Event-Illumination Collaborative Low-light Image Enhancement with a High-resolution Real-world Dataset
Supplementary Material

Contents

A. More details of our RLE Dataset

We will release all datasets to support the development of the event-based vision community.

A.1. Our RLE Dataset

We collect 7,888 event-image pairs with a resolution of 1024×768, a wide illumination range from 0.1 to 1000 lux (as shown in Fig. 2) in both indoor and outdoor conditions, along with low-light event streams captured during the exposure time of the RGB camera. This wide illumination range ensures our dataset's diversity, providing strong data support for enhancing a model's ability to handle various real-world inputs. Our dataset includes various scenes, such as libraries, buildings, and streets.

A.2. Samples

We provide sampled data from our RLE dataset as shown in Fig. 4.

A.3. Comparative Analysis of Real-World Datasets

As shown in **??**, our RLE dataset represents a significant advancement by balancing multiple competing factors. Unlike SDE [11], we provide high-resolution data and near-perfect temporal alignment. Unlike other datasets, we provide a quantified spatial error, enabling robust validation.

(i) Field of View vs. Alignment Precision. Our dual-beam splitter optical design enables the simultaneous capture of spatially aligned low-light images, event streams, and normal-light references. A key characteristic of this high-precision setup is the extended optical path length required to accommodate the beam splitters [8, 9]. To ensure optimal image quality and avoid peripheral occlusion (vignetting) from the optical assembly, we employ telephoto lenses rather than short-focal-length wide-angle lenses. While this design choice focuses our dataset on object-centric and medium-field scenes rather than ultra-wide-angle surveillance views, it guarantees artifact-free, center-aligned data essential for rigorous restoration benchmarking.

(ii) Data Volume and Scale. Unlike previous datasets [11] that rely on lower-resolution event sensors (e.g., 346×260), our RLE dataset captures high-resolution (1024×768) event streams with high temporal fidelity. This results in a significantly larger data footprint per sequence. Consequently, we prioritized scene diversity and signal quality over the sheer quantity of raw sequences. While the total number of samples is constrained compared to low-resolution alternatives, this trade-off ensures that each sequence in our dataset provides superior spatial detail for precise restoration tasks.

Figure 1. Real-world deployment of our imaging system, demonstrating the physical layout of the sensors and the optical pathway.

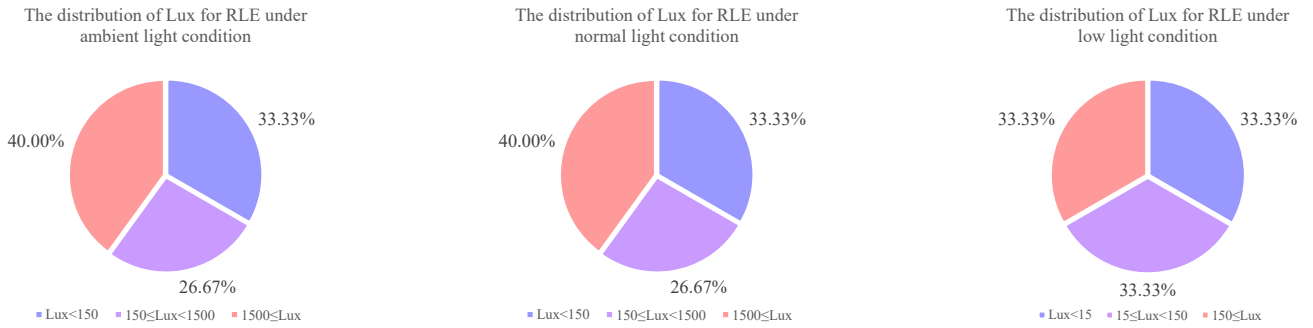

Figure 2. Illumination distribution in the filming environment.

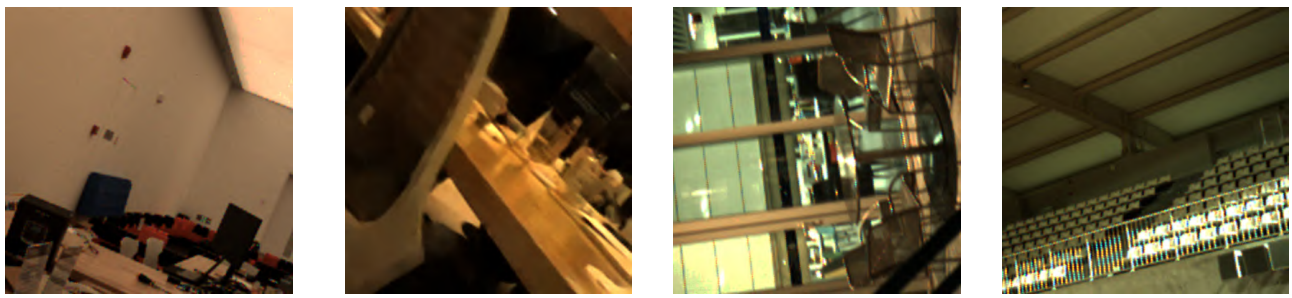

Figure 3. Samples of normal-light images from the SDE dataset that exhibit **color distortion** and l**ow resolution**.

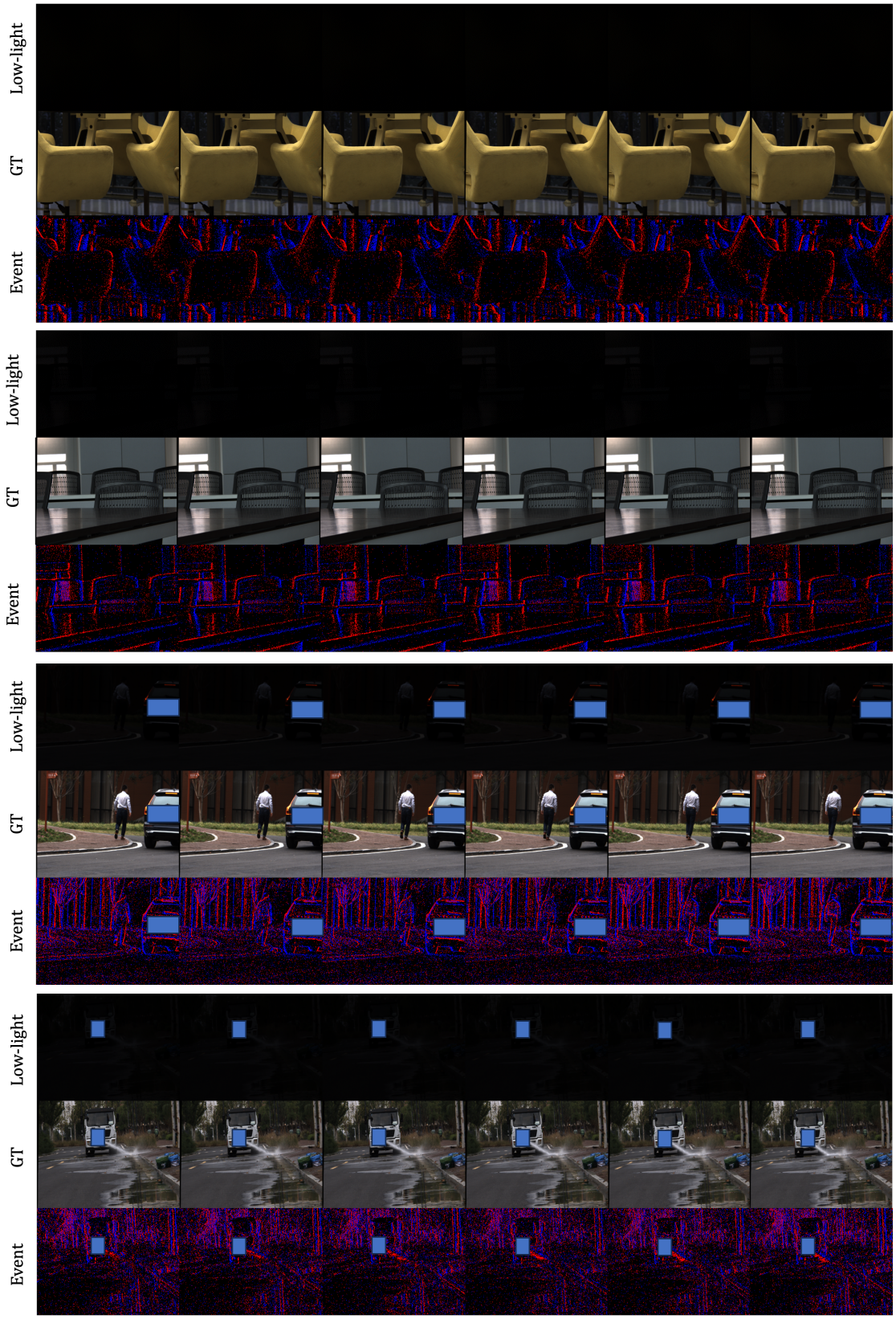

Figure 4. Samples from our RLE dataset.

B. More experimental results.

B.1. Ablation Studies on the Number of Blocks per Layer

We conduct an ablation study to evaluate how the number of blocks in each layer affects model performance. Specifically, we vary the block numbers in different layers while keeping other hyperparameters fixed.

In Tab. 1, our method with the {1,2,2} block configuration achieves a good balance between performance and model complexity. Compared to smaller architectures ({1,1,1}) and larger ones ({3,3,3}, {6,6,6}), it achieves superior PSNR while maintaining competitive SSIM. Notably, our model only requires **2.13M** parameters, significantly fewer than deeper or wider variants. This parameter efficiency is mainly attributed to the *event-illumination collaborative interaction*, which enables more effective feature learning with minimal architectural overhead. Increasing model capacity beyond this point yields diminishing returns, indicating that our design fully exploits the information from the event and illumination.

$\{N_1, N_2, ..., N_n\}$	PSNR	SSIM	Parm.(M)
$\{1, 1, 1\}$	20.11	0.7189	1.57
$\{1, 2, 2\}$(Ours)	23.63	0.7670	2.13
$\{3, 3, 3\}$	23.35	0.7681	3.01
$\{6, 6, 6\}$	23.71	0.7598	7.23
$\{1, 2, 2, 4\}$	23.69	0.7699	12.72
$\{2, 2, 4, 6\}$	22.88	0.7431	20.30

Table 1. Ablation Study on the Number of Blocks per Layer.

B.2. Ablation Studies on the EICI Module

To validate the architectural design of our Event-Illumination Collaborative Interaction (EICI) module, particularly the "Backward Injection" (BI) mechanism, we compare it against several baselines on the RLE dataset.

- **Case 6 (Ours w/o BI):** This removes the Backward Injection, creating a standard unidirectional (one-way) cross-attention fusion.
- **Case 6-large:** We scale up Case 6 to have a comparable parameter count to the SOTA method, EvLight, to show that the performance gap is not due to parameter count.
- **Retinexformer-event:** We adapt a strong image-based backbone (Retinexformer) to accept event data via a standard cross-attention module.

As shown in Tab. 2, simply removing Backward Injection (Case 6) causes a **>3.3dB PSNR drop**, proving that a standard unidirectional fusion fails. Naively adding event fusion to a strong backbone (Retinexformer-event) also yields poor results. This confirms that our bidirectional, collaborative design is not a redundant complication but a necessary and highly effective architecture for this task.

Method	PSNR	SSIM	Params(M)
Case 6 (Ours w/o BI)	20.24	0.6920	1.94
Retinexformer	20.33	0.6858	1.61
Retinexformer-event	20.57	0.6935	2.21
Case 6-large	22.04	0.7356	21.77
EvLight [11]	22.68	0.7201	22.73
Ours (EIC-LIE)	**23.63**	**0.7670**	**2.13**

Table 2. Ablation study on the EICI module. The catastrophic drop in performance for Case 6 (Ours w/o BI) demonstrates that a standard unidirectional cross-attention is insufficient. Our Backward Injection is critical for effective feature decomposition and refinement, and its design, not parameter count, is the key to our performance.

Case (Event Filter)	PSNR	SSIM
Case 1: No Filter	20.92	0.7064
Simpler Filters		
Case 2: w. Median Filter (pre)	20.43	0.6982
Case 3: w. Gaussian Filter (pre)	21.37	0.6988
Case 4: w. Illumination Adaptive Gating	21.55	0.7122
Case 5: w. IAEF (Fixed Function)	22.12	0.7309
Stronger Baselines		
Case 6: w. EDFormer [6] (pre)	21.33	0.7390
Case 7: w. ADFN [13]	22.28	0.7351
Case 8: Ours (w. IAEF)	**23.63**	**0.7670**

Table 3. Ablation study on the IAEF module. Our IAEF significantly outperforms all alternatives, including simple filters, adaptive gating, and strong SOTA filtering networks like ADFN and EDFormer. This proves the value of our specialized, illumination-guided dynamic filtering design.

B.3. Ablation Studies on the IAEF Module

We conducted a comprehensive ablation to validate our Illumination-Aware Event Filter (IAEF), comparing it against simpler traditional filters, adaptive gating mechanisms, and stronger, more established filtering baselines.

- **Simpler Filters:** We use Gaussian and Median filters as pre-processing steps.
- **Illum. Gating:** A simple module that uses illumination to create a gating (attention) map, as suggested by reviewers.
- **Fixed Function:** Replaces the learned kernel generation in IAEF with a fixed exponential function based on

Method	Task	PSNR	SSIM	Params(M)	FLOPs(G)	Runtime(ms)
Uformer [17]	General	20.58	0.704	50.88	178.9	487
X-Restormer [2]	General	20.89	0.713	25.98	328.6	812
SFHFormer-L [7]	General	20.17	0.695	57.61	101.4	377
Ours	**Task-Specific**	**23.63**	**0.767**	**2.13**	**70.9**	**298**

Table 4. Comparison with general-purpose restoration Transformers on the RLE dataset. Unified backbones perform significantly worse (>**2.74dB drop**) and are far more complex and slower.

brightness.

- **Strong Baselines:** We replace IAEF with ADFN [13] (a strong adaptive filtering network) and EDFormer [6] (a SOTA event denoising Transformer).

The results in Tab. 3 are clear. Simple filters (Cases 2-4) are insufficient. Even strong, established filtering networks (Cases 6, 7) perform significantly worse than our IAEF. This is because our IAEF is specifically architected to leverage the stable, global illumination statistics from the image to guide the dynamic filtering of event features within the network, a design that is more effective than generic preprocessing or adaptive filtering.

B.4. Comparison with Unified Transformer Backbones

We adapted several SOTA general-purpose image restoration Transformers to our task. We modified them to accept both the low-light image and event data as inputs.

As shown in Tab. 4, the unified restoration models fail on this task, performing significantly worse than our method while being orders of magnitude larger, more computationally expensive, and slower. This result strongly validates our approach, demonstrating that a general Transformer cannot effectively leverage the unique properties of event and illumination data. Our specialized, lightweight architecture is a highly efficient and necessary design.

C. More details of our EIC-LIE.

In this section, we detail some modules in our EIC-LIE framework.

C.1. Illumination Prior

The illumination prior L_p is derived by applying a pixel-wise maximum operation across all three channels of the input image $I \in \mathbb{R}^{H\times W\times 3}$, formulated as:

$$L_p = \max(I), \tag{1}$$

where the $\max(\cdot)$ function computes the maximum value at each spatial location across the color channels, resulting in a single-channel intensity map $L_p \in \mathbb{R}^{H\times W}$.

C.2. Illumination Estimator

Specifically, we first perform Illumination Estimator to estimate coarse lit-up image features $\mathbf{F}_i$ and initial illumination features $\mathbf{F}_l$ as:

$$(\mathbf{F}_i, \mathbf{F}_l) = \mathcal{C}(\mathbf{I}, \mathbf{L}_p), \tag{2}$$

where $\mathcal{C}$ denotes the illumination estimator consists of point-wise and depth-wise convolution layers.

C.3. Event-Illumination Collaborative Interaction

Note that we simplify the multi-head split in the main text for clarity. This section provides a comprehensive description of the complete multi-head collaborative bidirectional interaction mechanism.

Forward Gathering. The detail information flow of forward gathering $(\mathbf{T}', \mathbf{A}) = \mathcal{G}(\mathbf{X}, \mathbf{T})$.

Firstly, the input tokens $\mathbf{X} \in \mathbb{R}^{N\times C}$ and $\mathbf{T} \in \mathbb{R}^{N\times C}$ are split into k heads:

$$\mathbf{X} = [\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_k],\ \ \mathbf{T} = [\mathbf{T}_1,\ \mathbf{T}_2,\ \cdots,\ \mathbf{T}_k], \tag{3}$$

where $\mathbf{X}_i \in \mathbb{R}^{N\times d_k}, \mathbf{T}_i \in \mathbb{R}^{N\times d_k}, d_k = \frac{C}{k}$, and $i = 1, 2, \cdots, k$. For each $head_i$, three fully connected (fc) layers without $bias$ are used to linearly project $\mathbf{T}_i$ into $query$ elements $\mathbf{Q}_i \in \mathbb{R}^{N\times d_k}$; and $\mathbf{X}_i$ into *key* elements $\mathbf{K}_i \in \mathbb{R}^{N\times d_k}$ and *value* elements $\mathbf{V}_i \in \mathbb{R}^{N\times d_k}$ as

$$\mathbf{Q}_i = \mathbf{T}_i\mathbf{W}^{\mathrm{T}}_{\mathbf{Q}_i},\ \ \mathbf{K}_i = \mathbf{X}_i\mathbf{W}^{\mathrm{T}}_{\mathbf{K}_i},\ \ \mathbf{V}_i = \mathbf{X}_i\mathbf{W}^{\mathrm{T}}_{\mathbf{V}_i}, \tag{4}$$

where $\mathbf{W}_{\mathbf{Q}_i}$, $\mathbf{W}_{\mathbf{K}_i}$, and $\mathbf{W}_{\mathbf{V}_i} \in \mathbb{R}^{d_k\times d_k}$ represent the learnable parameters of the fc layers and T denotes the matrix transpose. Then, we calculate the attention matrix $\mathbf{A}_i$ and final output $\mathbf{T}'$ as:

$$\mathbf{A}_i = \mathrm{Softmax}(\frac{\mathbf{K}_i^{\mathrm{T}}\mathbf{Q}_i}{\alpha_i}),\ \ \mathbf{O}_i = \mathbf{V}_i\mathbf{A}_i, \tag{5}$$

$$\mathbf{T}' = \mathrm{Concat}(\mathbf{O}_i)\mathbf{W}^{\mathrm{T}}_{\mathbf{O}} + \mathbf{P} + \mathbf{T}, \tag{6}$$

where $\alpha_i \in \mathbb{R}^1$ is a learnable parameter that adaptively scales the matrix multiplication, Concat denotes the concatenation operation. To be specific, k heads are concatenated to pass through an fc layer $\mathbf{W}_{\mathbf{O}}$ and then add a positional encoding $\mathbf{P} \in \mathbb{R}^{N\times C}$ (learnable parameters) to produce the output tokens $\mathbf{T}' \in \mathbb{R}^{N\times C}$.

Backward Injection. The detail information flow of backward injection $\mathbf{X}' = \mathcal{I}(\mathbf{T}', \mathbf{A}) + \mathbf{X}$.

Firstly, the input tokens $\mathbf{T}' \in \mathbb{R}^{N\times C}$ is split into k heads:

$$\mathbf{T}' = [\mathbf{T}'_1,\ \mathbf{T}'_2,\ \cdots,\ \mathbf{T}'_k], \tag{7}$$

where $\mathbf{T}'_i \in \mathbb{R}^{N\times d_k}, d_k = \frac{C}{k}$, and $i = 1, 2, \cdots, k$. For each $head_i$, one fully connected (fc) layers without $bias$ are used to linearly project $\mathbf{T}'_i$ into *value* elements $\mathbf{V}'_i \in \mathbb{R}^{N\times d_k}$ as

$$\mathbf{V}'_i = \mathbf{T}'_i\mathbf{W}^{\mathrm{T}}_{\mathbf{V}'_i}, \tag{8}$$

where $\mathbf{W}_{\mathbf{V}'_i} \in \mathbb{R}^{d_k\times d_k}$ represents the learnable parameters of the fc layer. Then, we calculate the final output $\mathbf{X}'_i$ as:

$$\mathbf{X}'_i = \mathbf{A}_i^{\mathrm{T}}\mathbf{V}'_i,\ \ \mathbf{X}' = \mathrm{Concat}(\mathbf{X}'_i) + \mathbf{X}, \tag{9}$$

where $\mathbf{X}' \in \mathbb{R}^{N\times C}$ denotes the output tokens.

D. More details of the physical invariants

In this paper, the physical invariants W and C are applied to assist in the qualitative analysis of enhancement results across various methods. According to previous studies **[3, 10, 15], the property C may be interpreted as describing object color regardless of intensity, while the property W functions as an edge detector specific to changes in spectral distribution.** These invariants are derived from the Kubelka-Munk light transport theorem [4] and are based on the color invariance analysis by Geusebroek et al. [3].

Different from [4, 10], Wang et al. [15] propose a learning-based prior estimation module to predict the observed energy E, then compute H, C, and W from the input image. They utilize physical invariants prior to guiding the diffusion process. **In our work, we apply the pretrained prior estimation module [15] and compute H, C, and W from the input image.**

D.1. The derivation of the physical invariants

Following the Wang et al. [15], given wavelength λ, the energy of the incoming spectrum at spatial location $\mathbf{x}$ on the image plane is modeled as

$$E(\lambda, \mathbf{x}) = e(\lambda, \mathbf{x}) \left((1 - i(\mathbf{x}))^2 R_\infty(\lambda, \mathbf{x}) + i(\mathbf{x}) \right), \quad (10)$$

where $e(\lambda, \mathbf{x})$ denotes the spectrum of the light source, $i(\mathbf{x})$ the specular reflection, and $R_\infty(\lambda, \mathbf{x})$ the material reflectivity. Note that when the object is matte, i.e. $i(\mathbf{x}) \approx 0$, Eq. (10) can be reduced to

$$E(\lambda, \mathbf{x}) = e(\lambda, \mathbf{x}) R_\infty(\lambda, \mathbf{x}), \quad (11)$$

which is the same as the Retinex model. It means that the Retinex theory is a special case of Eq. (10).

First of all, we denote some variables for simplicity

$$E^\lambda = \frac{\partial E(\lambda, \mathbf{x})}{\partial \lambda}, \; R_\infty^\lambda = \frac{\partial R_\infty(\lambda, \mathbf{x})}{\partial \lambda}, \quad (12)$$

$$E^{\lambda\lambda} = \frac{\partial^2 E(\lambda, \mathbf{x})}{\partial \lambda^2}, \; R_\infty^{\lambda\lambda} = \frac{\partial^2 R_\infty(\lambda, \mathbf{x})}{\partial \lambda^2}. \quad (13)$$

Intuitively, E represents spectral intensity, E^λ signifies spectral slope, and $E^{\lambda\lambda}$ denotes spectral curvature.

Through simplifying assumptions, we can obtain a series of invariants from Eq. (10). The primary idea is to eliminate i and e, retaining solely R_∞. As R_∞ is about material property and is independent of illumination, the derived variable will exhibit illumination invariance.

- Assuming *equal energy* illumination, i.e., $e(\lambda, \mathbf{x})$ is reduced to λ-independent $\tilde{e}(\mathbf{x})$, and Eq. (10) is reduced to

$$E(\lambda, \mathbf{x}) = \tilde{e}(\mathbf{x}) \left((1 - i(\mathbf{x}))^2 R_\infty(\lambda, \mathbf{x}) + i(\mathbf{x}) \right), \quad (14)$$

substituting Eq. (14) into $E^\lambda / E^{\lambda\lambda}$ gives

$$\frac{E^\lambda}{E^{\lambda\lambda}} = \frac{\tilde{e}(\mathbf{x})(1 - i(\mathbf{x}))^2 R_\infty^\lambda}{\tilde{e}(\mathbf{x})(1 - i(\mathbf{x}))^2 R_\infty^{\lambda\lambda}} = \frac{R_\infty^\lambda}{R_\infty^{\lambda\lambda}}, \quad (15)$$

where illumination properties i and e are eliminated. As the material property R_∞ is independent of illumination, it establishes the illumination-invariance of $E^\lambda / E^{\lambda\lambda}$. Now, derive the illumination invariant H,

$$H = \arctan\left(E^\lambda / E^{\lambda\lambda} \right). \quad (16)$$

- *Further* assuming that the surface is *matte*, i.e. $i(\mathbf{x}) \approx 0$, then Eq. (10) is reduced to

$$E(\lambda, \mathbf{x}) = \tilde{e}(\mathbf{x}) R_\infty(\lambda, \mathbf{x}), \quad (17)$$

similarly, derive another illumination invariant C,

$$C = \log\left(\frac{(E^\lambda)^2 + (E^{\lambda\lambda})^2}{E(\lambda, \mathbf{x})^2} \right) = \log\left(\frac{(R_\infty^\lambda)^2 + (R_\infty^{\lambda\lambda})^2}{R_\infty(\lambda, \mathbf{x})^2} \right). \quad (18)$$

- *Further* assuming *uniform* illumination, i.e., $\tilde{e}(\mathbf{x})$ is reduced to a parameter $\bar{e}$, and Eq. (10) is reduced to

$$E(\lambda, \mathbf{x}) = \bar{e} R_\infty(\lambda, \mathbf{x}), \quad (19)$$

Similarly, derive the illumination invariant W,

$$W = \tan\left(\left| \frac{\partial E(\lambda, \mathbf{x})}{\partial \mathbf{x}} \frac{1}{E(\lambda, \mathbf{x})} \right| \right) = \tan\left(\left| \frac{\partial R_\infty(\lambda, \mathbf{x})}{\partial \mathbf{x}} \frac{1}{R_\infty(\lambda, \mathbf{x})} \right| \right). \quad (20)$$

Learning-based Prior Estimation Module. Wang et al. [15] follow Gaussian color models [4] and CIConv [10] to obtain priors from RGB images. First, Wang et al. [15] estimate the observed energy $\hat{E}$ along with its derivatives $\hat{E}^\lambda$ and $\hat{E}^{\lambda\lambda}$ via linear mapping:

$$\begin{bmatrix} \hat{E}(x, y) \\ \hat{E}^\lambda(x, y) \\ \hat{E}^{\lambda\lambda}(x, y) \end{bmatrix} = \mathcal{W} \begin{bmatrix} R(x, y) \\ G(x, y) \\ B(x, y) \end{bmatrix}, \quad (21)$$

where x and y denote positions in the image, and $\mathcal{W}$ is a 3×3 matrix. In [4, 10], $\mathcal{W}$ is manually defined. Wang et al. [15] instead learn it from the distribution of natural images through a prior-to-image framework.

The spatial derivative $\partial E / \partial \mathbf{x}$ in Eq. (20) is computed in both the x- and y-direction, denoted as $\partial E / \partial \mathbf{x} = (E_x, E_y)$, with its magnitude given by $|\partial E / \partial \mathbf{x}| = \sqrt{E_x^2 + E_y^2}$. Finally, E, E_x, and E_y are estimated by convolving $\hat{E}$ with Gaussian color smoothing and derivative filters of scale σ. σ is predicted from the input image. Similarly, E^λ is obtained from $\hat{E}^\lambda$, and $E^{\lambda\lambda}$ is obtained from $\hat{E}^{\lambda\lambda}$. Now we can compute H, C, and W from the input image.

E. More details of implementations

E.1. Implementation Details

We implement our framework via the PyTorch 1.8 platform on a NVIDIA RTX 4090 GPU. We adopt AdamW [12] optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to optimize our network. We train our network with patch size 128×128 and batch size 24. The initial learning rate is 2×10^{-4}, which changes to 5×10^{-4}, and finally decreases to 1×10^{-6}, following Cosine Annealing scheme. Note that we set the bin size B of event voxels as 6. The training objective is to minimize the Charbonnier loss [1] between the ground truth and the enhanced image.

E.2. Datasets

We use five datasets, including SDE-indoor, SDE-outdoor, SDSD-indoor, SDSD-outdoor, and RLE, to validate the effectiveness of our proposed method, and for a fair comparison, we retrained the previous six methods on the RLE dataset.

SDE: The SDE dataset [11] utilizes a robotic alignment system equipped with a DAVIS346 event camera to capture spatially paired low-light and normal-light sequences. Through a matching alignment strategy, it achieves high-precision temporal alignment across 91 sequences. The resolution of both image and event data is 346×260. Following the original setup, 76 sequences are allocated to the training set, while the remaining 15 sequences are designated for the testing set.

SDSD: The SDSD dataset [14] employs a camera mounted on an electromechanical rail system to capture 150 paired normal-light and low-light video sequences, with an original resolution of 1920×1080. We utilize the Liang et al. [11] released version, which includes synthetic event data and has a resolution of 346×260. Following the same dataset partitioning strategy, 125 sequences are used for training, and 25 sequences are reserved for testing.

RLE: The RLE dataset (Ours) utilizes an optical system comprising two RGB cameras and one event camera to capture 7,888 pairs of low-light and normal-light images with a resolution of 1024×768, along with event data during the exposure time. The dataset encompasses complex motions across various scenarios. We allocate 4,877 pairs for training and 3,011 pairs for testing.

E.3. Loss Function

We use the Charbonnier loss [1] to train our network in an end-to-end fashion:

$$\mathcal{L}_{\text{char}} = \sqrt{\left\| \hat{\mathbf{I}} - \mathbf{G} \right\|^2 + \epsilon^2} \tag{22}$$

where $\hat{\mathbf{I}}$ is the predicted image, and G is the ground truth. The constant ϵ is empirically set to 10^{-3} for all the experiments as in [18].

E.4. Metrics

We utilize the widely recognized Peak Signal-to-Noise Ratio (**PSNR**) [5] and Structural Similarity Index (**SSIM**) [16] as our primary metrics for evaluating image enhancement.

References

[1] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, pages 168–172. IEEE, 1994. 8

[2] Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 5

[3] J-M Geusebroek, Rein Van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Transactions on Pattern analysis and machine intelligence*, 23 (12):1338–1350, 2001. 7

[4] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in computer vision: fundamentals and applications*. John Wiley & Sons, 2012. 7

[5] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44(13): 800–801, 2008. 8

[6] Bin Jiang, Bo Xiong, Bohan Qu, M Salman Asif, You Zhou, and Zhan Ma. Edformer: Transformer-based event denoising across varied noise levels. In *European Conference on Computer Vision*, pages 200–216. Springer, 2024. 4, 5

[7] Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *European Conference on Computer Vision*, pages 381–402. Springer, 2024. 5

[8] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024. 1

[9] Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pages 433–451. Springer, 2025. 1

[10] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4399–4409, 2021. 7

[11] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: A large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23–33, 2024. 1, 4, 8

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8

[13] Hao Shen, Zhong-Qiu Zhao, and Wandi Zhang. Adaptive dynamic filtering network for image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2227–2235, 2023. 4, 5

[14] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9700–9709, 2021. 8

[15] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26057–26066, 2024. 7

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

[17] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 5

[18] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 8