Nah, Hyun Kim, and Mu Lee 2017) and Transformers (Vaswani et al. 2017; Wang et al. 2022; Zamir et al. 2022), perform poorly in low-light scenarios. This is because, in low-light conditions, the extended exposure time required to compensate for brightness exacerbates motion blur and significantly loses edge and texture details. Compared to deblurring under normal exposure, models face a more challenging task in estimating the blurring process in low-light conditions. Simply combining low-light enhancement and deblurring tasks in tandem often leads to suboptimal results due to the accumulation and amplification of errors (e.g., noise is enhanced and blur is solidified) (Zhou, Li, and Change Loy 2022).

Event cameras, with their unique advantages of high tem-

poral resolution (HTR) and high dynamic range (HDR), offer new possibilities for addressing this problem (Ruedi 1996; Lichtsteiner and Delbruck 2005) (Lichtsteiner 2003). Their HTR features capture fine motion trajectories to guide deblurring, while their HDR features perceive scene structure in extremely dark environments to aid low-light enhancement (Brandli et al. 2014). The precise spatiotemporal information provided by these two features perfectly complements the rich texture and color of RGB frames.

However, effectively fusing the advantages of these two heterogeneous modalities remains an open challenge. Existing event-*RGB* fusion methods, including some pioneering work (Kim et al. 2024), mostly follow a "staged fusion" strategy. In this paradigm, the network processes the two modalities in independent streams, exchanging information only at specific, discrete nodes, as shown in Figure 1. This discontinuous fusion approach fundamentally limits the network's ability to learn a deep, collaborative feature representation. Between the two fusion nodes, the network is forced to independently learn suboptimal single-modal features, making the fusion itself more of a "patching" operation than a deep integration. This approach fails to fully exploit the fine-grained spatiotemporal correlations between the two data sets, severely limiting its performance in scenarios that rely heavily on complementary information, such as intense motion or extreme dimming. Therefore, we believe that a more optimal solution should implement deep interaction throughout, allowing the features of the two modalities to co-evolve at every layer of the network processing, thereby building a comprehensive and robust understanding of degraded scenes.

To overcome the limitations of "staged fusion," we propose a novel restoration framework, *CompEvent*, to achieve "full-process fusion." The core idea of *CompEvent* is to leverage the inherent coupling properties of complex algebra to achieve deep interaction between modalities. Specifically, we unify the low-light blur *RGB* features and high-temporal-resolution event features as the real and imaginary parts of a complex tensor. To achieve this full-process fusion paradigm, *CompEvent*'s architecture consists of two core complex-domain components. The first is the Complex Temporal Alignment Gated Recurrent Unit, which extends the GRU mechanism, known for its temporal processing capabilities, to the complex domain. It aligns and fuses video and event streams through complex convolutions, robustly processing temporal information in a recursive manner. The time-aligned features are then fed into the second core component, the Complex Space-Frequency Learning (CSFL) module. This module, serving as the backbone of the network, collaboratively performs spatial and frequency domain processing in a unified complex domain, achieving joint restoration by deeply integrating the spatial structure and frequency representations of the scene. Using complex operations to process the Fourier spectrum avoids the information fragmentation caused by the forced separation of real and imaginary components in traditional real networks (as shown in Figure 1). Leveraging the holistic representational power of complex networks, *CompEvent* internalizes modal fusion into its fundamental operations, enabling full-

process spatiotemporal fusion and maximizing complementary learning between modalities.

In summary, our contributions are as follows:

- We propose *CompEvent*, a complex-valued event-*RGB* video restoration framework that integrates modal fusion throughout the entire process of feature extraction, alignment, and restoration, fully leveraging the complementary advantages of event and *RGB*.
- We design the Complex Temporal Alignment Gated Recurrent Unit, which organically combines the inherent fusion capabilities of complex operations with the temporal modeling advantages of recurrent neural networks, achieving temporal alignment that is robust to severely degraded videos.
- We construct the Complex Spatial-Frequency Learning module, which synergistically processes spatial structure and frequency representations in the unified complex domain. This module can more effectively utilize the fused multimodal information to jointly correct motion blur and low-light effects.

Experiments on multiple benchmarks show that *CompEvent* outperforms state-of-the-art methods on the joint task of low-light video enhancement and deblurring, validating its effectiveness.

## Related Work

### Motion Deblurring

Traditional video deblurring relies on frames alone. Early methods were model-based (Levin et al. 2009). With the rapid development of deep learning (Li et al. 2023, 2024a,b; Jiang, Xu, and Wang 2024; Zhu et al. 2024; Li et al. 2025a,b), modern approaches leverage deep learning networks, from CNNs (Nah, Hyun Kim, and Mu Lee 2017) to Transformers (Wang et al. 2022; Zamir et al. 2022). However, in low-light scenes, long exposure times worsen motion blur and severely degrade edge and texture detail (Kim et al. 2024), making motion estimation and detail recovery extremely difficult. To address this, researchers use event cameras, which asynchronously record brightness changes with high temporal resolution, capturing motion trajectories lost in blurred frames. Events thus serve as effective motion priors for handling severe blur (Qi et al. 2024; Liang et al. 2023).

### Low-light Enhancement

Low-light enhancement methods are also predominantly frame-based. They include Retinex-based models (Land 1977; Wei et al. 2018) that decompose images into illumination and reflection, and zero-reference methods (Guo et al. 2020; Jiang et al. 2021) that learn enhancement without ground truth. However, applying these methods directly to blurry low-light videos amplifies noise and artifacts (Zhou, Li, and Change Loy 2022). Event cameras, due to their high dynamic range, can preserve scene structure even in underexposed areas. Introducing events provides structural priors absent in frames, aiding detail restoration and brightness improvement without blindly enhancing degradation (Fu et al. 2024; Xu et al. 2024; Liu et al. 2025; Sun et al. 2025).

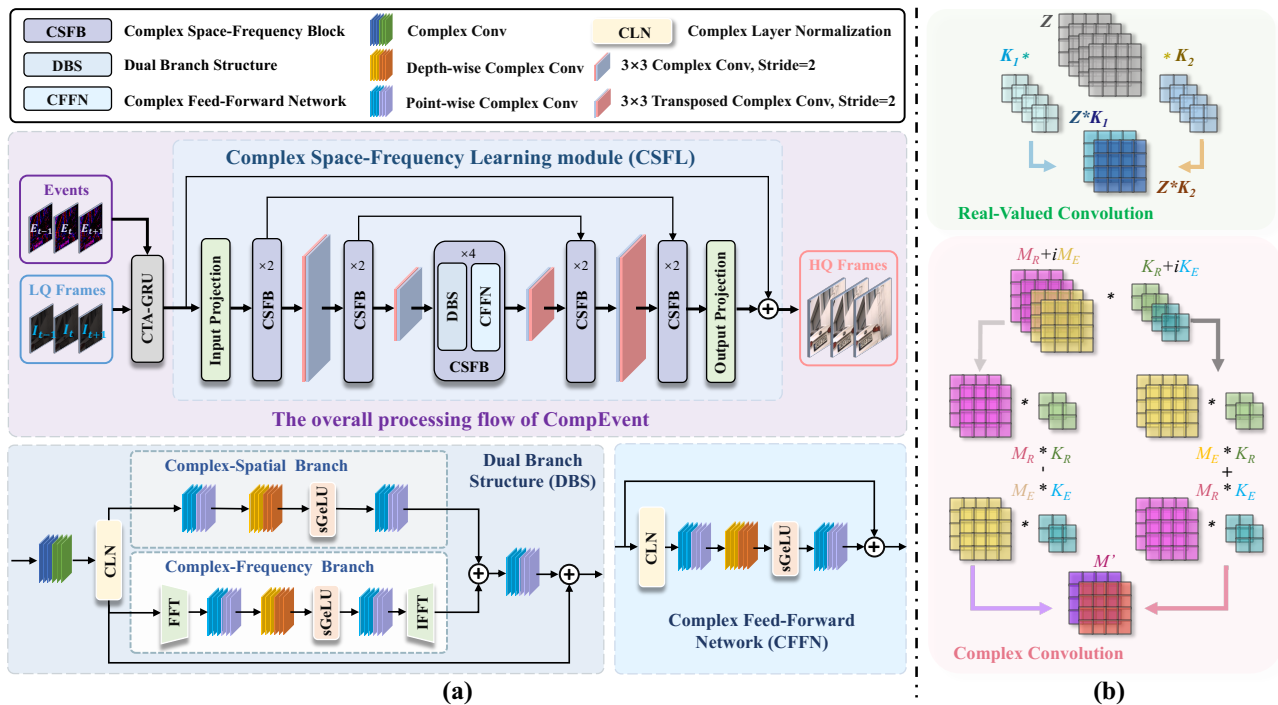


Figure 2: (a) Overall architecture of the CompEvent framework. The Complex Temporal Alignment Gated Recurrent Unit (CTA-GRU) is shown in Figure 3. (b) Comparison between the operations of complex convolution and real-valued convolution.

### Joint Low-light Enhancement and Deblurring

Low-light and motion blur are physically coupled, motivating unified solutions. Frame-based methods such as LED-Net (Zhou, Li, and Change Loy 2022) and JUDE (Vo and Park 2025) improve upon cascaded models via joint architectures and deep algorithm but remain limited by frame information (Kim et al. 2024). (Kim et al. 2024) introduced the RELED dataset and ED-TFA, a staged event-frame fusion module. Yet staged fusion—common in prior work—restricts cross-modal interaction to specific stages, limiting recovery under extreme degradation. We instead propose a full-process fusion strategy using complex neural networks, enabling continuous spatiotemporal integration to more effectively solve this joint restoration task.

## Methodology

### Overall Framework

Figure 2 illustrates the overall architecture of the proposed CompEvent framework. At each time step  $t$ , the low-light blurry RGB frame  $I_t \in \mathbb{R}^{H \times W \times 3}$  and the corresponding event representation  $E_t \in \mathbb{R}^{H \times W \times C_E}$  are processed by two separate embedding networks:  $\mathcal{F}_R$  and  $\mathcal{F}_I$ , each composed of several convolutional layers. These networks extract the real and imaginary components of the complex-valued representation, which are then combined into a complex tensor:

$$Z_t = \mathcal{F}_R(I_t) + i \cdot \mathcal{F}_I(E_t)$$

where  $i$  is the imaginary unit, and  $H, W, C$  represent the height, width, and number of channels, respectively. This representation is not a simple concatenation of the two

modalities along the channel dimension. Instead, it leverages the complex convolution algorithm to inherently promote the joint learning of features from both the real and imaginary parts, thereby achieving more effective information fusion. In real-valued convolution, the convolution operation between the feature map  $Z$  and the kernel  $K$  is expressed as  $Z * K$ , where  $*$  represents the convolution operation. In complex convolution, the convolution result  $M'$ , of the complex-valued feature maps  $M = M_R + iM_E$  and the complex kernel  $K = K_R + iK_E$  is:

$$\begin{aligned} M' &= K * M \\ &= (K_R * M_R - K_E * M_E) + \\ &\quad i \cdot (K_R * M_E + K_E * M_R) \end{aligned}$$

As shown in the Figure 2, complex convolution jointly operates on the real part  $M_R$  and the imaginary part  $M_E$  via a shared kernel consisting of  $K_R$  and  $K_E$  (Luo et al. 2025). This operation mechanism allows each complex output  $M'$  to depend on both modalities, achieving tighter fusion than real convolution. Furthermore, compared to real convolution, this shared structure reduces the number of parameters by nearly 50% while enhancing cross-modal learning. Complex convolution naturally supports a "full-process fusion" strategy, promoting the continuous interaction of RGB and event features throughout the network.

The overall processing flow of CompEvent is shown in Figure 2: CompEvent receives three consecutive frames of complex features  $\{Z_{t-1}, Z_t, Z_{t+1}\}$ . These are first input into the CTA-GRU module for temporal alignment. This module models temporal relationships in the complex-

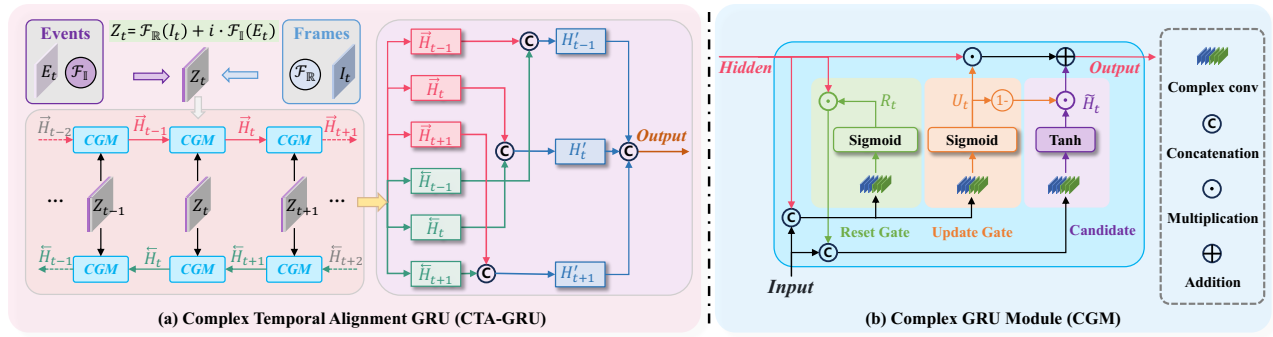


Figure 3: Complex Temporal Alignment Gated Recurrent Unit (CTA-GRU). It consists of multiple cascaded Complex GRU Modules (CGM) with a bidirectional architecture, enabling the fusion of temporal information from both past and future frames.

valued domain and robustly aligns the features by incorporating the context of the previous and next frames. Subsequently, the aligned features  $\{H'_{t-1}, H'_t, H'_{t+1}\}$  are concatenated and fed into the CSFL module. CSFL uses a hierarchical U-Net structure, combines spatial and frequency information modeling details, outputs a restored residual map for each frame, and adds it to the input image to ultimately generate a clear and bright video sequence.

### Complex Temporal Alignment GRU

Accurate temporal alignment of features becomes highly challenging under severe motion blur and low-light noise. Traditional optical flow methods (Horn and Schunck 1981; Lucas and Kanade 1981), rely on the assumptions of constant brightness and spatial smoothness. However, these assumptions often fail under conditions of blur caused by long exposures and noise caused by low light, resulting in motion estimation failures. To address this, we adopt recurrent neural networks (RNNs) to model temporal relations implicitly, avoiding explicit motion estimation. Specifically, we employ gated recurrent units (GRUs), which flexibly regulates information flow through a gating mechanism and can maintain higher stability in the face of uncertainty and noise (Zhou, Li, and Change Loy 2022).

To better utilize the high temporal resolution of event data, we extend GRU to the complex-valued domain and propose CTA-GRU. Let  $Z_t \in \mathbb{C}^{H \times W \times C}$  denote the complex-value input at time  $t$ , and  $H_{t-1} \in \mathbb{C}^{H \times W \times C}$  the previous hidden state. The complex reset and update gates are computed as:

$$R_t = \sigma_c(CCConv_r([Z_t, H_{t-1}]))$$

$$U_t = \sigma_c(CCConv_u([Z_t, H_{t-1}]))$$

where  $CCConv$  denotes complex convolution,  $\sigma_c$  is the complex sigmoid, and  $[\cdot, \cdot]$  denotes channel-wise concatenation. The complex candidate hidden state is:

$$\tilde{H}_t = \tanh_c(CCConv_h([Z_t, R_t \odot H_{t-1}]))$$

and the updated hidden state becomes:

$$H_t = (1 - U_t) \odot H_{t-1} + U_t \odot \tilde{H}_t$$

Where,  $\odot$  is the complex Hadamard product. We adopt split activations: for input  $z = x + iy$ , we define  $\sigma_c(z) =$

$\sigma(x) + i\sigma(y)$ , which has been proven effective and stable in practice (Nah, Hyun Kim, and Mu Lee 2017; Zamir et al. 2022).

The core advantage of the CTA-GRU lies in its gating mechanism being driven by complex convolutions, meaning that the reset and update gate decisions are based on a deep fusion of RGB features (real part) and event features (imaginary part). For example, when processing a fast-moving object, the event stream of the current frame (the imaginary part of  $Z_t$ ) provides a clear motion trajectory. This information is passed to the reset gate  $R_t$  via the complex convolution  $CCConv_r$ . This gate "realizes" that the features corresponding to the object's old position in the previous hidden state  $H_{t-1}$  are outdated, and thus generates a smaller gate value to "reset" or ignore this information.

To fully utilize the contextual information in a video, we use bidirectional CTA-GRU (Wang et al. 2022; Jiang et al. 2021). A forward pass processes  $t-1 \rightarrow t \rightarrow t+1$ , yielding  $\{\vec{H}_{t-1}, \vec{H}_t, \vec{H}_{t+1}\}$ , while a backward pass processes  $t+1 \rightarrow t \rightarrow t-1$ , yielding  $\{\overleftarrow{H}_{t+1}, \overleftarrow{H}_t, \overleftarrow{H}_{t-1}\}$ . The aligned feature at  $t$  is:

$$H'_t = \text{concat}(\vec{H}_t, \overleftarrow{H}_t)$$

In this way, the features of each frame incorporate information from both the past and the future. Figure 3 illustrates the overall architecture of the proposed CTA-GRU. The aligned features  $\{H'_{t-1}, H'_t, H'_{t+1}\}$  are then concatenated and fed into CSFL module for space-frequency restoration.

### Complex Space-Frequency Learning

After temporal alignment through the CTA-GRU module, the features are fed into the backbone network for final image restoration. The mixed degradation of low light and motion blur exhibits different characteristics in different image domains: low light primarily affects low-frequency components (such as overall brightness and contrast) (Huang et al. 2022; Chen and Jin 2023), while motion blur primarily manifests as an attenuation of high-frequency components (such as edges and texture) (Manolakis and Ingle 2011). Therefore, an ideal restoration network should be able to synergistically address the fine spatial structure and systematic frequency characteristics of the image.

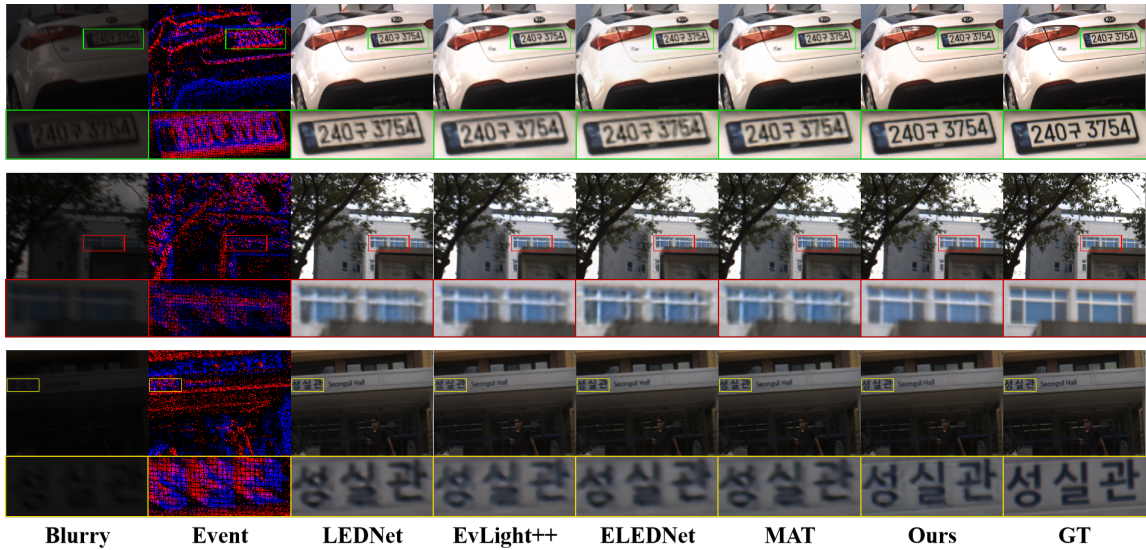


Figure 4: Qualitative comparisons on the RELED dataset. Zoom in for better view.

The proposed Complex Space-Frequency Learning (CSFL) module adopts an encoder-decoder architecture shown in Figure 2, where downsampling and upsampling are handled by complex convolutions and transposed convolutions, respectively. Skip connections restore details at each level. Each Complex Space-Frequency Block (CSFB) processes input  $X_{l-1} \in \mathbb{C}^{H \times W \times C}$  as follows. First, features are normalized via complex layer normalization (CLN) (See the supplementary material for details) :

$$X'_l = \text{CLN}(X_{l-1})$$

We use whitening-based CLN (Trabelsi et al. 2017) to preserve correlations between real and imaginary components, ensuring stability during training .

The features are then fed into a Dual Branch Structure (DBS) . The first branch of DBS is the Complex Space Branch, which aims to learn spatially varying, deeply fused representations, extracting and fusing features simultaneously. Its main structure draws on depth-wise separable convolution (Chollet 2017; Howard et al. 2017)—an efficient form of convolution widely used in modern network architectures. We extend it to the complex domain to process the fused features. The operation can be expressed as:

$$Y_{\text{spatial}} = \text{CConv}_{\text{pw}2}(\text{sGeLU}(\text{CConv}_{\text{dw}}(\text{CConv}_{\text{pw}1}(X'))))$$

Where,  $\text{CConv}_{\text{pw}1}$  and  $\text{CConv}_{\text{pw}2}$  are  $1 \times 1$  complex point-wise convolutions used for channel mapping and mixing;  $\text{CConv}_{\text{dw}}$  is a complex depth-wise separable convolution used to efficiently extract spatial features. Based on the complex activation function of GeLU , the sGeLU also employs a split activation strategy, applying the real-valued GeLU function to the real and imaginary parts of the complex input separately.

Another parallel branch of DBS is the Complex Frequency Branch, which processes systematic degradations. Features are first transformed via a two-dimensional fast Fourier transform (FFT) :

$$\mathcal{F}(X'_l) = \text{FFT2D}(X'_l)$$

The resulting complex spectrum  $\mathcal{F}(X'_l)$  is then processed in a complex convolutional network with a similar spatial branching structure :

$$\mathcal{F}_{\text{proc}} = \text{CConv}_{\text{pw}2}(\text{sGeLU}(\text{CConv}_{\text{dw}}(\text{CConv}_{\text{pw}1}(\mathcal{F}(X'_l))))))$$

Then, the processed spectrum is converted back to the spatial domain via an inverse Fourier transform (IFFT) :

$$Y_{\text{freq}} = \text{IFFT2D}(\mathcal{F}_{\text{proc}})$$

The entire spectrum is processed holistically without separating real and imaginary parts, preserving the spectral structure and enabling adaptive corrections in the complex domain.

Outputs from both branches are combined with the input via a residual connection:

$$X''_l = X_{l-1} + Y_{\text{spatial}} + Y_{\text{freq}}$$

Finally,  $X''_l$  passes through a Complex Feed-Forward Network (CFFN) for further refinement:

$$X_l = \text{CFFN}(\text{CLN}(X''_l)) + X''_l$$

Where,  $\text{CFFN}(X) = \text{CConv}_{\text{pw}4}(\text{sGeLU}(\text{CConv}_{\text{pw}3}(X)))$ ,  $\text{CConv}_{\text{pw}3}$  and  $\text{CConv}_{\text{pw}4}$  are  $1 \times 1$  complex point-wise convolutions. CFFN performs nonlinear transformations directly in the complex domain. It not only models complex feature relationships like real-valued FFN, but also preserves and utilizes the phase information of the signal, thus possessing richer representation capabilities (Bassey, Qian, and Li 2021). By stacking multiple such CSFBs, our network can deeply and collaboratively process spatial and frequency domain information at different scales, thereby achieving effective restoration of low-light blurry videos.

## Experiments and Analysis

### Datasets

We evaluate CompEvent on the RELED real-world dataset (Kim et al. 2024) and the LOL-Blur synthetic

Methods		Input	RELED		LOL-Blur	
			PSNR	SSIM	PSNR	SSIM
Low-Light Enhancement	SNRNet (Xu et al. 2022)	F	26.47	0.851	20.25	0.815
	LLFormer (Wang et al. 2023)	F	26.62	0.862	20.68	0.832
	RetinexFormer (Cai et al. 2023)	F	26.66	0.865	20.83	0.817
	SDSDNet (Wang et al. 2021)	F	28.47	0.887	21.34	0.832
	EvLight++ (Chen et al. 2024)	F+E	30.87	0.888	24.99	0.880
Motion Deblur	MPRNet (Zamir et al. 2021)	F	26.89	0.867	21.35	0.825
	MIMOUNet+ (Cho et al. 2021)	F	26.52	0.866	21.12	0.821
	NAFNet (Chen et al. 2022)	F	26.77	0.862	21.28	0.818
	RNN-MBP (Zhu et al. 2022)	F	29.52	0.902	23.58	0.862
	DSTNet (Pan et al. 2023)	F	29.59	0.903	23.63	0.864
	e-SLNet (Wang et al. 2020)	F+E	19.45	0.663	17.05	0.738
	REDNet (Xu et al. 2021)	F+E	29.19	0.903	23.25	0.859
	EFNet (Sun et al. 2022)	F+E	29.85	0.905	23.92	0.867
	MAT (Xu et al. 2025)	F+E	31.22	0.896	25.15	<u>0.882</u>
	UEVD (Kim et al. 2022)	F+E	29.93	0.905	24.08	0.869
	REFID (Sun et al. 2023)	F+E	30.10	0.913	24.55	0.875
Joint (Frame-based)	LEDNet (Zhou, Li, and Change Loy 2022)	F	30.36	0.887	<u>25.74</u>	0.850
Joint (Event-guided)	ELEDNet (Kim et al. 2024)	F+E	<u>31.30</u>	<u>0.925</u>	25.04	0.873
	<b>Ours</b>	F+E	<b>32.51</b>	<b>0.928</b>	<b>28.73</b>	<b>0.907</b>

Table 1: The quantitative results on RELED and LOL-Blur. ‘‘F’’ denotes image frame-based methods, while ‘‘F+E’’ represents frame-based methods integrated with event-guided information. Best and second-best results are boldfaced and underlined.

dataset (Zhou, Li, and Change Loy 2022). Training details and hyperparameters are provided in the supplementary material. (1) **RELED** (Kim et al. 2024) is the first large-scale real-world benchmark built for the joint low-light enhancement and motion deblurring tasks. The dataset is acquired through an optical beam splitting system that can simultaneously record low-light blurry videos, the corresponding high-quality clear images, and high-fidelity event streams. (2) **LOL-Blur** (Zhou, Li, and Change Loy 2022) is a large-scale synthetic dataset that provides low-light blurry image and clear image pairs for the joint low-light enhancement and deblurring tasks. To adapt our event-based approach, we generate the corresponding event streams for it using the ESIM simulator (Rebecq, Gehrig, and Scaramuzza 2018).

### Comparison with State-of-the-Art Methods

To comprehensively evaluate the CompEvent framework, we conduct extensive comparisons against state-of-the-art methods across a variety of tasks and modalities. These baselines are systematically categorized into four categories: single low-light enhancement methods, single motion deblurring methods, frame-only joint restoration methods, and event-guided joint restoration methods. The detailed description of the experimental setup for the baseline comparison method is provided in the supplementary material.

**Real-World Dataset (RELED):** As shown in Table 1, CompEvent achieves a PSNR of 32.51 dB and an SSIM of 0.928 on the real-world RELED dataset, outperforming all single-task methods, including EvLight++ (30.87 / 0.888) for low-light enhancement and MAT (31.22 / 0.896) for deblurring, as well as joint frameworks such as LEDNet (30.36 / 0.887) and ELEDNet (31.30 / 0.925). As shown in Figure 4, CompEvent produces sharper structures with fewer artifacts

Model Variant	PSNR	SSIM
(a) CompEvent (Full)	32.51	0.928
(b) w/o Complex (Concat)	31.39 (-1.12↓)	0.901 (-0.027↓)
(c) w/o GRU (Static)	30.87 (-1.64↓)	0.885 (-0.043↓)
(d) w/o GRU (Concat)	31.93 (-0.58↓)	0.914 (-0.014↓)
(e) w/o Freq. Branch	31.76 (-0.75↓)	0.908 (-0.020↓)

Table 2: Ablation study of the core components of CompEvent on the RELED dataset.

under complex lighting and motion conditions. These results demonstrate the effectiveness of our method for real-world video restoration.

**Synthetic Dataset (LOL-Blur):** As shown in Table 1, CompEvent achieves 28.73 / 0.907 PSNR/SSIM, surpassing the best single-task baselines EvLight++ (24.99 / 0.880) and MAT (25.15 / 0.882) by large margins. Compared to joint methods LEDNet (25.74 / 0.850) and ELEDNet (25.04 / 0.873), CompEvent improves by up to +3.69 dB and +0.057 SSIM. Qualitative results in Figure 5 confirm its superior texture and structural recovery under synthetic degradation.

### Ablation Studies

We conduct ablation studies on the RELED dataset to evaluate the effectiveness of each component in our CompEvent framework. By systematically removing or substituting key modules, we quantify their contributions.

**Effectiveness of Complex Full-Process Fusion.** To evaluate our fusion design, we construct a real-valued variant (b) without Complex (Concat), replacing complex convolution with a real-valued counterpart of similar parameter size, where RGB and event features are concatenated in the channel dimension. Results in Table 2 show that this variant

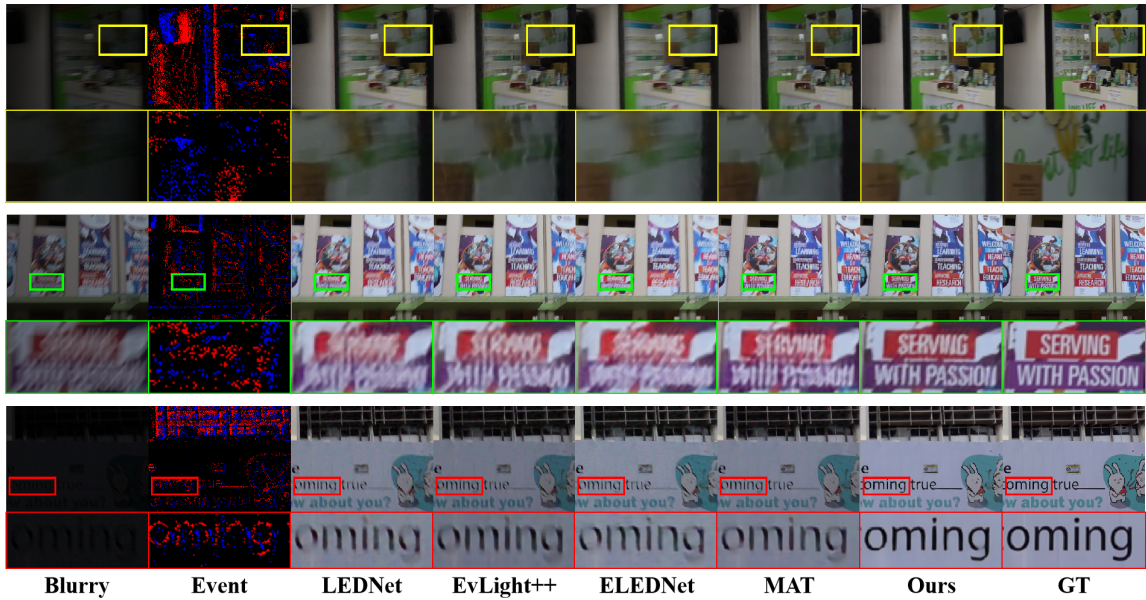


Figure 5: Qualitative comparisons on the LOL dataset. Zoom in for better view.

achieves a 1.12 dB decrease in PSNR, demonstrating that the performance improvement does not come from more parameters, but rather benefits from the inherent fusion mechanism brought by complex convolution throughout the network, which enables the network to more effectively utilize the complementary advantages of the two modalities to cope with complex mixed degradations.

**Effectiveness of CTA-GRU for Temporal Modeling.** We designed two variants to evaluate the contribution of CTA-GRU: (c) without GRU (Static), completely removing the GRU; and (d) without GRU (Concat), replacing the GRU model with cross-frame concatenation. Results in Table 2 show that removing the CTA-GRU significantly degrades model performance, demonstrating the critical importance of modeling temporal context for video restoration. Furthermore, compared to cross-frame concatenation, CTA-GRU achieves dynamic alignment between frames through a gating mechanism, demonstrating greater robustness to large motion and inter-frame misalignment.

**Effectiveness of CSFL Module.** We validated the design of the Complex Space-Frequency Learning (CSFL) module. This module consists of a spatial branch and a frequency branch. To evaluate the contribution of the frequency branch, we constructed a variant (e) w/o Freq. Branch, which removes the frequency branch while retaining the spatial branch. Results in Table 2 show that this variant significantly degrades performance, validating that separately modeling blur (high-frequency) and low-light (low-frequency) degradation in the frequency domain is an effective strategy for addressing this mixed degradation problem. The complete CSFL module achieves the best overall performance through space-frequency co-processing.

**Effectiveness of Complex Layer Normalization (CLN).** To evaluate CLN, we test (f) w/o CLN (Separate), apply-

Model Variant	PSNR	SSIM
(a) CompEvent (Full)	32.51	0.928
(f) w/o CLN (Separate)	31.98 (-0.53↓)	0.915 (-0.013↓)
(g) w/o CLN (Concat)	31.55 (-0.96↓)	0.904 (-0.024↓)

Table 3: Ablation study of the Complex Layer Normalization (CLN) on the RELED dataset.

ing real-valued normalization to real and imaginary parts separately, and (g) w/o CLN (Concat), concatenating and normalizing them jointly. Results in Table 3 show that both alternatives lead to performance degradation, with (g) performing the worst. This suggests that directly treating complex features as ordinary real-valued vectors for normalization destroys their algebraic structure and weakens their expressive power. In contrast, CLN normalizes the variance of the real and imaginary parts by using a whitening transformation and considers their covariance to decorrelate them, improving training stability and final performance.

## Conclusion

We propose CompEvent, a complex-valued neural network framework for joint low-light video enhancement and deblurring, enabling holistic full-process fusion of event data and RGB frames. It features two core components: the Complex Temporal Alignment GRU for efficient temporal alignment and recursive fusion via complex operations, and the Complex Space-Frequency Learning module for synergistic spatial and frequency domain processing in a unified complex domain. Throughout the process, CompEvent overcomes "staged fusion" limitations by leveraging the inherent complex-valued operations of complex convolution to fuse RGB and event information at every step, enabling full-process spatiotemporal fusion. Extensive experimental results on several benchmarks demonstrate the effectiveness of the proposed method.

## Acknowledgments

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

## References

- Bassey, J.; Qian, L.; and Li, X. 2021. A survey of complex-valued neural networks. *arXiv preprint arXiv:2101.12249*.
- Brandli, C.; Berner, R.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2014. A  $240 \times 180 \times 130$  db  $3 \mu\text{s}$  latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10): 2333–2341.
- Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; and Zhang, Y. 2023. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12504–12513.
- Chen, H.; and Jin, Z. 2023. Low-Light Enhancement in the Frequency Domain. *arXiv preprint arXiv:2306.16782*.
- Chen, K.; Liang, G.; Li, H.; Lu, Y.; and Wang, L. 2024. Ev-light++: Low-light video enhancement with an event camera: A large-scale real-world dataset, novel method, and more. *arXiv preprint arXiv:2408.16254*.
- Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *European conference on computer vision*, 17–33. Springer.
- Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4641–4650.
- Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.
- Fu, X.; Cao, C.; Xu, S.; Zhang, F.; Wang, K.; and Zha, Z.-J. 2024. Event-driven heterogeneous network for video deraining. *International Journal of Computer Vision*, 132(12): 5841–5861.
- Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1780–1789.
- Horn, B. K.; and Schunck, B. G. 1981. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203.
- Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Huang, J.; Liu, Y.; Zhao, F.; Yan, K.; Zhang, J.; Huang, Y.; Zhou, M.; and Xiong, Z. 2022. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*, 163–180. Springer.
- Jiang, S.; Xu, S.; and Wang, X. 2024. Rbsformer: Enhanced transformer network for raw image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6479–6488.
- Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30: 2340–2349.
- Kim, T.; Jeong, J.; Cho, H.; Jeong, Y.; and Yoon, K.-J. 2024. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, 433–451. Springer.
- Kim, T.; Lee, J.; Wang, L.; and Yoon, K.-J. 2022. Event-guided deblurring of unknown exposure time videos. In *European Conference on Computer Vision*, 519–538. Springer.
- Land, E. H. 1977. The retinex theory of color vision. *Scientific american*, 237(6): 108–129.
- Levin, A.; Weiss, Y.; Durand, F.; and Freeman, W. T. 2009. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE conference on computer vision and pattern recognition*, 1964–1971. IEEE.
- Li, D.; Liu, Y.; Fu, X.; Xu, S.; and Zha, Z.-J. 2024a. Fouriermamba: Fourier learning integration with state space models for image deraining. *arXiv preprint arXiv:2405.19450*.
- Li, D.; Luo, C.; Bao, Y.; Yang, G.; Xiao, J.; Fu, X.; and Zha, Z.-J. 2025a. Enhanced Pansharpening via Quaternion Spatial-Spectral Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10908–10918.
- Li, D.; Zhu, J.; Fu, X.; Guo, X.; Liu, Y.; Yang, G.; Liu, J.; and Zha, Z.-J. 2024b. Noise-Assisted Prompt Learning for Image Forgery Detection and Localization. In *European Conference on Computer Vision*, 18–36. Springer.
- Li, D.; Zhu, J.; Liu, Y.; Lu, X.; Fu, X.; Liu, J.; Liu, A.; and Zha, Z.-J. 2025b. Learnable frequency decomposition for image forgery detection and localization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 1359–1367.
- Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.
- Liang, J.; Yang, Y.; Li, B.; Duan, P.; Xu, Y.; and Shi, B. 2023. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10615–10625.
- Lichtsteiner, P. 2003.  $64 \times 64$  event-driven logarithmic temporal derivative silicon retina. In *Program 2003 IEEE Workshop on CCD and AIS*.
- Liu, K.; Zhong, M.; Xu, S.; Sun, Z.; Zhu, J.; Ge, C.; Wang, X.; Lu, X.; Fu, X.; and Zha, Z.-J. 2025. Event-conditioned dual-modal fusion for motion deblurring. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1482–1492.

- Lucas, B. D.; and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, 674–679.
- Luo, C.; Li, D.; Ma, X.; Lu, X.; Wang, Z.; Tan, J.; and Fu, X. 2025. PanComplex: leveraging complex-valued neural networks for enhanced pansharpening. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 1702–1710.
- Manolakis, D. G.; and Ingle, V. K. 2011. *Applied digital signal processing: theory and practice*. Cambridge university press.
- Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.
- Pan, J.; Xu, B.; Dong, J.; Ge, J.; and Tang, J. 2023. Deep discriminative spatial and temporal network for efficient video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22191–22200.
- Qi, Y.; Zhu, L.; Zhao, Y.; Bao, N.; and Li, J. 2024. Deblurring neural radiance fields with event-driven bundle adjustment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9262–9270.
- Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. Esim: an open event camera simulator. In *Conference on robot learning*, 969–982. PMLR.
- Sun, L.; Sakaridis, C.; Liang, J.; Jiang, Q.; Yang, K.; Sun, P.; Ye, Y.; Wang, K.; and Gool, L. V. 2022. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*, 412–428. Springer.
- Sun, L.; Sakaridis, C.; Liang, J.; Sun, P.; Cao, J.; Zhang, K.; Jiang, Q.; Wang, K.; and Van Gool, L. 2023. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18043–18052.
- Sun, Z.; Xu, S.; Liu, K.; Tian, R.; Fu, X.; and Zha, Z.-J. 2025. EVDM: Event-based Real-world Video Deblurring with Mamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13793–13803.
- Trabelsi, C.; Bilaniuk, O.; Zhang, Y.; Serdyuk, D.; Subramanian, S.; Santos, J. F.; Mehri, S.; Rostamzadeh, N.; Bengio, Y.; and Pal, C. J. 2017. Deep complex networks. *arXiv preprint arXiv:1705.09792*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Vo, T.; and Park, C. Y. 2025. Deep Joint Unrolling for Deblurring and Low-Light Image Enhancement (JUDE). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2696–2705. IEEE.
- Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*, 155–171. Springer.
- Wang, R.; Xu, X.; Fu, C.-W.; Lu, J.; Yu, B.; and Jia, J. 2021. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9700–9709.
- Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; and Lu, T. 2023. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI conference on artificial intelligence*, 2654–2662.
- Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.
- Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*.
- Xu, F.; Yu, L.; Wang, B.; Yang, W.; Xia, G.-S.; Jia, X.; Qiao, Z.; and Liu, J. 2021. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2583–2592.
- Xu, S.; Sun, Z.; Zhong, M.; Cao, C.; Liu, Y.; Fu, X.; and Chen, Y. 2025. Motion-adaptive Transformer for Event-based Image Deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8942–8950.
- Xu, S.; Sun, Z.; Zhu, J.; Zhu, Y.; Fu, X.; and Zha, Z.-J. 2024. Demosaicformer: Coarse-to-fine demosaicing network for hybridevs camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1126–1135.
- Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2022. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17714–17724.
- Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.
- Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.
- Zhou, S.; Li, C.; and Change Loy, C. 2022. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, 573–589. Springer.
- Zhu, C.; Dong, H.; Pan, J.; Liang, B.; Huang, Y.; Fu, L.; and Wang, F. 2022. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *Proceedings of the AAAI conference on artificial intelligence*, 3598–3607.
- Zhu, J.; Li, D.; Fu, X.; Yang, G.; Huang, J.; Liu, A.; and Zha, Z.-J. 2024. Learning discriminative noise guidance for image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7739–7747.