Seeing the Unseen: Zooming in the Dark with Event Cameras

Dachun Kai[1], **Zeyu Xiao**[2], **Huyue Zhu**[1], **Jiaxiao Wang**[1], **Yueyi Zhang**[3], **Xiaoyan Sun**[1,4*]
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[2]National University of Singapore
[3]Miromind AI
[4]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
dachunkai@mail.ustc.edu.cn, sunxiaoyan@ustc.edu.cn

Abstract

This paper addresses low-light video super-resolution (LVSR), aiming to restore high-resolution videos from low-light, low-resolution (LR) inputs. Existing LVSR methods often struggle to recover fine details due to limited contrast and insufficient high-frequency information. To overcome these challenges, we present RetinexEVSR, the first event-driven LVSR framework that leverages high-contrast event signals and Retinex-inspired priors to enhance video quality under low-light scenarios. Unlike previous approaches that directly fuse degraded signals, RetinexEVSR introduces a novel bidirectional cross-modal fusion strategy to extract and integrate meaningful cues from noisy event data and degraded RGB frames. Specifically, an illumination-guided event enhancement module is designed to progressively refine event features using illumination maps derived from the Retinex model, thereby suppressing low-light artifacts while preserving high-contrast details. Furthermore, we propose an event-guided reflectance enhancement module that utilizes the enhanced event features to dynamically recover reflectance details via a multi-scale fusion mechanism. Experimental results show that our RetinexEVSR achieves state-of-the-art performance on three datasets. Notably, on the SDSD benchmark, our method can get up to 2.95 dB gain while reducing runtime by 65% compared to prior event-based methods.

Code — https://github.com/DachunKai/RetinexEVSR

1 Introduction

Video super-resolution (VSR) aims to restore high-resolution (HR) videos from low-resolution (LR) inputs. While existing methods (Zhou et al. 2024) get good results on general videos, they often fail under low-light conditions. However, such conditions are common in real-world applications, such as video surveillance, where zooming in on distant license plates or human faces at night is often required. Other important scenarios include remote sensing (Xiao et al. 2025) and night videography (Yue, Gao, and Su 2024; Li et al. 2025a). Therefore, it is essential to develop VSR algorithms specifically designed for low-light videos.

To achieve VSR from low-light videos, *i.e.*, low-light VSR (LVSR), a straightforward approach is to first apply low-light video enhancement (LVE) (Li et al. 2023), followed by VSR methods, which we refer to as the *cascade* strategy. However, this approach has a major drawback in that the pixel errors introduced during the LVE stage are propagated and amplified in the VSR step, thus degrading the overall performance. An alternative strategy is to perform VSR first and then apply LVE. However, the quality deteriorates because the super-resolved frames suffer from weakened textures, amplified noise, and low contrast. To address these issues, Xu et al. (2023b) proposed the first one-stage LVSR model that directly learns a mapping from low-light LR inputs to well-lit HR outputs. However, as shown in Fig. 1, these methods still suffer from severe artifacts, structural distortions, and inaccurate illumination.

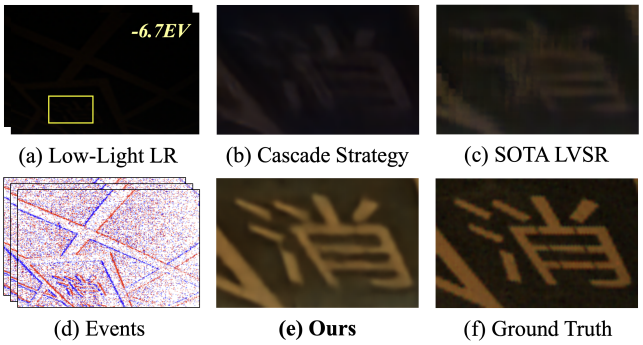

Figure 1: An example (a) from an extremely low-light (-6.7 EV) LR sample, enhanced by (b) SOTA LVE (Li et al. 2023) + VSR (Xu et al. 2024) methods; (c) SOTA one-stage LVSR method (Lu et al. 2023); and (e) our event-based approach. It can be observed that only our method produces well-lit, high-quality results with clearly recognizable text.

LVSR is a very challenging problem. It is difficult to rely solely on low-light LR frames to restore high-quality HR videos due to the inherent lack of sufficient contrast to distinguish fine textures, as well as the lack of high-frequency details in LR frames. In addition, sudden lighting changes at night, such as flashes from streetlights or car headlights, further exacerbate the problem. Recently, event signals captured by event cameras have been used for low-light enhancement (Liang et al. 2023), super-resolution (Kai, Zhang, and Sun 2023), and high dynamic range imaging (Han et al.

*Corresponding author.

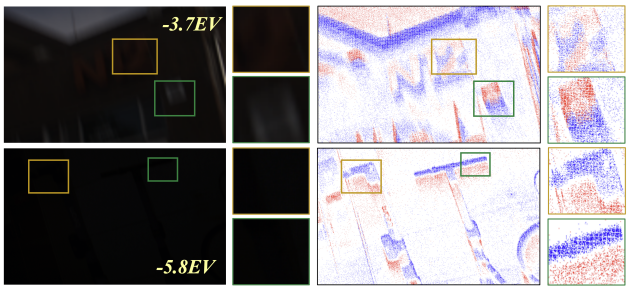

Figure 2: In low light, both RGB and event signals degrade: the RGB frame suffers from severe illumination and detail loss, and the event data contains noise and trailing artifacts.

2023). Compared with standard cameras, event cameras offer a very high dynamic range (120 dB), high temporal resolution (about 1 μs), and rich "moving edge" information (Gallego et al. 2020). These characteristics enable event signals to provide complementary cues, such as sharp edges and motion details, even at night, for LVSR. Motivated by these advantages, we propose including event signals as auxiliary information to improve LVSR performance.

However, while event signals offer valuable information, effectively integrating them into LVSR remains challenging. As shown in Fig. 2, not only are RGB frames heavily degraded under low-light conditions, but event data also suffers from noise, temporal trailing effects, and spatially non-stationary distributions (Liu et al. 2025b). Directly fusing such degraded event signals with low-quality RGB frames inevitably introduces noise and artifacts into the reconstructed results. Therefore, how to effectively extract and fuse meaningful information from both degraded signals is of paramount importance for event-based LVSR.

To achieve this, we first argue that the degradation in both modalities mainly arises from insufficient lighting, and that relying solely on event data is inadequate to address these issues without additional low-light priors. To address this, we draw inspiration from Retinex decomposition (Wei et al. 2018), which separates a low-light image into illumination and reflectance. Illumination provides smooth, low-noise global lighting cues, while reflectance preserves intrinsic scene content but lacks fine details in LR settings. Based on this insight, we propose a Retinex-inspired Bidirectional Fusion (RBF) strategy: illumination guides the refinement of noisy events, and enhanced events are then used to recover reflectance details, as illustrated in Fig. 3(c). This bidirectional process enables effective mutual guidance between RGB and event modalities.

To this end, we present RetinexEVSR, an innovative LVSR network that integrates high-contrast event signals with Retinex-inspired priors to enhance video quality under low-light conditions. In our RetinexEVSR, the input frames are first decomposed into illumination and reflectance components. Guided by the proposed RBF strategy, we introduce an Illumination-guided Event Enhancement (IEE) module, which progressively refines event features through multi-scale fusion with illumination, enabling hierarchical guid-

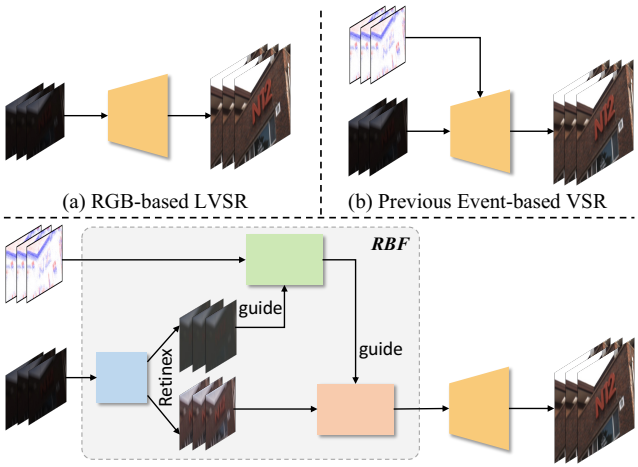

Figure 3: Comparison of LVSR strategies. (a) RGB-based method (Xu et al. 2023b) directly super-resolves low-light frames. (b) Previous event-based methods (Lu et al. 2023; Kai et al. 2024) directly fuse two degraded modalities. (c) Our RBF strategy first uses illumination to guide event refinement and then leverages the refined events to enhance reflectance, enabling effective information integration.

ance from coarse to fine levels. The refined events are then passed to the Event-guided Reflectance Enhancement (ERE) module to recover reflectance details. This module adopts a dynamic attention mechanism to inject high-frequency information from events into the reflectance stream via multi-scale fusion. Finally, the illumination, enhanced reflectance, and refined event features are jointly used to guide the upsampling process, reducing information loss and improving reconstruction quality. Experimental results on three datasets demonstrate the effectiveness of our proposed RetinexEVSR, which remains robust even under extreme darkness and severe motion blur. To summarize, our main contributions are:

- We present RetinexEVSR, the first event-driven scheme for LVSR. Our RetinexEVSR leverages event signals and Retinex-inspired priors to restore severely degraded RGB inputs under low-light conditions.
- We introduce a novel RBF strategy to enable effective cross-modal fusion between RGB and event signals, addressing the challenge of combining degraded inputs.
- We propose the IEE and ERE modules to progressively enhance event and reflectance features, enabling coarse-to-fine guidance and detailed texture restoration.
- RetinexEVSR achieves state-of-the-art performance on three datasets, including synthetic and real-world data.

2 Related Work

Video Super-Resolution. As a fundamental computer vision task, VSR technology has made remarkable progress in recent years (Li et al. 2025b; Wei et al. 2025; Xie et al. 2025). The essential challenge in VSR is to predict the missing details of the current HR frame from other unaligned

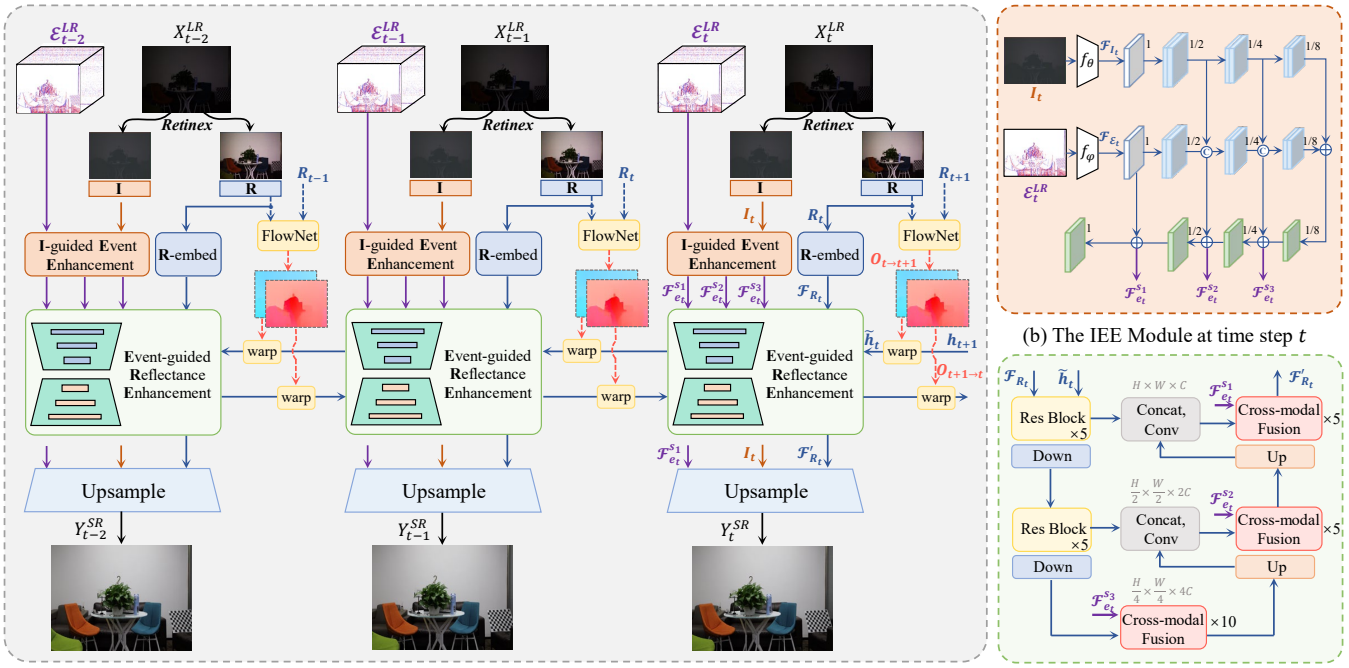

Figure 4: Network architecture of RetinexEVSR. (a) The model takes low-light LR frames and corresponding events as input, and outputs HR frames with well-lit details. Each frame is decomposed into illumination and reflectance, and optical flow is estimated from reflectance for temporal alignment. (b) At each time step, the IEE module uses illumination to guide event enhancement. (c) The refined event features are then used in the ERE module to enhance reflectance features.

frames. To achieve this, many advanced alignment (Tang et al. 2024) and propagation (Du et al. 2025) methods have been devised. However, these methods often perform poorly under low-light conditions due to issues such as amplified noise and weakened textures. To address this, some works (Xu et al. 2023a; Gao et al. 2024) have proposed joint learning of low-light enhancement and super-resolution. Xu et al. (2023b) introduced the first one-stage LVSR framework that directly learns a mapping from low-light LR videos to normal-light HR videos. However, the method still struggles with large displacements and motion blur, resulting in severe temporal inconsistency.

Low-Light Video Enhancement. To achieve LVE, a common strategy is to apply low-light image enhancement (LIE) methods to each frame independently. In recent years, a large number of CNN-based (Wu et al. 2025a,b; Ju et al. 2025) and Transformer-based (Wang et al. 2023a; Cai et al. 2023) LIE methods have emerged. Among them, the Retinex model (Wei et al. 2018) is a popular tool for LIE, where an observed image X can be expressed as $X = R \odot I$. Here, R and I represent reflectance and illumination maps, respectively, and $\odot$ denotes element-wise multiplication. Cai et al. (2023) introduced Retinexformer, the first Transformer-based method that uses illumination derived from Retinex theory to guide the modeling of long-range dependencies in the self-attention mechanism.

However, frame-by-frame LIE often causes temporal flickering and jitter effects due to dynamic illumination changes in low-light videos. To address these, many one-stage LVE methods (Zhu et al. 2024a,b) have been proposed. Li et al. (2023) proposed an efficient pipeline named FastLLVE, which leverages the look-up table technique to effectively maintain inter-frame brightness consistency. However, they still face limitations in using temporal redundancy in low-light videos due to difficulties in extracting distinct features for motion estimation.

Event-based Vision. Event cameras are bio-inspired sensors that offer several advantages over standard RGB cameras, including ultra-high temporal resolution (about 1μs) (Xiao et al. 2024a), high dynamic range (120 dB), and low power (5 mW). They have been widely used for tasks like frame interpolation (Liu et al. 2025a; Sun et al. 2025; Liu et al. 2025c), deblurring (Yang et al. 2024, 2025), and low-light enhancement (Zhang et al. 2024; Kim et al. 2024).

More closely related to our work, recent studies (Xiao et al. 2024c,b; Yan et al. 2025; Kai et al. 2025; Xiao and Wang 2025) have introduced event signals for VSR. For instance, Jing et al. (2021) proposed the first event-based VSR method, named E-VSR, which uses events for frame interpolation followed by VSR, enhancing overall performance. Kai et al. (2024) introduced EvTexture, utilizing high-frequency information from events to improve texture restoration. While these methods perform well under normal-light conditions, they struggle in low-light scenarios. The challenge of training with event data for VSR under low-light conditions remains largely unexplored.

Type	Method	SDSD-in			SDSD-out			SDE-in			SDE-out		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
I	Retinexformer + MIA-VSR	27.11	0.8341	0.3962	19.81	0.6634	0.5051	16.68	0.4053	05533	16.20	0.3981	0.5507
	FastLLVE + MIA-VSR	16.32	0.7224	0.5120	20.22	0.6500	0.5598	13.91	0.3300	0.6367	13.79	0.3030	0.6452
	Retinexformer + IART	27.09	0.8331	0.3938	19.84	0.6638	0.5051	16.68	0.4053	0.5533	16.17	0.3967	0.5473
	FastLLVE + IART	16.33	0.7217	0.5095	20.25	0.6508	0.5578	13.91	0.3300	0.6373	13.77	0.3011	0.6407
II	MIA-VSR + Retinexformer	25.10	0.8457	0.3644	23.80	0.7583	0.4316	16.82	0.4607	0.4716	15.94	0.4011	0.4942
	MIA-VSR + FastLLVE	23.94	0.8261	0.4033	13.71	0.5364	0.4303	16.39	0.4875	0.5028	15.80	0.4104	0.5200
	IART + Retinexformer	25.30	0.8500	0.3541	24.07	0.7566	0.4192	17.73	0.4339	0.4952	16.00	0.4088	0.4799
	IART + FastLLVE	24.03	0.8304	0.3944	13.61	0.5345	0.4261	16.38	0.4869	0.5017	15.75	0.4108	0.5105
III	EvLight + EGVSR	26.57	0.8220	0.3944	20.59	0.6752	0.4214	19.61	0.5939	0.5120	19.21	0.5296	0.5198
	EvLowLight + EGVSR	18.61	0.6783	0.4638	14.75	0.5147	0.5465	19.81	0.5626	0.5923	17.46	0.4389	0.6393
	EvLight + EvTexture	26.15	0.8127	0.3823	19.65	0.6504	0.4214	19.54	0.5700	0.5070	19.58	0.5230	.5227
	EvLowLight + EvTexture	18.46	0.6452	0.4634	14.52	0.4615	0.5545	19.32	0.5118	0.5813	17.52	0.4187	0.6312
IV	EGVSR + CoLIE	13.53	0.6844	0.3888	23.39	0.7351	0.4064	15.27	0.3096	0.4963	14.50	0.2404	0.5151
	EGVSR + Zero-IG	16.86	0.6904	0.4533	9.77	0.3957	0.4934	19.06	0.5056	0.5491	18.03	0.4607	0.5559
	EvTexture + CoLIE	9.46	0.2671	0.5767	23.12	0.7457	0.4014	15.27	0.3096	0.4963	14.43	0.2404	0.4941
	EvTexture + Zero-IG	10.85	0.3747	0.5755	9.78	0.4006	0.4680	19.06	0.5056	0.5491	18.16	0.4723	0.5337
V	BasicVSR++	25.90	0.8496	0.3481	22.87	0.7115	0.4200	19.91	0.6128	0.5123	19.44	0.5768	0.5307
	DP3DF	27.23	0.8445	0.3413	22.25	0.7299	0.4161	16.99	0.4007	0.5122	14.89	0.3523	0.5114
	MIA-VSR	15.71	0.6619	0.4863	19.57	0.6777	0.4545	19.06	0.5284	0.5492	16.82	0.4834	0.5703
	IART	23.74	0.8331	0.3699	24.15	0.7400	0.4260	19.33	0.5671	0.5008	19.82	0.6109	0.4886
	FMA-Net	27.53	0.8680	0.3300	23.93	0.7473	0.4084	20.31	0.6334	**0.4578**	19.86	0.6156	0.4709
VI	EGVSR	27.24	0.8559	0.3698	23.71	0.7483	0.4167	19.78	0.5780	0.5734	19.92	0.5716	0.5256
	EvTexture	27.33	0.8776	0.3286	24.20	0.7587	0.4166	20.29	0.6301	0.4869	19.75	0.6046	0.4977
	RetinexEVSR (Ours)	**30.28**	**0.8932**	**0.3149**	**25.15**	**0.7737**	**0.3933**	**21.24**	**0.6525**	0.4627	**20.68**	**0.6541**	**0.4382**

Table 1: Quantitative comparison on SDSD and SDE datasets for $4\times$ LVSR. All methods are retrained on the same dataset. All results are calculated on the RGB channel. **Bold** and underlined numbers indicate the best and second-best performance.

Method	Enhancement + VSR		VSR + Enhancement		Joint Enhancement and VSR						
	Retinexformer + IART	EvLight + EvTexture	MIA-VSR + FastLLVE	EGVSR + CoLIE	DP3DF	MIA-VSR	IART	FMA-Net	EGVSR	EvTexture	**Ours**
PSNR↑	15.43	17.22	19.30	18.76	27.02	24.48	26.36	27.61	26.90	28.07	**28.92**
SSIM↑	0.5861	0.6506	0.7123	0.7462	0.8406	0.8194	0.8402	0.8611	0.8473	0.8604	**0.8707**
LPIPS↓	0.5749	0.5760	0.5834	0.5073	0.4625	0.5383	0.4731	0.4633	0.5060	0.4837	**0.4612**
tOF↓	6.55	8.87	6.76	6.07	5.65	5.86	5.93	4.70	5.27	4.69	**4.60**
TCC↑$_{\times 10}$	1.50	1.26	0.80	1.88	2.88	2.40	2.85	3.17	2.96	3.14	**3.31**
Params (M)	1.61+13.41	22.73+8.90	16.60+11.11	2.58+0.13	28.86	16.60	13.41	9.62	**2.58**	8.90	8.07
FLOPs (G)	21.3+2778.4	241.5+1141.1	1755.5+87.0	226.9+8.7	775.3	1755.5	2778.4	1941.3	226.9	1141.1	**159.1**
Runtime (ms)	17.3+1666.8	50.8+126.9	1159.1+28.6	181.7+7.5	52.7	1159.1	1666.8	596.3	181.7	126.9	**44.5**

Table 2: Quantitative comparison on RELED for $4\times$ LVSR. FLOPs and runtime are computed on one 256×320 LR frame.

3 Method

RetinexEVSR Framework. We propose a novel neural network, named RetinexEVSR, to address the challenge of VSR under low-light conditions by leveraging high-contrast event signals and Retinex-inspired priors. The architecture of RetinexEVSR is illustrated in Fig. 4(a). The input consists of a LR image sequence $\{X_t^{LR}\}_{t=1}^{T}$ with T frames and the corresponding event data $\{\mathcal{E}_t^{LR}\}_{t=1}^{T}$. The network outputs a super-resolved, well-lit image sequence $\{Y_t^{SR}\}_{t=1}^{T}$.

At a given time step t, the input frame X_t^{LR} is first decomposed into illumination I_t and reflectance R_t via a Retinex-based LIE model, such as SCI (Ma et al. 2022, 2025). The illumination I_t is used to guide event feature extraction within the IEE module, producing multi-scale event features. In our implementation, we use three scales: $\{\mathcal{F}_{e_t}^{s_1}, \mathcal{F}_{e_t}^{s_2}, \mathcal{F}_{e_t}^{s_3}\}$, where s_1 corresponds to the largest spatial scale. The reflectance R_t is fed into the R-embed layer, which consists of five Residual Blocks adopted from (Wang et al. 2018), to extract the feature representation $\mathcal{F}_{R_t}$. This feature is then enhanced by events in the ERE module, yielding the enhanced reflectance feature $\mathcal{F}'_{R_t}$. Finally, features from events, illumination, and reflectance are fused to guide upsampling, producing the final output Y_t^{SR}.

From a temporal perspective, RetinexEVSR employs a bidirectional recurrent framework (Chan et al. 2021), where inter-frame optical flow serves as a *bridge* for temporal alignment and feature propagation. Unlike prior methods (Xu et al. 2023b), we compute flow from reflectance maps instead of raw inputs, as they offer higher contrast and enable more accurate alignment under low-light conditions. For example, between timestamps t and $t+1$, flows $O_{t+1\to t}$ and $O_{t\to t+1}$ are computed between R_t and R_{t+1}. In backward propagation, the feature h_{t+1} is warped to time t using $O_{t\to t+1}$ via a backward warping operation, producing the aligned feature $\tilde{h}_t$, which is then fed into the ERE module for reflectance enhancement.

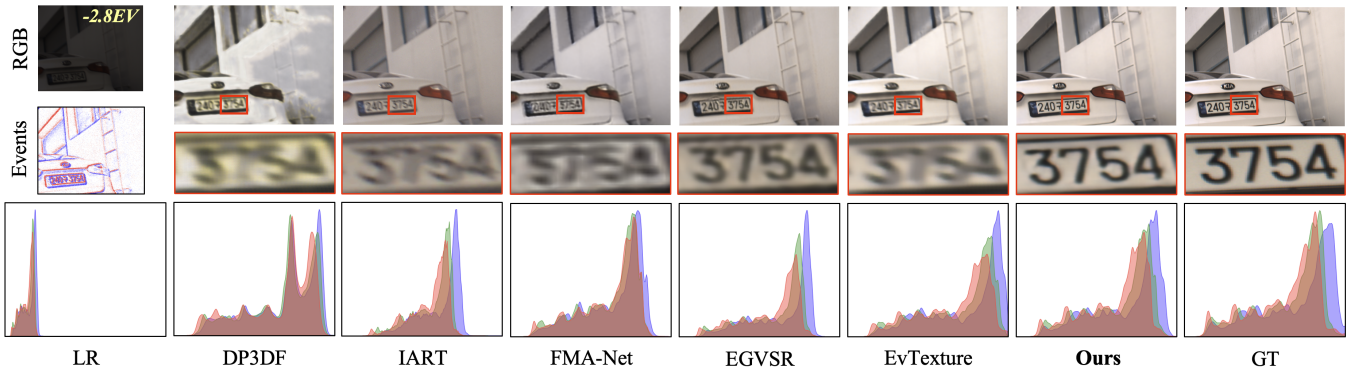

Figure 5: Qualitative comparison on RELED for 4× LVSR. The bottom row is the statistical distribution of the RGB channels. Our method recovers clearer license plate numbers and more faithful colors that better match the ground truth.

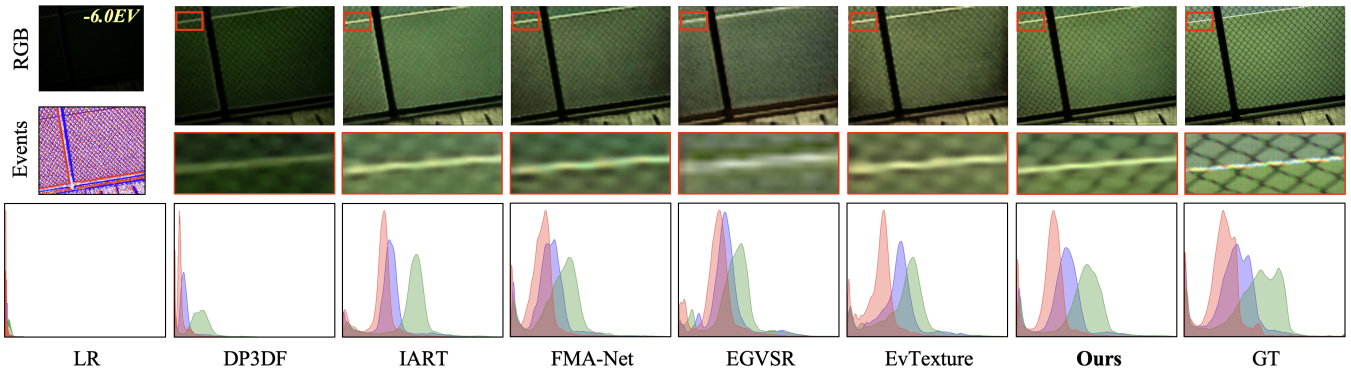

Figure 6: Qualitative comparison on SDE for 4× LVSR. Our method effectively restores well-lit images with fine details.

Illumination-guided Event Enhancement. Under low-light conditions, both event signals and RGB frames suffer significant degradation. Directly fusing them for LVSR often leads to artifacts due to the compounded noise and distortions from both modalities. To address this, we propose the IEE module, which leverages the illumination map as a global lighting prior to *light up* the event feature extraction and suppress low-light noise. As shown in Fig. 4(b), at time step t, given the illumination map I_t and LR event $\mathcal{E}_t^{LR}$, the IEE module first extracts shallow features using two symmetric branches: f_θ for illumination and f_φ for events. Both branches adopt lightweight residual blocks:

$$\mathcal{F}_{I_t} = f_\theta(I_t), \quad \mathcal{F}_{\mathcal{E}_t} = f_\varphi(\mathcal{E}_t^{LR}), \tag{1}$$

where $\mathcal{F}_{I_t}$ and $\mathcal{F}_{\mathcal{E}_t}$ denote the initial features from illumination and events. However, $\mathcal{F}_{\mathcal{E}_t}$ still suffers from trailing artifacts and noise. To refine event features, we adopt a multi-scale fusion strategy inspired by (Wang et al. 2023b). Convolutions with varying kernel sizes are used to extract features at four spatial scales: full, half, quarter, and one-eighth resolution (*i.e.*, 1, $1/2$, $1/4$, and $1/8$), enabling the network to perceive illumination-aware cues across multiple receptive fields. At each scale, illumination features guide the fusion process via channel-wise concatenation and convolution, allowing the network to recalibrate event representations based on lighting priors. The fused features are then progressively upsampled from coarse to fine in a top-down refinement pathway. At each stage, they are combined with finer-scale event features to recover spatial details while maintaining illumination consistency. We retain the top three scales after fusion as the final enhanced event features: $\{\mathcal{F}_{e_t}^{s_1}, \mathcal{F}_{e_t}^{s_2}, \mathcal{F}_{e_t}^{s_3}\}$, where s_1 corresponds to the largest spatial scale. This hierarchical strategy effectively enhances event representations under low-light conditions, providing reliable guidance for subsequent reconstruction.

Event-guided Reflectance Enhancement. In Retinex-based LIE, reflectance is commonly used as the target since it carries well-lit content and structural information. However, in the LVSR setting, it often lacks high-frequency details. To compensate for this, we propose the ERE module, which utilizes refined event features—enhanced by the IEE module—to supplement reflectance features with high-frequency cues. As illustrated in Fig. 4(c), the ERE module adopts an 'encoder–bottleneck–decoder' architecture. To incorporate temporal information, we introduce the temporally propagated feature $\widetilde{h}_t$ into the input. Additionally, event and reflectance features are dynamically fused in both the bottleneck and decoder stages through an attention-based cross-modal fusion (Li et al. 2024) block. This design enables the network to selectively inject informative structures from events into the reflectance stream while suppressing noise specific to either modality. After processing through the ERE module, the original reflectance feature $\mathcal{F}_{R_t}$ is en-

Datasets	Methods	NIQE↓	PI↓	CLIP-IQA↑	Q-Align↑
SDE-in	DP3DF	**6.8206**	7.5631	0.1540	1.3184
	IART	10.6221	9.0256	0.1889	1.3207
	EGVSR	9.0954	8.1349	**0.3063**	1.4150
	EvTexture	8.5623	7.4788	0.1993	1.2285
	Ours	7.0684	**7.2035**	0.2588	**1.6426**
SDE-out	DP3DF	7.5242	7.1929	0.1510	1.2910
	IART	7.1097	7.1529	0.1342	1.5479
	EGVSR	9.6327	8.5254	0.2537	1.5928
	EvTexture	8.0480	8.4553	0.2377	1.6209
	Ours	**6.7292**	**7.0141**	**0.2618**	**1.7432**

Table 3: Generalization to real-world SR on the SDE dataset.

Method			SDSD-in PSNR↑	SDSD-in SSIM↑	SDSD-in LPIPS↓	#Params (M)
Break -down		(a) w/o IEE	28.27	0.8642	0.3274	7.26
		(b) w/o ERE	27.31	0.8422	0.3304	6.38
	*	(c) Full Model	30.28	0.8932	0.3149	8.07
IEE		(d) $scale = 1$	28.64	0.8772	0.3313	7.28
		(e) $scale = 2$	28.83	0.8801	0.3239	7.71
	*	(f) $scale = 3$	30.28	0.8932	0.3149	8.07
ERE		(g) single-scale	28.04	0.8553	0.3325	8.05
		(h) w/o fusion	29.78	0.8911	0.3172	8.06
Retinex Model		(i) URetinex	29.62	0.8873	0.3294	8.09
	*	(j) SCI	30.28	0.8932	0.3149	8.07
Optical Flow		(k) from $\{X_t\}$	29.85	0.8908	0.3196	8.07
	*	(l) from $\{R_t\}$	30.28	0.8932	0.3149	8.07

Table 4: Ablation study of model components on SDSD-indoor. $*$ indicates the setting used in our final model.

riched with detailed textures and contrast information from the event features. The output, denoted as $\mathcal{F}^{'}_{R_t}$, also serves as the updated temporal feature h_t for the next frame, enabling continuous refinement. This enhancement not only improves the perceptual quality of the reconstructed frames but also provides stronger guidance for the final restoration. Further details about the fusion block are provided in the appendix.

Loss Function. We follow the previous study (Kai et al. 2024) and adopt the Charbonnier loss (Lai et al. 2017) as the training loss function, which is defined as:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\left\|Y_t^{GT} - Y_t^{SR}\right\|^2 + \varepsilon^2}, \tag{2}$$

where $\varepsilon = 1 \times 10^{-12}$ is set for numerical stability.

4 Experiments

Datasets. We first follow the previous LVSR method DP3DF (Xu et al. 2023b) and use the SDSD dataset (Wang et al. 2021), which provides paired low-light and normal-light videos. Since SDSD does not include event signals, we simulate events using the vid2e event simulator (Gehrig et al. 2020) with a noise model based on ESIM (Rebecq, Gehrig, and Scaramuzza 2018). We further train and evaluate our method on two real-world event datasets: SDE (Liang et al. 2024) and RELED (Kim et al. 2024). SDE contains over 30K image-event pairs captured under varying lighting conditions in indoor and outdoor scenes. RELED introduces

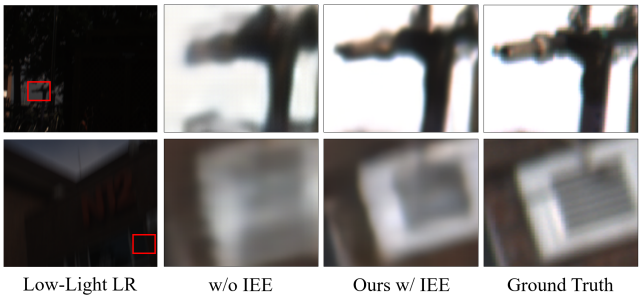

Figure 7: Ablation study of IEE. The full model produces sharper structures and finer details.

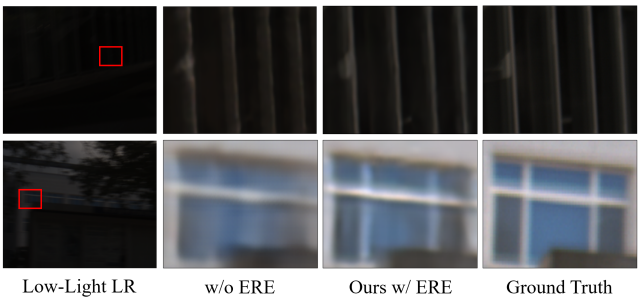

Figure 8: Ablation study of ERE. The model equipped with the ERE module can recover visually clearer results.

severe motion blur caused by long exposures in low-light, making it more challenging. Event data is converted into voxel grids (Zhu et al. 2021) with 5 temporal bins and downsampled using the same bicubic interpolation as the frames.

Implementation Details. Our model is trained from scratch on each dataset. During training, we use 15 input frames with a mini-batch size of 8 and apply center-cropping to both input frames and event voxels to a size of 64×64. Data augmentation is performed with random horizontal and vertical flips. The model is trained for 300K iterations using the Adam optimizer and Cosine Annealing learning rate scheduler. We adopt the Charbonnier loss (Lai et al. 2017) for supervision and use SpyNet (Ranjan and Black 2017) to compute optical flow. We use SCI (Ma et al. 2022) as our Retinex decomposition model. For SpyNet and SCI, the initial learning rate is 2.5×10^{-5}, frozen for the first 5K iterations. The initial learning rate for other modules is 2×10^{-4}. Training is conducted on 2 NVIDIA RTX4090 GPUs, taking about four days per dataset to converge.

Comparisons with State-of-the-Art Methods

Baselines. We compare our method with both RGB-based and event-based SOTA methods, covering two strategies: cascade and one-stage LVSR. For RGB-based VSR, we include BasicVSR++ (Chan et al. 2022), DP3DF (Xu et al. 2023b), MIA-VSR (Zhou et al. 2024), IART (Xu et al. 2024), and FMA-Net (Youk, Oh, and Kim 2024). For event-based VSR, we compare with EGVSR (Lu et al. 2023) and EvTexture (Kai et al. 2024). We also include RGB-based low-light enhancement methods: Retinexformer (Cai et al. 2023), FastLLVE (Li et al. 2023), CoLIE (Chobola et al.

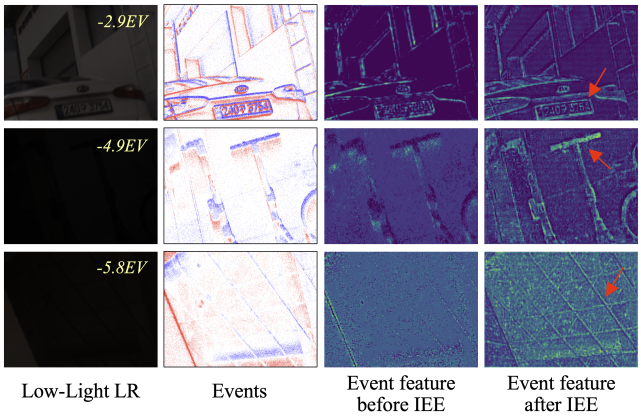

Figure 9: Event features before and after IEE. The module effectively enhances details and suppresses trailing effects.

2024), and Zero-IG (Shi et al. 2024), as well as event-based methods: EvLowLight (Liang et al. 2023) and Ev-Light (Liang et al. 2024). As shown in Tab. 1, these baselines are grouped into six categories: (I) RGB-based enhancement + VSR, (II) RGB-based VSR + enhancement, (III) Event-based enhancement + VSR, (IV) Event-based VSR + enhancement, (V) RGB-based joint enhancement and VSR, and (VI) Event-based joint enhancement and VSR. For category (IV), due to the lack of HR events after the first stage, we use CoLIE and Zero-IG as substitutes. Note that all methods are retained on the same dataset for fair comparison.

Quantitative Results. Tabs. 1 and 2 report comparisons in spatial quality (PSNR, SSIM, LPIPS), temporal consistency (tOF (Chu et al. 2020), TCC (Chi et al. 2020)), and computational cost. Our RetinexEVSR consistently outperforms all baselines across all datasets. Compared to EvTexture, it improves PSNR by **2.95**, 0.95, 0.95, 0.93, and 0.85 dB on five datasets, while reducing FLOPs by **86.1%** and runtime by 64.9%, using fewer parameters.

Qualitative Results. We also perform qualitative comparisons on these datasets. The visual results are shown in Figs. 5 and 6. It is obvious that our method enhances illumination to a well-lit level and restores textural details more accurately, while suppressing artifacts. Moreover, the histograms of the RGB channels in each figure show that our method produces color distributions more closely matching those of the ground-truth images.

Generalization to Real-world SR. Following prior VSR studies (Kai et al. 2024), our main experiments use bicubic degradation. To assess real-world generalization, we test the SDE-trained model on SDE without downsampling. NIQE (Zhang, Zhang, and Bovik 2015), PI (Blau et al. 2018), CLIP-IQA (Wang, Chan, and Loy 2023), and Q-Align (Wu et al. 2024) are used for no-reference evaluation. As shown in Tab. 3, our method achieves SOTA results on most metrics, especially on SDE-outdoor, highlighting its strong generalization to real-world low-light videos.

Ablation Study

We conduct comprehensive ablation studies on the SDSD-indoor dataset due to its good convergence and stable performance. The ablation results are summarized in Tab. 4.

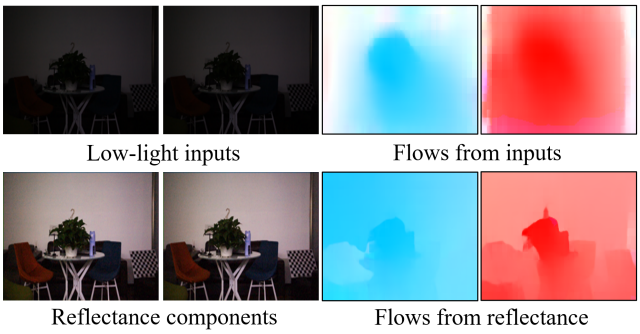

Figure 10: Analysis of optical flow calculation methods. Reflectance-based flow yields sharper edges in low light.

Break-down Ablations. Tab. 4(a-c) shows the effect of the IEE and ERE modules. The full model achieves the best performance with only a moderate increase in parameters.

The IEE Module. Tab. 4(e-f) analyzes the effect of scale in IEE. Using illumination to guide and extract event features from three scales gives the best performance, outperforming the single-scale setup by 1.64 dB. As shown in Fig. 7, the IEE module helps recover sharper structures and finer details. Fig. 9 further verifies that the IEE module effectively enhances event features by reducing trailing effects.

The ERE Module. Tab. 4(g-h) examines the impact of using single-scale and unfused features in ERE. The results show that multi-scale fusion in our full model yields notable gains. As illustrated in Fig. 8, the full model equipped with the ERE module restores clearer and sharper details.

Retinex Model. Tab. 4(i-j) compares different Retinex models. Although URetinex-Net (Wu et al. 2022) has more parameters, its supervised nature limits generalization. Our model with the unsupervised SCI (Ma et al. 2022, 2025) achieves better results, improving PSNR by 0.66 dB.

Optical Flow. Tab. 4(k-l) compares optical flow computed from either low-light frames X_t or their reflectance R_t. Using reflectance improves PSNR by 0.43 dB, thanks to clearer structures that enhance edge localization. Fig.10 shows that flow from reflectance captures sharper edges, making it more reliable for alignment in low-light scenes.

5 Conclusion

In this paper, we present RetinexEVSR, the first event-driven framework for LVSR. Our method leverages Retinex-inspired priors, coupled with a novel RBF strategy, to effectively fuse degraded RGB and event signals under low-light conditions. Specifically, it includes an IEE module that treats the illumination component, decomposed from the input frames, as a global lighting prior to enhance event features. The refined events are then utilized in the ERE module to enhance reflectance details by injecting high-frequency information. Extensive experiments demonstrate that RetinexEVSR achieves state-of-the-art performance on three datasets, including both synthetic and real-world datasets, and generalizes well to unseen degradations, highlighting its potential for low-light video applications.

Acknowledgments

We acknowledge funding from the National Natural Science Foundation of China under Grants 62472399 and 62021001.

References

Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; and Zelnik-Manor, L. 2018. The 2018 PIRM challenge on perceptual image super-resolution. In *ECCVW*.

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 *ICCV*.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*.

Chi, Z.; Mohammadi Nasiri, R.; Liu, Z.; Lu, J.; Tang, J.; and Plataniotis, K. N. 2020. All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling. In *ECCV*.

Chobola, T.; Liu, Y.; Zhang, H.; Schnabel, J. A.; and Peng, T. 2024. Fast Context-Based Low-Light Image Enhancement via Neural Implicit Representations. In *ECCV*.

Chu, M.; Xie, Y.; Mayer, J.; Leal-Taixé, L.; and Thuerey, N. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM TOG*.

Du, S.; Xia, M.; Liu, C.; Wang, X.; Wang, J.; Wan, P.; Zhang, D.; and Ji, X. 2025. PatchVSR: Breaking Video Diffusion Resolution Limits with Patch-wise Video Super-Resolution. In *CVPR*.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE TPAMI*.

Gao, J.; Yue, Z.; Liu, Y.; Xie, S.; Fan, X.; and Liu, R. 2024. A Dual-Stream-Modulated Learning Framework for Illuminating and Super-Resolving Ultra-Dark Images. *IEEE TNNLS*.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to events: Recycling video datasets for event cameras. In *CVPR*.

Han, J.; Yang, Y.; Duan, P.; Zhou, C.; Ma, L.; Xu, C.; Huang, T.; Sato, I.; and Shi, B. 2023. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE TPAMI*.

Jing, Y.; Yang, Y.; Wang, X.; Song, M.; and Tao, D. 2021. Turning frequency to resolution: Video super-resolution via event cameras. In *CVPR*.

Ju, X.; Zou, Y.; Li, X.; Wang, Z.; Ma, J.; Jiang, Z.; and Liu, J. 2025. Illumination Refinement via Textual Cues: A Prompt-Driven Approach for Low-Light NeRF Enhancement. *IEEE TCSVT*.

Kai, D.; Lu, J.; Zhang, Y.; and Sun, X. 2024. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *ICML*.

Kai, D.; Zhang, Y.; and Sun, X. 2023. Video Super-Resolution Via Event-Driven Temporal Alignment. In *ICIP*.

Kai, D.; Zhang, Y.; Wang, J.; Xiao, Z.; Xiong, Z.; and Sun, X. 2025. Event-Enhanced Blurry Video Super-Resolution. In *AAAI*.

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 *ECCV*.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*.

Li, W.; Wu, G.; Wang, W.; Ren, P.; and Liu, X. 2023. FastLLVE: Real-Time Low-Light Video Enhancement with Intensity-Aware Look-up Table. In *ACM MM*.

Li, X.; Liu, J.; Chen, Z.; Zou, Y.; Ma, L.; Fan, X.; and Liu, R. 2024. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *ECCV*.

Li, X.; Wang, Z.; Zou, Y.; Chen, Z.; Ma, J.; Jiang, Z.; Ma, L.; and Liu, J. 2025a. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *CVPR*.

Li, Z.; Liao, J.; Tang, C.; Zhang, H.; Li, Y.; Bian, Y.; Sheng, X.; Feng, X.; Li, Y.; Gao, C.; et al. 2025b. USTC-TD: A test dataset and benchmark for image and video coding in 2020s. *IEEE TMM*.

Liang, G.; Chen, K.; Li, H.; Lu, Y.; and Wang, L. 2024. Towards Robust Event-guided Low-Light Image Enhancement: A Large-Scale Real-World Event-Image Dataset and Novel Approach. In *CVPR*.

Liang, J.; Yang, Y.; Li, B.; Duan, P.; Xu, Y.; and Shi, B. 2023. Coherent event guided low-light video enhancement. In *ICCV*.

Liu, H.; Xu, J.; Chang, Y.; Zhou, H.; Zhao, H.; Wang, L.; and Yan, L. 2025a. TimeTracker: Event-based Continuous Point Tracking for Video Frame Interpolation with Nonlinear Motion. In *CVPR*.

Liu, H.; Xu, J.; Peng, S.; Chang, Y.; Zhou, H.; Duan, Y.; Zhu, L.; Tian, Y.; and Yan, L. 2025b. NER-Net+: Seeing Motion at Nighttime with an Event Camera. *IEEE TPAMI*.

Liu, Y.; Chen, Z.; Yan, H.; Ma, D.; Tang, H.; Zheng, Q.; and Pan, G. 2025c. E-NeMF: Event-based Neural Motion Field for Novel Space-time View Synthesis of Dynamic Scenes. In *ICCV*.

Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; and Wang, L. 2023. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *CVPR*.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*.

Ma, L.; Ma, T.; Xu, C.; Liu, J.; Fan, X.; Luo, Z.; and Liu, R. 2025. Learning with self-calibrator for fast and robust low-light image enhancement. *IEEE TPAMI*.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *CVPR*.

Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. ESIM: an open event camera simulator. In *CORL*.

Shi, Y.; Liu, D.; Zhang, L.; Tian, Y.; Xia, X.; and Fu, X. 2024. ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images. In *CVPR*.

Sun, C.; Zhang, J.; Wang, Y.; Ge, H.; Xia, Q.; Yin, B.; and Yang, X. 2025. Exploring Historical Information for RGBE Visual Tracking with Mamba. In *CVPR*.

Tang, Q.; Zhao, Y.; Liu, M.; Jin, J.; and Yao, C. 2024. Semantic Lens: Instance-Centric Semantic Alignment for Video Super-resolution. In *AAAI*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*.

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 *ICCV*.

Wang, T.; Zhang, K.; Shen, T.; Luo, W.; Stenger, B.; and Lu, T. 2023a. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *AAAI*.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCVW*.

Wang, Y.; Li, B.; Zhang, G.; Liu, Q.; Gao, T.; and Dai, Y. 2023b. Lrru: Long-short range recurrent updating networks for depth completion. In *ICCV*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. In *BMVC*.

Wei, S.; Li, F.; Tang, S.; Zhao, Y.; and Bai, H. 2025. EvEnhancer: Empowering Effectiveness, Efficiency and Generalizability for Continuous Space-Time Video Super-Resolution with Events. In *CVPR*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2024. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *ICML*.

Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; and Jiang, J. 2022. URetinex-Net: Retinex-based Deep Unfolding Network for Low-light Image Enhancement. In *CVPR*.

Wu, X.; Hou, X.; Lai, Z.; Zhou, J.; Zhang, Y.-n.; Pedrycz, W.; and Shen, L. 2025a. A codebook-driven approach for low-light image enhancement. *Engineering Applications of Artificial Intelligence*.

Wu, X.; Lai, Z.; Hou, X.; Zhou, J.; Zhang, Y.-n.; and Shen, L. 2025b. LightQANet: Quantized and Adaptive Feature Learning for Low-Light Image Enhancement. *arXiv preprint arXiv:2510.14753*.

Xiao, P.; Zhang, Y.; Kai, D.; Peng, Y.; Zhang, Z.; and Sun, X. 2024a. Estme: Event-driven spatio-temporal motion enhancement for micro-expression recognition. In *ICME*.

Xiao, Y.; Yuan, Q.; Jiang, K.; Chen, Y.; Wang, S.; and Lin, C.-W. 2025. Multi-Axis Feature Diversity Enhancement for Remote Sensing Video Super-Resolution. *IEEE TIP*.

Xiao, Z.; Kai, D.; Zhang, Y.; Sun, X.; and Xiong, Z. 2024b. Asymmetric Event-Guided Video Super-Resolution. In *ACM MM*.

Xiao, Z.; Kai, D.; Zhang, Y.; Zha, Z.-J.; Sun, X.; and Xiong, Z. 2024c. Event-Adapted Video Super-Resolution. In *ECCV*.

Xiao, Z.; and Wang, X. 2025. Event-based Video Super-Resolution via State Space Models. In *CVPR*.

Xie, R.; Liu, Y.; Zhou, P.; Zhao, C.; Zhou, J.; Zhang, K.; Zhang, Z.; Yang, J.; Yang, Z.; and Tai, Y. 2025. STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution. In *ICCV*.

Xu, K.; Yu, Z.; Wang, X.; Mi, M. B.; and Yao, A. 2024. Enhancing Video Super-Resolution via Implicit Resampling-based Alignment. In *CVPR*.

Xu, M.; Zhuang, C.; Lv, F.; Lu, F.; and Cloud, H. C. B. C. 2023a. Joint Low-light Enhancement and Super Resolution with Image Underexposure Level Guidance. In *BMVC*.

Xu, X.; Wang, R.; Fu, C.-W.; and Jia, J. 2023b. Deep parametric 3d filters for joint video denoising and illumination enhancement in video super resolution. In *AAAI*.

Yan, H.; Lu, Z.; Chen, Z.; Ma, D.; Tang, H.; Zheng, Q.; and Pan, G. 2025. EvSTVSR: Event Guided Space-Time Video Super-Resolution. In *AAAI*.

Yang, W.; Wu, J.; Li, L.; Dong, W.; and Shi, G. 2025. Asymmetric Hierarchical Difference-aware Interaction Network for Event-guided Motion Deblurring. In *AAAI*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; Dong, W.; and Shi, G. 2024. Learning Frame-Event Fusion for Motion Deblurring. *IEEE TIP*.

Youk, G.; Oh, J.; and Kim, M. 2024. FMA-Net: Flow-Guided Dynamic Filtering and Iterative Feature Refinement with Multi-Attention for Joint Video Super-Resolution and Deblurring. In *CVPR*.

Yue, Z.; Gao, J.; and Su, Z. 2024. Unveiling Details in the Dark: Simultaneous Brightening and Zooming for Low-Light Image Enhancement. In *AAAI*.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE TIP*.

Zhang, Z.; Ma, Y.; Chen, Y.; Zhang, F.; Gu, J.; Xue, T.; and Guo, S. 2024. From Sim-to-Real: Toward General Event-based Low-light Frame Interpolation with Per-scene Optimization. In *SIGGRAPH Asia*.

Zhou, X.; Zhang, L.; Zhao, X.; Wang, K.; Li, L.; and Gu, S. 2024. Video Super-Resolution Transformer with Masked Inter&Intra-Frame Attention. In *CVPR*.

Zhu, A. Z.; Wang, Z.; Khant, K.; and Daniilidis, K. 2021. EventGAN: Leveraging large scale image datasets for event cameras. In *ICCP*.

Zhu, L.; Yang, W.; Chen, B.; Zhu, H.; Meng, X.; and Wang, S. 2024a. Temporally Consistent Enhancement of Low-Light Videos via Spatial-Temporal Compatible Learning. *IJCV*.

Zhu, L.; Yang, W.; Chen, B.; Zhu, H.; Ni, Z.; Mao, Q.; and Wang, S. 2024b. Unrolled Decomposed Unpaired Learning for Controllable Low-Light Video Enhancement. In *ECCV*.