IEEE Open Journal of Signal Processing

Received 24 September 2025; revised 24 January 2026; accepted 2 February 2026. Date of publication 6 February 2026; date of current version 18 February 2026. The review of this article was arranged by Associate Editor Bruno Zatt.

Digital Object Identifier 10.1109/OJSP.2026.3662627

EcDiff-LLIE: Event-Conditional Diffusion Model for Structure-Preserving Low-Light Image Enhancement

RAMNA MAQSOOD, PAULO NUNES (Member, IEEE), LUÍS DUCLA SOARES (Senior Member, IEEE), AND CAROLINE CONTI (Member, IEEE)

[1]Instituto de Telecomunicações, Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal

CORRESPONDING AUTHOR: RAMNA MAQSOOD (e-mail: ramna.maqsood@lx.it.pt).

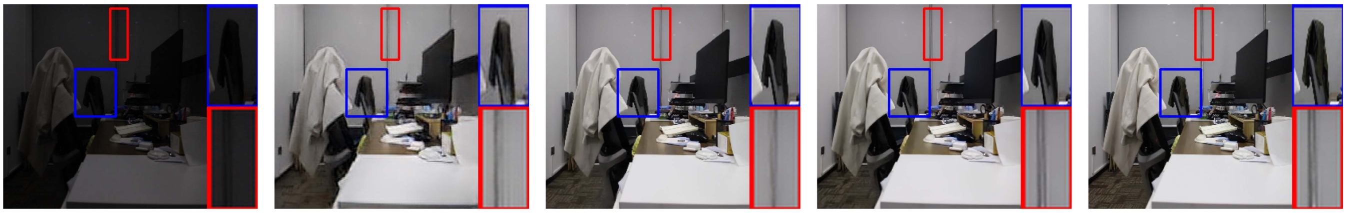
This work was supported in part by the National funds through FCT – Fundação para a Ciência e a Tecnologia, I.P. and in part by EU funds through Project/support UID/50008/2025 –Instituto de Telecomunicações, with DOI identifier 10.54499/UID/50008/2025

ABSTRACT Low-light image enhancement (LLIE) aims to restore the visual quality of poorly illuminated images by recovering fine details and textures while suppressing noise and artifacts. Recently, diffusion models have shown superior generative capabilities for LLIE. However, existing diffusion-based methods condition the denoising process only on low-light images or features derived from them (e.g., structural or illumination maps). Since the low-light images are severely degraded, this limits the denoising model's ability to restore fine structure and reduce artifacts. In this work, we show that the event data captured simultaneously with the low-light images provides complementary high-dynamic-range and high-temporal-resolution structural information that can overcome this limitation. Therefore, we propose EcDiff-LLIE, a novel event-conditional diffusion framework for LLIE. At its core, we introduce a multimodality denoising network that conditions on both low-light images and concurrent event streams. To effectively fuse the two modalities, we design a cross-modality attention block that bridge their domain differences, while also enabling long-range dependency modeling for improved structural preservation. Experiments on the synthetic SDSD and real-world SDE datasets show significant improvements in quantitative evaluation metrics. Furthermore, evaluation on the high-resolution real-world HUE dataset further shows the generalization ability of the proposed framework.

INDEX TERMS Low-light image enhancement, event camera, diffusion model, cross-modality self-attention.

I. INTRODUCTION

Low-light image enhancement (LLIE) aims to improve illumination and restore the perceptual quality of images captured under insufficient lighting. In fact, LLIE can be regarded as a denoising problem [1], [2], since enhancing the image quality is only one component of the task with the other component being noise suppression. With the rise of deep learning (DL) approaches [3], [4], many LLIE challenges have been significantly alleviated. Nevertheless, DL-based methods typically establish a direct deterministic mapping from low-light to normal-light image. While these approaches can restore brightness and details, they lack adaptive control over the degree of enhancement and may yield visually unappealing results such as over-smoothing or structural distortions due to the ill-posed nature of this problem. To address these challenges, recent efforts have turned to diffusion models (DMs) [5]. Unlike deterministic mappings, DMs rely on a hierarchical architecture of denoising autoencoders that enhance images through iterative denoising. By reversing the diffusion process step by step, DMs map randomly sampled Gaussian noise to target images. This iterative framework allows them to capture the full distribution of possible enhancements, effectively handling uncertainty and reducing artifacts. Most existing LLIE approaches [7], [8], [9] leverage DMs by conditioning the denoising process on the degraded low-light image. While this strategy



(a) Low-light Image (b) SNR-aware [6] (c) LLDiff [7] (d) **EcDiff-LLIE** (e) Normal-light (GT)

FIGURE 1. Comparison of LLIE methods: (a) Low-light image; (b) Deterministic method, SNR-aware [6] suffering from over-smoothing (blue rectangles); (c) Diffusion-based LLDiff [7] struggling to recover fine details (red rectangles); (d) Proposed framework, EcDiff-LLIE, that preserves textures and restores fine structures effectively.

improves stability, it also makes the models heavily dependent on degraded inputs. However, existing DM-based LLIE methods are still constrained by two major limitations. First, their UNet-based denoising networks [10] primarily rely on convolutional blocks, which capture local neighborhood information but struggle to model long-range dependencies, despite the demonstrated effectiveness of multi-scale designs in image restoration tasks [11], [12]. Second, preserving high-frequency (HF) structures such as textures and edges is critical for LLIE. Yet, when relying solely on degraded RGB inputs, existing methods often smooth out or misinterpret these fine structures as noise, particularly in extremely dark regions, resulting in significant detail loss (see Fig. 1).

A fundamental limitation of these methods, whether deterministic learning or DMs, is their sole reliance on the intensity information from a standard 2D camera. 2D camera struggles to capture sufficient photons under low light, resulting in detail-deficient images. To overcome this sensor limitation, a pioneering effort in this direction is EvLight [13], which introduced a large-scale dataset pairing low-light images with asynchronous event streams from an event camera. This setup involves a colocated standard 2D camera and an event camera pointing at the same scene. Event cameras operate on a fundamentally different principle, they asynchronously report per-pixel logarithmic brightness changes (events) with high temporal resolution and a high dynamic range. This allows them to capture fine details even in conditions where a standard camera fails. While EvLight demonstrated the promise of events for LLIE, the key challenge lies in how to integrate them with intensity images in a meaningful way. Naive strategies such as concatenation or early fusion are ineffective, since events and images follow very different data distributions [14]. At the same time, DMs are not naturally suited to handle such heterogeneous inputs, as they were designed primarily for dense image data.

To overcome these limitations, we propose EcDiff-LLIE, a novel event-conditional diffusion framework that utilizes both local and global contextual information as well as complementary modality information for the LLIE task. Our core insight is that the event stream provides fine details, which can be used to explicitly guide the denoising process of DMs, preventing it from smoothing over fine details. Unlike existing diffusion-based LLIE methods that rely solely on learned priors or auxiliary structural cues [15], [16], we introduce event conditional encoders and decoders to explicitly inject fine-grained structural information into the denoising process. Specifically, the core of our framework is a multimodality denoising network (MMDN). Its design is motivated by two fundamental challenges. First, effectively fusing information from the dense intensity image and the sparse asynchronous event stream is not straightforward. Second, event data is rich in HF details but requires specialized processing to extract adequate scene features without introducing noise. To address the former, we design a cross-modality attention block (CMAB) that combines ResNet-based local feature extraction with transformer-based self-attention (SA), enabling the network to integrate event features with 2D image features while capturing rich local and global context. To address the second challenge, we introduce a lightweight feature enhancement block (FEB) that processes the sparse event stream efficiently before conditioning the encoders and decoders stages of our denoising network. Our contributions can be summarized as follows:

- We address the problem of LLIE with the aid of event data and, for this, we propose a novel framework EcDiff-LLIE. To the best of our knowledge, this is the first work to leverage asynchronous event data to explicitly enhance structural and fine details during the denoising process.
- We introduce a novel multimodality denoising network that jointly extracts local and global features. Within this network, we design (i) a cross-modality attention block to effectively fuse dense image features with event features, and (ii) a lightweight feature enhancement block to efficiently process sparse event data.

II. RELATED WORK

Recently, deep learning has achieved great success in various restoration tasks, such as dark video understanding [17], [18], image deblurring [19], snow removal [20] and light-field imaging enhancement [21]. Existing approaches can be broadly categorized into deterministic and non-deterministic methods. Deterministic methods for LLIE can be broadly divided into single-modality approaches and multi-modal approaches. Single-modality methods rely solely on standard 2D images. Retinexformer [6] operates on the principle of decomposing an image into illumination and reflectance maps. Retinexformer advanced this concept with a single-stage architecture that directly estimates and refines these maps within an end-to-end trainable model. SNR-Aware [4] introduced a signal-to-noise-ratio (SNR) aware network to dynamically adjust enhancement intensity per pixel. Another method, Uformer [22], applied a U-shaped hierarchical transformer to capture long-range dependencies for effective image restoration. These methods, however, are inherently limited by the information contained in low-light images and struggle to recover fine textures in extremely dark regions. To overcome the limitations of single-modality inputs, several recent methods incorporate additional sensors, particularly event cameras. For example, eSL-Net [3] integrates events through a spatially-adaptive learning framework for efficient enhancement, while EvLight [13] uses a multi-scale holistic fusion branch combined with SNR-guided regional feature selection to adaptively merge image and event information.

Diffusion Models for LLIE: Denoising diffusion probabilistic models (DDPMs) [5] have emerged as a powerful framework for generative modeling and image restoration [23]. They operate through a forward process that incrementally adds noise to a clean image, and a reverse process where a UNet denoiser [10] is trained to iteratively recover the high-quality image. While vanilla DDPMs are unconditional, conditional variants [24] have been developed to guide the generation for specific tasks such as LLIE, by incorporating auxiliary information. For instance, LLDiff [7] utilizes color maps and latent degradation representations derived from low-light images to steer the diffusion process. DiffLL [8] employs a wavelet-based conditional DM to enhance fine details. Transdiff [15] conditions DMs with a structure prior map extracted by the advanced edge detection network LDC [25]. A key limitation of these approaches is their dependence on conditioning signals derived either from the low-light image itself or from hand-crafted prior models. Consequently, they remain constrained by the information loss inherent in the low-light input domain. To the best of our knowledge, no existing work has effectively integrated event data as a conditioning signal for diffusion-based LLIE. In this work, we bridge this gap with our EcDiff-LLIE, a novel framework that conditions the diffusion process not only on the low-light image stream but also on the concurrent event stream.

III. PROPOSED METHODOLOGY

The overall pipeline of our proposed event-conditional diffusion framework, termed EcDiff-LLIE, is illustrated in Fig. 2. Our goal is to restore the visual quality of images captured under poor illumination by learning the conditional distribution $p(x_0|x_E, x_L)$ that maps a low-light image x_L and its corresponding voxelized event stream x_E to a high-quality output. While built upon the DDPM [5], our core contribution is a novel MMDN. This network is specifically designed to effectively condition the reverse denoising process on both x_L and x_E, overcoming the modality gap to restore fine details and enhance visual quality.

Algorithm 1: Training on data.

```
Input: Normal-light image, $x_0$, low-light image, $x_L$, event
       voxel grid, $x_E$; total timesteps, $T$
Output: Optimized parameter $\theta$
while not converged do
    Sample a training pair $(x_0, x_E, x_L)$
    $t \sim \text{Uniform}(\{1, \dots, T\})$
    $\epsilon \sim \mathcal{N}(0, I)$
    $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$
    $\mathcal{L}_{DM} = \left\| \epsilon - \epsilon_\theta(x_t, x_E, x_L, t) \right\|_2^2$
    Perform gradient descent steps on $\nabla_\theta \mathcal{L}_{DM}$
end
return $\theta$
```

Algorithm 2: Sampling on data.

```
Input: Low-light image $x_L$, event voxel grid $x_E$; total
       timesteps $T$, and the number of implicit sampling
       step $S$
Output: Restore clean image $\hat{x_0}$
$\hat{x_T} \sim \mathcal{N}(0, I)$; for $i = S : 1$ do
    $t = (i-1).T/S + 1$
    $t_{next} = (i-2).T/S + 1$ if $i > 1$, else $0$
    $\hat{x}_t \leftarrow \sqrt{\bar{\alpha}_{t_{next}}} \left( \frac{\hat{x}_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_{\theta(\hat{x_t}, x_E, x_L, t)}}{\sqrt{\bar{\alpha}_t}} \right)$
            $+ \sqrt{1-\bar{\alpha}_{t_{next}}}\, \epsilon_{\theta(\hat{x_t}, x_E, x_L, t)}$
end
return $\hat{x_0}$
```

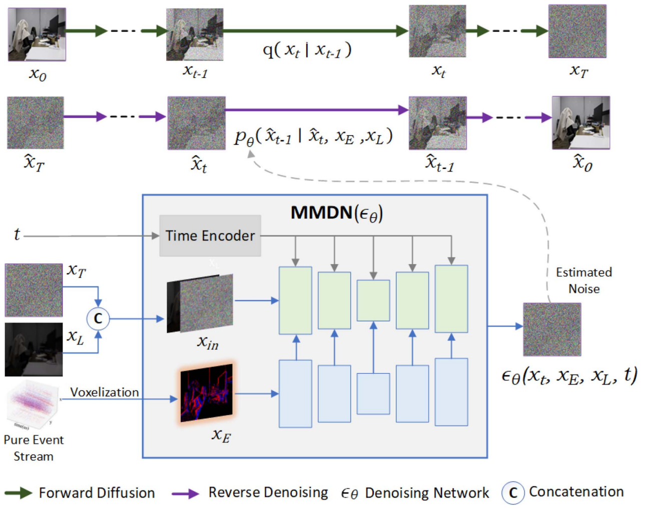


FIGURE 2. Overall pipeline of proposed EcDiff-LLIE.

Forward Diffusion: The forward process is a fixed Markov chain that gradually adds Gaussian noise to a ground-truth (GT) normal-light image x_0 over T timesteps. Given x_0, the forward process q produces a sequence of increasingly noisy latents $x_1, x_2, \dots, x_T$, where $x_T \sim \mathcal{N}(0, \mathbf{I})$ is nearly pure noise. Each step in the chain is defined as (1):

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

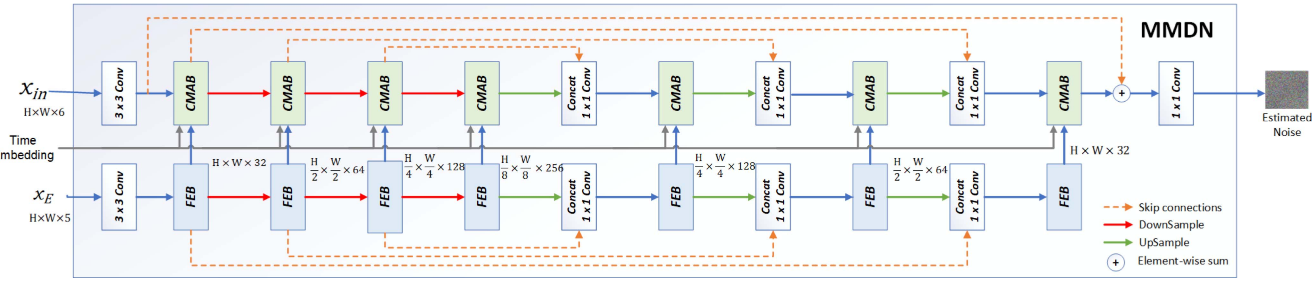

MMDN
x_{in}
H×W×6
Time Embedding
x_E
H×W×5
3 x 3 Conv
3 x 3 Conv
CMAB
FEB
H × W × 32
$\frac{H}{2} \times \frac{W}{2} \times 64$
$\frac{H}{4} \times \frac{W}{4} \times 128$
$\frac{H}{8} \times \frac{W}{8} \times 256$
Concat 1 x 1 Conv
$\frac{H}{4} \times \frac{W}{4} \times 128$
$\frac{H}{2} \times \frac{W}{2} \times 64$
H × W × 32
1 x 1 Conv
Estimated Noise
Skip connections
DownSample
UpSample
Element-wise sum

FIGURE 3. Detailed illustration of the proposed MMDN.

Here, $\beta_t \in (0, 1)_{t=1}^T$ is a pre-defined variance schedule controlling the amount of noise added at each step, T is the total number of diffusion steps, and $\mathbf{I}$ represents the identity matrix. A key property of this process is that we can directly sample x_t for any arbitrary timestep t in closed form as (2):

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}\right) \tag{2}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$.

Reverse Denoising: The restoration task is achieved by reversing the forward process. Starting from pure random Gaussian noise $\hat{x}_T \sim \mathcal{N}(0, \mathbf{I})$, the goal is to iteratively denoise it over T steps to generate a clean image $\hat{x}_0$. In conditional DMs [24], this reverse process is learned by a neural network parameterized by θ and guided by conditioning inputs. In our framework, the reverse process is conditioned on both the low-light image x_L and the event data x_E. Thus, at each step t, our model learns the conditional transition as (3):

$$p_\theta(\hat{x}_{t-1} \mid \hat{x}_t, x_E, x_L) = \mathcal{N}\left(\hat{x}_{t-1}; \mu_\theta(\hat{x}_t, x_E, x_L, t), \sigma_t^2\mathbf{I}\right) \tag{3}$$

Following [5], we set the variance σ_t^2 to $\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ and train our model to predict the noise added during the forward process. The mean of the reverse distribution μ_θ is then derived from this prediction via (4). The pseudo-code for training is shown in Algorithm 1.

$$\mu_\theta(\hat{x}_t, x_E, x_L, t) = \frac{1}{\sqrt{\alpha_t}}\left(\hat{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\hat{x}_t, x_E, x_L, t)\right) \tag{4}$$

In the training phase, the goal of the MMDN is to optimize the parameters θ of the network ϵ_θ to promote the estimated noise vector $\epsilon_\theta(x_t, x_E, x_L, t)$ close to Gaussian noise like [5], which is formulated as (5):

$$\mathcal{L}_{DM} = ||\epsilon - \epsilon_\theta(x_t, x_E, x_L, t)||_2^2 \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the sampled noise at time t. Note, we adopt the standard DDPM noise prediction loss as the sole training objective. This loss is fundamental to DMs and is required to learn the reverse denoising process. During inference, while standard DDPM requires many iterative network predictions (typically $T = 1000$ steps), making the reverse inference process computationally expensive, we instead adopt denoising diffusion implicit models (DDIM) [26] for efficient sampling as demonstrated in Algorithm 2.

Multimodality Denoising Network: The cornerstone of EcDiff-LLIE is the design of ϵ_θ, which we name MMDN. Overall training pipeline is shown in Fig. 3. This process is crucial since the quality of the generated samples depends directly on the denoiser's ability to restore details and preserve structural consistency during iterative denoising. In existing approaches [27], [28], the denoiser is conditioned only on low-light images, which limits its ability to recover fine structures, especially under low-light conditions. To address this, recent work [15] has attempted to provide structural priors into the denoising network, where edges are explicitly provided as auxiliary guidance. However, such approaches rely on handcrafted or pre-computed priors that may be unreliable in highly degraded scenarios. In contrast, we design an MMDN utilizing event data as an additional sensing modality. By conditioning both the encoder and decoder stages of the denoising network on events, our method captures fine details of the scene in low-light scenarios.

The fusion of event data and images presents two primary challenges. First, a significant modality gap exists: event streams represent asynchronous brightness changes (a differential signal), while images capture absolute intensity (an integral signal). Second, event data is inherently sparse and noisy, yet rich in high-temporal-resolution details. Directly conditioning the ϵ_θ with event data alongside low-light images is ineffective, as demonstrated in our ablation studies (Section V). To address this, our MMDN processes events in a dedicated parallel pathway through FEBs. These enhanced event features are then fused with image features at corresponding scales via CMABs. The CMAB is specifically designed to bridge the modality gap by allowing image features to adaptively query and attend to relevant information from the event features.

Our proposed MMDN follows a UNet encoder-decoder structure to enable multi-scale processing. Features are downsampled (indicated by red arrows in Fig. 3) via strided convolutions and upsampled (green arrows) via transposed convolutions. More specifically, in the event path, we obtain the event voxel grid $x_E \in \mathbb{R}^{H\times W\times B}$ following [29] by assigning the polarity of each event to the two closest voxels. Here, H and W denote height and width, respectively, and B refers to the temporal bin. We set $B = 5$ based on experiments with $B = 3, 5, 10, 15$, where larger values provide no further gain while increasing computational cost due to sparser event distribution per bin. In the image path, the noisy image, x_t, and its low-light, x_L, counterpart are concatenated to form $x_{in} \in \mathbb{R}^{H\times W\times 6}$. After that, image features, $F \in \mathbb{R}^{H\times W\times 32}$ of

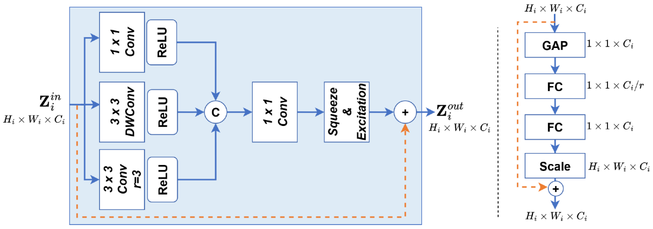


FIGURE 4. **Left: Detailed diagram of the FEB showing the flow of features. Right: Architecture of the SE layer used within FEB.**

x_{in}, and event features, $Z \in \mathbb{R}^{H \times W \times 32}$, of event voxel grid, x_E, are extracted with a 3×3 convolution (Conv) layer.

Feature Enhancement Block: Since the event voxel grid is inherently sparse and noisy, directly fusing it with image features can lead to misalignment and degraded representations. To mitigate this, we propose FEB to transform the x_E into a robust, multi-scale feature map before fusion. As shown in Fig. 4, given an input feature map Z_i^{in}, FEBs capture multi-scale spatial-channel information while maintaining computational efficiency. The output of the i-th FEB is denoted as $\mathbf{Z}_i^{\text{out}} = \text{FEB}_i(\mathbf{Z}_i^{\text{in}})$, where $\mathbf{Z}_i^{\text{in}}, \mathbf{Z}_i^{\text{out}} \in \mathbb{R}^{H \times W \times C}$, and $i = 1, \ldots, N$ indexes the FEBs in the network. Each FEB processes its input through three parallel Conv paths: a 1×1 Conv to aggregate the information in the channel direction, a 3×3 depthwise Conv (DWConv) to model fine local details such as edges and textures, and a 3×3 dilated Conv to capture long-range contextual information. The outputs are concatenated along the channel dimension, fused back to C channels with a 1×1 Conv, and further refined using a Squeeze-and-Excitation (SE) block [30]. The SE block performs global average pooling (GAP) across each channel to generate a channel descriptor, followed by a lightweight bottleneck with a reduction factor $r = 16$ to reduce parameters. A subsequent fully-connected (FC) layer and sigmoid activation produce channel-wise scaling factors, which are applied to the input features to adaptively recalibrate channel responses. This allows the SE block to emphasize informative channels and suppress less useful ones. Finally, this design allows the FEB to provide multi-scale, context-aware features that are particularly suited for representing sparse and noisy event data, thereby supplying the CMABs with rich features.

Cross-modality Attention Block: As shown in Fig. 5, the CMAB fuses image features F_i with enhanced event features Z_i^{out} through a cross-attention mechanism. The input image feature map, F_i, is first processed by a ResNet block that consists of two Conv layers with group normalization and nonlinear activations. To incorporate global context, we first apply layer normalization to both the events, Z_i^{out}, and image features, F_i. F_i is projected into a query, $Q = W_Q F_i$, while Z_i^{out} is projected into a key, $K = W_K Z_i^{\text{out}}$, and a value, $V = W_V Z_i^{\text{out}}$. Here, the image features query the event features to retrieve relevant complementary information. The projection matrices W_Q, W_K, and W_V are implemented as a combination of point-wise convolution (to mix channel information) and depthwise convolution (to preserve local spatial context). This design allows the projections to aggregate both channel-wise interactions and local spatial context, which is crucial for aligning the two modalities. The timestep embedding t is processed through a FC layer and incorporated into CMAB to enable noise-aware cross-modality attention. Since image features are corrupted by diffusion noise while event features provide structural information, timestep conditioning allows the attention mechanism to adaptively modulate the correlation between noisy image Q and event K across diffusion steps. Cross-attention is then computed across the direction of the layers as in (6)

$$\hat{F}_i = V_i \cdot \text{Softmax}\left(\frac{K_i \cdot Q_i}{d}\right) + F_i, \tag{6}$$

where d is a learnable scaling parameter. As in standard multi-head attention [31] blocks, channels are split into multiple heads to compute attention maps. We use the CMAB at multiple levels of MMDN (as shown in Fig. 3) to fuse event information aggregated across receptive fields of varying size. Following [9], our feed-forward network consists of separable point-wise Conv and DWConv layers to maximize computational efficiency. Given a feature $\hat{F}_i$ after layer normalization, the output feature map can be expressed as (7):

$$F_i^{\text{out}} = \hat{F}_i + \text{Conv}\left(\text{G}\left(\mathcal{F}(\text{Conv}(\hat{F}_i))\right) \odot \mathcal{F}(\text{Conv}(\hat{F}_i))\right) \tag{7}$$

where $\mathcal{F}$ denotes DWConv layers, G denotes GELU activation function and $\odot$ is element-wise multiplication.

IV. EXPERIMENTAL SETUP

Datasets:

SDE – *Spatio-temporally aligned dual-modality event-image dataset* [13]: This dataset provides paired low-light images and events, together with corresponding normal-light images, captured using a DAVIS346 event camera (346 × 260 resolution). It consists of 91 paired image–event sequences (43 indoor and 48 outdoor). Following [13], we use 76 sequences for training and 15 for testing. We focus on the low-light image–event pairs, as this setting directly reflects the practical challenge of enhancing degraded low-light inputs while utilizing the complementary event modality to recover structural details that cannot be obtained from images alone.

SDSD – *Seeing dynamic scene in the dark* [32] dataset: This dataset provides paired low-/normal-light video sequences at 1080 × 1920 resolution. Following [13], we first downsampled the videos to 346 × 260 to match the DAVIS346 resolution and then input the resized images to the video-to-event simulator [33] under the default noisy model. For a fair comparison, we follow the same train/test split (70 indoor and 80 outdoor) as in [13], [32].

Implementation Details: Our EcDiff-LLIE is implemented in PyTorch on an RTX 4080 GPU. We train our proposed model using the Adam optimizer. The initial learning rate is set to 1×10^{-3} and decays by a factor of 0.8 every 5×10^5 iterations to ensure stable convergence. For both SDE and SDSD, we adopt the same training strategy. On the SDSD dataset, our model converges after approximately $350k$

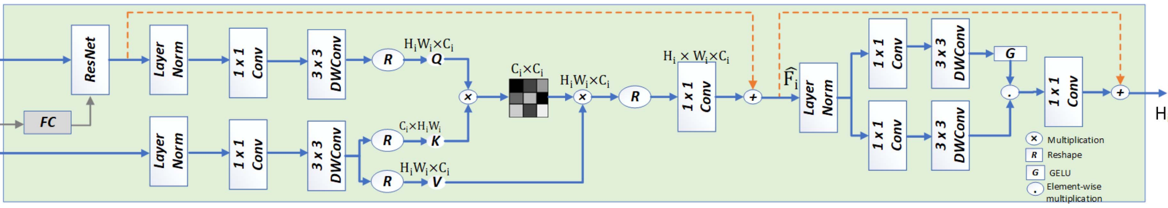


FIGURE 5. Illustration of each component in the CMAB.

TABLE 1. Results on Synthetic SDSD and Real-World SDE Datasets. Here **Bold** Values Indicate Best and underlined Second-Best Performing Models. Here * Indicates the Absence of the Events.

Input	Method	SDSD_IN			SDSD_OUT			SDE_IN			SDE_OUT		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Event-only	ET-Net (ICCV'21)	16.37	0.665	0.547	18.19	0.544	0.588	16.37	0.609	0.601	16.80	0.560	0.609
Image-only	Uformer (CVPR'22)	24.28	0.885	0.158	22.67	0.815	0.188	20.34	0.728	0.215	22.32	0.725	0.208
	SNR-Net (CVPR'22)	23.01	0.836	0.171	21.48	0.798	0.202	20.90	0.637	0.239	22.12	0.612	0.238
	Retinexformer (ICCV'23)	22.56	0.859	0.169	24.06	0.821	0.182	20.90	0.678	0.229	22.60	0.678	0.212
Event+Image	eSL-Net (ECCV'20)	23.98	0.875	0.164	23.58	0.801	0.189	21.71	0.722	0.209	22.74	0.705	0.206
	EvLight (CVPR'24)	24.40	0.892	0.145	26.67	0.835	0.157	22.04	0.735	0.197	22.80	0.731	0.186
Diffusion-based	DiffLL* (ACM'23)	23.98	0.886	0.155	24.89	0.835	0.169	22.01	0.698	0.210	20.34	0.707	0.216
	LLDiff* (PR'25)	25.68	0.875	0.150	25.35	0.836	0.161	**23.56**	0.742	0.171	21.35	0.721	0.201
	EcDiff-LLIE	**28.49**	**0.918**	**0.111**	**28.15**	**0.893**	**0.136**	23.22	**0.744**	**0.167**	**23.08**	**0.752**	**0.177**

iterations, whereas on the SDE dataset, convergence requires nearly 400k iterations. The batch size and patch size for both datasets were set to 6 and 200 × 200, respectively, and we adopt a cosine noise schedule for the variance in the diffusion process. We apply random cropping and horizontal and vertical flipping data augmentation. The total timesteps, T, and sampling steps, S, were set to 500 and 30 for the forward diffusion and reverse denoising processes, respectively.

Evaluation Metrics: To quantitatively assess the performance of our method, we employ three widely used full reference image quality metrics: peak signal-to-noise ratio (PSNR), structural similarity index metric (SSIM), and learned perceptual patch similarity (LPIPS).

Comparison Methods: We compare EcDiff-LLIE with state-of-the-art (SOTA) methods with publicly available code under four input settings to evaluate the impact of image-event modalities: (1) events only, ET-Net [34], (2) RGB images only, including Uformer [22], Retinexformer [6], and SNR-Net [4], (3) both RGB images and events, including EvLight [13] and eSL-Net [3], and (4) diffusion-based models conditioning only on low-light image, including DiffLL [8] and LLDiff [7]. Restored images are compared against normal-light GT using evaluation metrics.

Table 1 shows that our proposed EcDiff-LLIE achieves SOTA performance across all comparison methods on both datasets. Specifically, EcDiff-LLIE outperforms the previous best diffusion-based method, LLDiff, by 2.81 dB and 2.80 dB in PSNR on SDSD_IN and SDSD_OUT, respectively, and achieves an improvement of 4.29% and 4.14% in SSIM. On the real-world SDE dataset, while PSNR lags, our method consistently improves SSIM and achieves the lowest LPIPS values, indicating better structural and perceptual quality. This behavior is largely due to dataset characteristics as SDE images exhibit partial chromatic aberrations and Moiré patterns [13], which introduce small pixel-level misalignments and color distortions that affect distortion-based metrics like PSNR more than perceptual metrics. These results demonstrate that, even when PSNR is lower, our method provides visually faithful and structurally consistent restorations.

To better illustrate the performance of our method, we present qualitative comparisons on the SDE test sets in Fig. 6. The first two rows depict a challenging indoor low-light scenario. While all SOTA methods fail to recover scene details and suffer from extreme noise and artifacts, our method successfully restores a clean image.

Visual results on the SDSD dataset are shown in Fig. 7. As expected, event-only methods perform poorly. Image+event fusion methods restore images that are often blurred or oversmoothed. DMs without event conditioning correct the global brightness but fail to recover fine details. In contrast, our method restores clear edges and texture patterns, as highlighted in the blue rectangles. These results highlight the critical role of event-based supervision for LLIE and the effectiveness of conditioning DMs on event streams.

Generalization Ability: To assess the generalization capability of the proposed EcDiff-LLIE, we conduct experiments on the real-world high-resolution HUE dataset [35], which does not provide normal-light GT images. We evaluate the proposed method using models trained on both the synthetic SDSD and the real SDE datasets to analyze the impact of synthetic training on real-world data. Table 2 and Fig. 8 show that our method achieves superior performance in terms of the

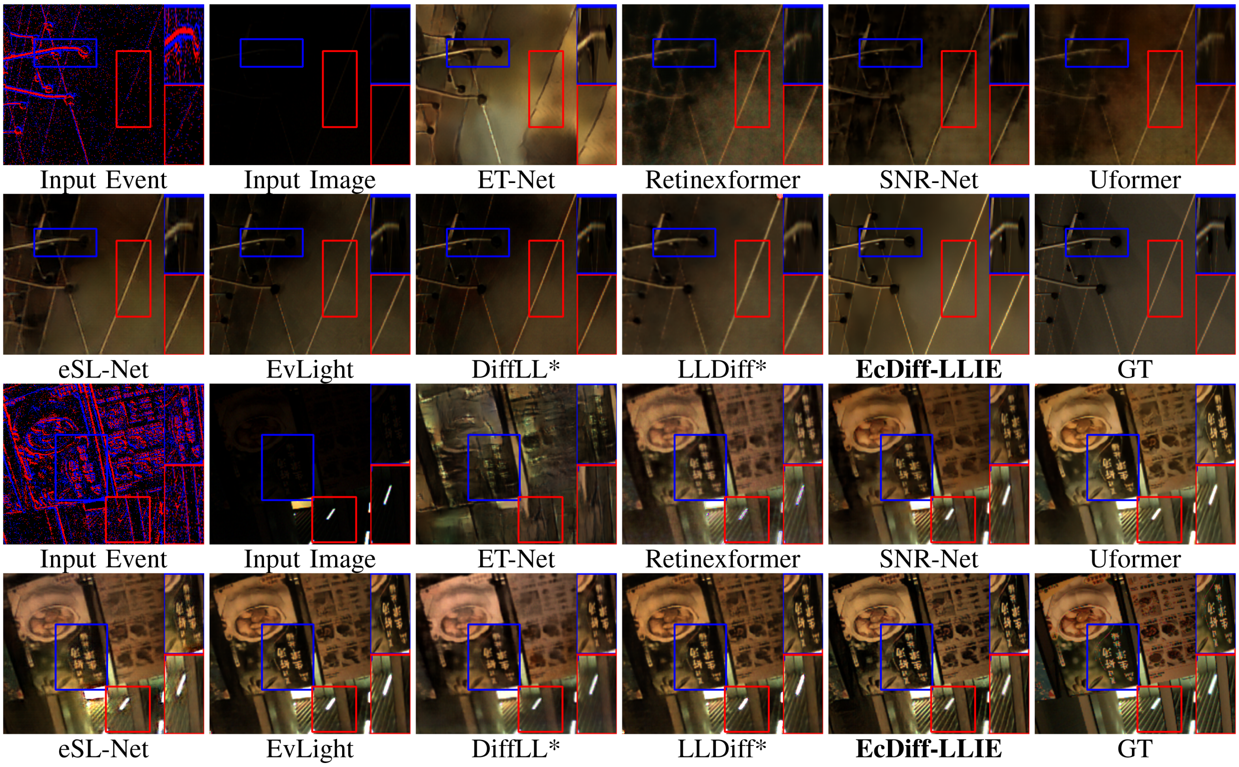


FIGURE 6. Comparison of different methods on real SDE dataset.

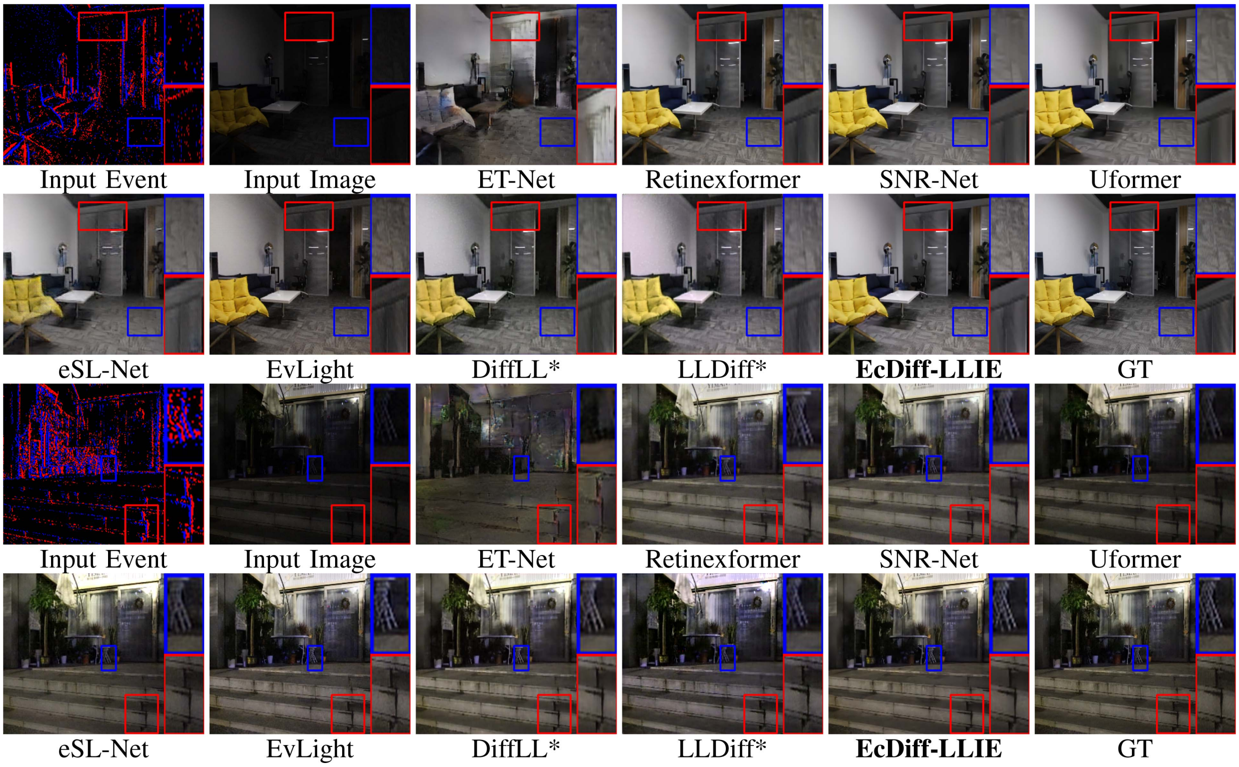


FIGURE 7. Comparison of different methods on synthetic SDSD dataset.

TABLE 2. **Generalization and Inference Efficiency of Diffusion-Based LLIE Methods. We Report BRISQUE, Inference Time Per Image (seconds), and Model Parameters (millions).**

Method	Params(M)↓	BRISQUE↓		Time (s)↓	
		512×512	1280×720	512×512	1280×720
DiffLL	22.08	12.89	12.97	0.47	1.18
LLDiff	208.71	24.09	26.78	1.18	3.89
EcDiff-SDE	**8.25**	**7.28**	**9.52**	**0.47**	**1.08**
EcDiff-SDSD	**8.25**	**7.43**	**9.67**	**0.47**	**1.08**

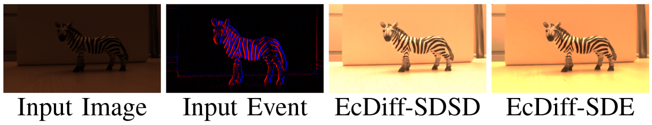

Input Image Input Event EcDiff-SDSD EcDiff-SDE

FIGURE 8. **Visual results on HUE dataset at a 1280 × 720 resolution. EcDiff-SDSD denotes results produced by the model trained on the SDSD dataset, while EcDiff-SDE denotes results from the model trained on the SDE dataset.**

TABLE 3. **Ablation Study in Terms of Metrics, Number of Parameters in Millions, and Inference Time in Seconds**

Settings	PSNR ↑	SSIM ↑	Params (M) ↓	Time (s) ↓
only-image conditioning	26.44	0.865	8.15	0.167
Baseline	26.70	0.870	7.95	0.161
MMDN w/o CMA	26.96	0.876	8.25	0.170
MMDN w/o SE	28.04	0.908	8.22	0.169
EcDiff-LLIE	28.94	0.918	8.25	0.170

no-reference image quality evaluation metric, BRISQUE, and the inference time, compared to benchmark diffusion-based LLIE methods.

V. ABLATION STUDY

We conduct an ablation study to measure the impact of proposed MMDN both in terms of evaluation metrics and computational complexity. The training details for all models are identical to those of EcDiff-LLIE and we evaluate their performance on SDSD_IN test set. Table 3 reports quantitative results together with model complexity in terms of parameters in million (M) and inference time (in seconds per image) on an RTX 4090 GPU with input size 346 × 260.

Image-only conditioning: In this experiment, we remove the event path from MMDN and condition the network only on low-light image x_L. In this setting, CMABs compute the Q, K and V solely from intensity input x_{in}. As shown in Table 3, MMDN conditioning only on low-light image results in a decrease in PSNR and SSIM compared with EcDiff-LLIE that includes event path in MMDN. Notably, the baseline remains competitive with many image-only methods [4], [6], showing that conditional DMs are inherently powerful.

Baseline: To examine whether the performance gain arises from the proposed MMDN or merely from conditioning a DM on event data, we introduce an event-conditioned diffusion *Baseline*. In this setting, the proposed MMDN is entirely removed. Instead, an encoder is used to extract features from the x_E, which are then provided to the denoising U-Net through a standard cross-attention conditioning mechanism commonly adopted in conditional DMs [36]. As reported in Table 3, this baseline outperforms the image-only DM, confirming the usefulness of event information. However, it remains consistently inferior to EcDiff-LLIE, indicating that generic event conditioning is insufficient to fully exploit the characteristics of events and intensity images.

MMDN w/o cross-modality attention (CMA): In this variant, we introduce the event path and process x_E through FEBs. However, instead of using the cross-modality attention mechanism, we adopt a simpler strategy. Specifically, we concatenate the image and event features and then apply SA (6) on the fused representation, where Q, K, and V are all derived from the concatenated features at each scale. As shown in Table 3, this model performs only slightly better than the *Baseline,* indicating that the event features are largely ignored. This occurs because the SA mechanism treats the concatenated features as a single, homogeneous input. Without a dedicated cross-attention mechanism, the richer modality (image) tends to dominate the fused space, so event features become underutilized.

MMDN w/o SE: This model keeps the event path but removes the SE layer from FEB, fusing the three parallel convolutional paths using only a final 1 × 1 Conv. This non-adaptive fusion leads to a significant performance drop of 0.90 dB in PSNR and 0.010 in SSIM compared to the complete EcDiff-LLIE framework. This confirms that the SE ability to adaptively recalibrate channel weights is essential for emphasizing informative event features and suppressing noise. Crucially, the SE layer provides this substantial performance boost with a negligible computational overhead. As shown in Table 3, the parameter count and inference time for the w/o SE model (8.22 M, 0.1694 s) are nearly identical to those of the full EcDiff-LLIE model (8.25 M, 0.1708 s). This shows that the FEB, even with the SE layer, is extremely lightweight and efficient for processing event data, adding minimal cost for a major gain in restoration quality.

Noise Robustness Comparison: To further assess the robustness of EcDiff-LLIE, we adopt a commonly used noise robustness evaluation protocol in the literature, in which Gaussian noise is added to the x_E to simulate degraded conditioning. Specifically, zero-mean Gaussian noise with a standard deviation of 0.02 is used as the reference baseline. A scaling coefficient α is then applied to the x_E standard deviation σ_E to control the noise strength, resulting in the noisy event representation as in (8):

$$x_E^{\text{noisy}} = x_E + \mathcal{N}\big(0, (\alpha \times \sigma_E)^2\big) \tag{8}$$

Varying α allows systematic control over the magnitude of the injected noise. As shown in Table 4, even under relatively strong noise ($\alpha = 1.0$), the restoration quality of EcDiff-LLIE slightly degrades, demonstrating that the proposed model remains stable under noisy event conditioning signals.

TABLE 4. Performance Evaluation Under Degraded Condition

	SDE_OUT			
Metric	α (0.02)	$\alpha(0.1)$	$\alpha(0.5)$	$\alpha(1.0)$
LPIPS ↓	0.1789	0.1840	0.1896	0.2080
SSIM ↑	0.7498	0.7434	0.7375	0.7083

TABLE 5. Ablation Study on Sampling Steps

	SDSD_IN			
Metric	S=10	S=30	S=40	S=50
LPIPS ↓	0.198	0.111	0.110	0.121
Time (s) ↓	0.105	0.170	0.367	0.577

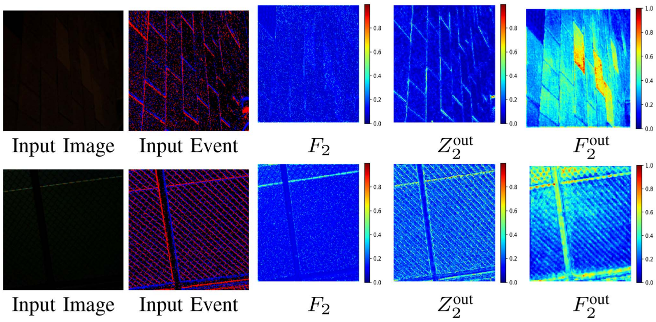


FIGURE 9. Feature visualization from the second CMAB block. Color scale indicates activation strength. The event stream provides complementary structural information that enriches the final fused features. Visualizations are zoomed for clarity.

Effect of Sampling Steps: We analyze the impact of sampling steps on image quality and inference time in Table 5. Increasing the number of steps significantly improves LPIPS from 10 to 30 steps, while further increasing the steps yields marginal or inconsistent gains and even slight degradation at 50 steps. Meanwhile, inference time increases substantially from 0.1708 s (30 steps) to 0.5770 s (50 steps). These results indicate that 30 sampling steps provide a favorable trade-off between image quality and computational cost.

Visualization of Network Elements: To validate the internal working mechanism of our proposed framework, we visualize the feature representations at the second CMAB block in Fig. 9. The heatmaps represent spatial feature energy, computed as the mean absolute activation across all channels. The image-only features F_2 extracted from the degraded low-light input appear noisy and lack clear structural definition. In contrast, the event features Z_2^{out} processed by our FEB module show strong activations concentrated around edges and structural boundaries. The final fused features F_2^{out} after CMAB exhibit significantly enhanced structural patterns compared to the image-only features, with clear preservation of edge information guided by the event stream. This progression from noisy image features to structurally coherent fused features shows that edge information from events successfully guides the DM to learn better features for low-light image restoration, thereby validating our core contribution of event-conditioned diffusion through the CMAB fusion module.

VI. CONCLUSION

In this paper, we introduced EcDiff-LLIE, a novel event-conditional diffusion framework for structure-preserving LLIE. Our approach utilizes a multimodality denoising network that jointly conditions on degraded low-light images and voxelized event streams. By extracting complementary local and global contextual features from the two modalities and fusing them through cross-modality attention blocks, the proposed method effectively exploits the strengths of both inputs. Furthermore, a lightweight feature enhancement block was designed to refine fine-grained event representations, enabling better preservation of details. Extensive experiments demonstrate that conditioning denoising networks on both modalities yields superior results compared to using RGB images alone, achieving consistent improvements over SOTA methods in both quantitative metrics and visual quality.

Limitations and Future Work: The proposed framework assumes spatially and temporally aligned low-light images and event streams. In the real-world, image–event misalignment due to sensor synchronization or calibration errors may occur, handling such cases is an important practical consideration and can be addressed by incorporating alignment or preprocessing modules prior to the fusion, which we leave for future work.

REFERENCES

[1] Y. Shi, D. Liu, L. Zhang, Y. Tian, X. Xia, and X. Fu, “ZERO-IG: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 3015–3024.

[2] L. Ma et al., “Learning with self-calibrator for fast and robust low-light image enhancement,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 9095–9112, Oct. 2025.

[3] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, “Event enhanced high-quality image recovery,” in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 155–171.

[4] C.-W. F. X. Xu, R. Wang, and J. Jia, “SNR-aware low-light image enhancement,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17714–17724.

[5] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.

[6] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, “Retinexformer: One-stage retinex-based transformer for low-light image enhancement,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12504–12513.

[7] T. Wang et al., “LLDiffusion: Learning degradation representations in diffusion models for low-light image enhancement,” *Pattern Recognit.*, vol. 166, 2025, Art. no. 111628.

[8] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, “Low-light image enhancement with wavelet-based diffusion models,” *ACM Trans. Graph.*, vol. 42, no. 6, pp. 1–14, 2023.

[9] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, “Diff-Retinex: Rethinking low-light image enhancement with a generative diffusion model,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12302–12311.

[10] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Med. Image Comput. Comput. Assist. Intervention*, Oct. 2015, pp. 234–241.

[11] Z. Jin, Y. Qiu, K. Zhang, H. Li, and W. Luo, “MB-TaylorFormer V2: Improved multi-branch linear transformer expanded by Taylor formula for image restoration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5990–6005, Jul. 2025.

[12] T. Wang et al., “GridFormer: Residual dense transformer with grid structure for image restoration in adverse weather conditions,” *Int. J. Comput. Vis.*, vol. 132, no. 10, pp. 4541–4563, 2024.

[13] G. Liang, K. Chen, H. Li, Y. Lu, and L. Wang, “Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 23–33.

[14] J. Weng, B. Li, and K. Huang, “Event-based image enhancement under high dynamic range scenarios,” in *Proc. Asian Conf. Comput. Vis.*, Dec. 2024, pp. 2456–2470.

[15] Y. Zhang et al., “TransDiff: Transformer-based diffusion model for low-light image enhancement,” *Neurocomputing*, vol. 655, 2025, Art. no. 131174.

[16] J. H. H. L. H. Z, J. Hou, Z. Zhu, and H. Yuan, “Global structure-aware diffusion process for low-light image enhancement,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, vol. 36, pp. 79734–79747.

[17] Z. Tu, Y. Liu, Y. Zhang, Q. Mu, and J. Yuan, “DTCM: Joint optimization of dark enhancement and action recognition in videos,” *IEEE Trans. Image Process.*, vol. 32, pp. 3507–3520, 2023.

[18] Y. Yin, M. Liu, R. Yang, Y. Liu, and Z. Tu, “Dark-DSAR: Lightweight one-step pipeline for action recognition in dark videos,” *Neural Netw.*, vol. 179, 2024, Art. no. 106622.

[19] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, “Adversarial spatio-temporal learning for video deblurring,” *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 291–301, Jan. 2019.

[20] K. Zhang, R. Li, Y. Yu, W. Luo, and C. Li, “Deep dense multi-scale network for snow removal using semantic and depth priors,” *IEEE Trans. Image Process.*, vol. 30, pp. 7419–7431, 2021.

[21] D. Liu, S. Li, Z. Xiao, P. An, and C. Shan, “L3FMamba: Low-light light field image enhancement with prior-injected state space models,” *IEEE Signal Process. Lett.*, vol. 32, pp. 3270–3274, 2025.

[22] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, “Uformer: A general U-Shaped transformer for image restoration,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17683–17693.

[23] K. Zhang, D. Li, W. Luo, W. Ren, and W. Liu, “Enhanced spatio-temporal interaction learning for video deraining: Faster and better,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1287–1293, Jan. 2023.

[24] H. Chung, B. Sim, and J. C. Ye, “Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12403–12412.

[25] X. Soria, G. Pomboza-Junez, and A. D. Sappa, “LDC: Lightweight dense CNN for edge detection,” *IEEE Access*, vol. 10, pp. 68281–68290, 2022.

[26] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *Proc. Int. Conf. Learn. Representations*, May 2021 doi: *arXiv:2010.02502*.

[27] H. Jiang, A. Luo, X. Liu, S. Han, and S. Liu, “LightenDiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2024, pp. 161–179.

[28] X. Lv, X. Dong, Z. Jin, H. Zhang, S. Song, and X. Li, “LDM: A diffusion model for low-light image enhancement,” in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, Oct. 2023, pp. 130–145.

[29] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 43, no. 6, pp. 1964–1980, Jun. 2021.

[30] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[32] C.-W. F. J. L. B. Y, R. Wang, X. Xu, and J. Jia, “Seeing dynamic scene in the dark: High-quality video dataset with mechatronic alignment,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9680–9689.

[33] Y. Hu, S.-C. Liu, and T. Delbruck, “v2e: From video frames to realistic DVS events,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 1312–1321.

[34] W. Weng, Y. Zhang, and Z. Xiong, “Event-based video reconstruction using transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 2543–2552.

[35] B. Ercan, O. Eker, A. Erdem, and E. Erdem, “HUE dataset: High-resolution event and frame sequences for low-light vision,” in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 174–191.

[36] T. Yang, C. Lan, Y. Lu, and N. Zheng, “Diffusion model with cross attention as an inductive bias for disentanglement,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 37, pp. 82465–82492.