Dynamic Reflection Removal from RGB Images via Event-Guided Multimodal Approach

Simone Melcarne, Sahar Husseini, and Jean-Luc Dugelay

Eurecom Research Center, Data Science Department, France
melcarne@eurecom.fr

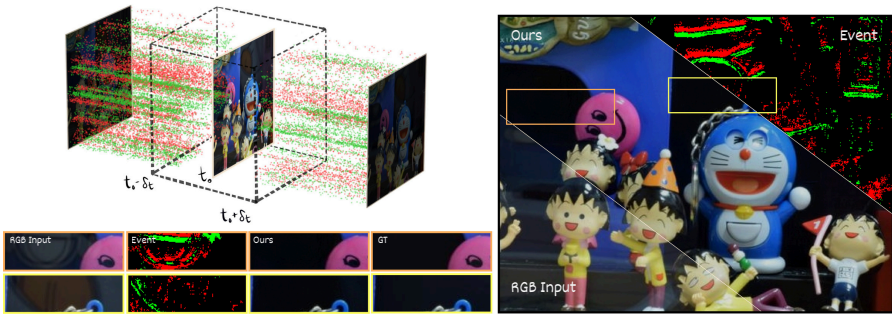


Fig. 1: A dual-camera system captures an RGB mixture along with a synchronized event stream that exclusively map the dynamic reflections. Our framework is able to remove reflection artifacts while preserving the content of the background.

Abstract. Single-image reflection removal is a relevant yet ill-posed task that aims to separate two overlapping layers from a single mixture observation. While decoupling the background from the reflection layer is inherently ambiguous when relying solely on a single modality, multi-modal solutions that leverage auxiliary modalities have emerged as a powerful alternative. In this work, we introduce a two-stage framework that pairs a single RGB image with a synchronized window of event data captured during the camera shot, under the assumption of static background and moving reflection. In the first stage the model extracts features from both the RGB frame and the synchronized event voxel grid to estimate the reflection layer. Then, the second stage reconstructs the clean background from the intermediate representations propagated through cross-stage skip connections. To train and validate our approach, we build a large-scale synthetic dataset and a preliminary real-world testing benchmark. Experimental results demonstrate that our framework effectively bridges the sim-to-real gap, showing good separation and generalization abilities compared to state-of-the-art methods. *Code and data will be released.*

Keywords: Computational Photography · Image Reflection Removal · Multimodal Fusion · Event-Based Cameras

1 Introduction

Capturing images through glass is a challenging situation that may happen in everyday photography (*e.g.,* museum displays, car windows, shop fronts, *etc.*). Every glass surface partially reflects the scene behind the camera, superimposing a reflection layer at the top of the desired background and potentially creating an issue for imaging systems. Beyond aesthetics, this contamination can have an impact on downstream tasks like image classification, object detection or semantic segmentation [27, 36, 50]. Single-image reflection removal (*i.e.,* SIRR) is a timely research area that tries to separate the *mixture* observation $\mathbf{M}$ into the transmission layer $\mathbf{T}$ and the reflection layer $\mathbf{R}$. This problem is intrinsically ill-posed as the same $\mathbf{M}$ is consistent with infinite valid $(\mathbf{T}, \mathbf{R})$ pairs [52]. The deep-learning community has made remarkable progress on SIRR by exploiting large-scale training data together with carefully designed priors to constrain the problem, however, when the reflection and transmission layers share similar content, a single RGB image simply does not carry enough information to resolve this ambiguity. For this purpose, a natural solution is to introduce a second complementary modality. Prior works have explored near-infrared images (NIR) [13], RAW sensor data [38] and flash/no-flash pairs [20]. Each of these approaches reduces the ambiguity of the task by incorporating additional constraints, while requiring more advanced hardware or carefully designed capture protocols. In this work, we take a similar approach and propose to use *event-based sensors* as the auxiliary modality. An event camera is a neuromorphic sensor that responds asynchronously to per-pixel brightness changes, emitting a stream of timestamped events rather than frames [7]. Under the assumption of a static background scene, where $\mathbf{T}$ remains time-invariant and cannot trigger the sensor, the asynchronous event stream captures *exclusively* the dynamic reflections $\mathbf{R}$ moving across the glass surface, providing the network with a clear clue that maps the shape, position, and trajectory of $\mathbf{R}$ over time. Equipped with this observation, we propose **E-SIRR**, an event-guided network for single-image reflection removal. Our architecture consists of (*i*) a *Reflection Estimation Network* (Stage 1) that fuses RGB and event features through a dedicated module and produce an intermediate reflection estimate $\hat{\mathbf{R}}$ and (*ii*) a *Background Restoration Network* (Stage 2) that restores the transmission layer $\hat{\mathbf{T}}$ by leveraging both $\mathbf{M}$ and $\hat{\mathbf{R}}$, further enriched by cross-stage connections. We summarize our main contributions as follows:

- To the best of our knowledge, we introduce the *first* event-guided framework specifically tailored for SIRR.
- We design a two-stage multimodal network that exploits the event signal when truly informative to perform the estimation and restoration tasks.
- We develop a large-scale *synthetic* dataset for training event-guided reflection removal models, featuring dynamic reflection simulation, synchronized RGB-event data, and ground-truth supervision; we also build a novel preliminary *real-world* dataset captured with an event camera for testing. Experiments demonstrate that our model reconstructs physically consistent reflection-free images compared with state-of-the-art single-modality baselines.

The remainder of this paper is organized as follows. Sec. 2 reviews related work on SIRR and event-based image restoration. Sec. 3 outlines the physical mixture formulation and the neuromorphic background theory. Sec. 4 details the proposed two-stage network architecture and the objective functions. Sec. 5 and Sec. 6 describe the synthetic and real-world datasets alongside the experimental setup. Finally, quantitative and qualitative results are analyzed in Sec. 7, followed by limitations and concluding remarks in Sec. 8 and 9.

2 Related Work

2.1 Single-Image Reflection Removal

To separate transmission and reflection layers, early optimization-based methods relied on handcrafted priors and physical assumptions, such as gradient sparsity, relative smoothness, and ghosting cues [2, 21, 24, 36]. Deep-learning based SIRR methods tackled the problem from a data-driven perspective. Zhang *et al.* [55] introduced one of the first end-to-end CNN approaches with perceptual and exclusion losses, while ERRNet [45] exploited misaligned real-world pairs through alignment-invariant training. Several subsequent methods improved the physical modeling and supervision strategy, including BDN [51], CRRN [41], CoRRN [43] and RAGNet [25], by incorporating reflection-aware guidance, multi-scale features, or cooperative decomposition. Other approaches like YTMT [14] and DSRNet [15] focused on the interaction between transmission and reflection streams, or explicitly localized reflection-dominant regions, as in LANet [6].

More recently, RDNet [57] revisited the problem by introducing a multi-column reversible encoder together with a transmission-rate-aware prompt generator, reporting strong performance on real-world benchmarks. The field has also been further advanced by improved real-world data collection and benchmarking, culminating in recent work that introduces larger paired datasets and stronger location-guidance frameworks [60]. Despite these advances, current methods that still rely solely on RGB input are forced to hallucinate the separation, leaving room for multimodal supervision [40, 52].

2.2 Event-Based Image Restoration and Enhancement

Event cameras have been applied to a wide range of image restoration tasks. Given their ability to capture high dynamic range (HDR) scenes with microsecond temporal resolution [7], these sensors constitute an effective modality in solving problems like de-blurring [3,4,17,28,34,58], super-resolution [8,10,18,33], HDR reconstruction [5, 9, 11, 35, 53] and low-light enhancement [16, 39, 49].

More closely related to our setup are methods leveraging event sensors for occlusion removal. For instance, Li *et al.* [23] address dense static occlusions by moving the camera to capture background information behind the obstacle via events. Zou *et al.* [61] tackle dynamic occlusions where a moving occluder triggers events across a static background. Recently, Zhang *et al.* [56] exploited

events for reflective flare removal, observing that attenuated specular reflection intensities often fall below the sensor contrast threshold, leaving the event stream naturally flare-free.

While these approaches share the common objective of artifact removal, they are designed for different degradation models and application scenarios. To the best of our knowledge, no existing work directly addresses event-guided reflection removal, making this study the first to explore this direction.

3 Background and Preliminaries

3.1 Mixture Image Formation

A mixture image $\mathbf{M}$ is commonly modeled as the linear combination of a transmission layer $\mathbf{T}$ and a reflection layer $\mathbf{R}$ [21]:

$$\mathbf{M} = \mathbf{T} + \mathbf{R} \tag{1}$$

To better capture real-world dynamics [42], several works have also proposed augmented formulations, introducing weighting coefficients ($\mathbf{M} = \alpha\mathbf{T} + \beta\mathbf{R}$) [41,51], alpha-matting masks ($\mathbf{M} = \mathbf{W} \odot \mathbf{T} + (\mathbf{1} - \mathbf{W}) \odot \mathbf{R}$) [46], or non-linear residuals ($\mathbf{M} = \mathbf{T} + \mathbf{R} + \phi(\mathbf{T}, \mathbf{R})$) [15]. While these solutions can improve physical accuracy, they also highlight the under-determined nature of the task.

In this work, we try to constrain and simplify the problem by introducing event data, a modality structurally complementary to the frame signal.

3.2 Neuromorphic Approach

An event camera is a bio-inspired sensor that responds *asynchronously* to per-pixel logarithmic brightness changes. Each pixel outputs an event e whenever the log-intensity change of the brightness in an image $\mathbf{I}$ exceeds a contrast threshold Θ [1, 7]:

$$\mathbf{e}_{\mathbf{I}}(x, y, t, \sigma) \coloneqq \log(\mathbf{I}(x, y, t)) - \log(\mathbf{I}(x, y, t - \delta t)) \geq \Theta \tag{2}$$

Each event is represented by the pixel coordinates (x, y), indicating where the brightness change occurred, a timestamp t, specifying the exact time of the event, and a polarity σ of value ± 1, indicating whether the brightness increased or decreased.

In the context of reflection removal, if we make the experimental assumption that the target background scene remains static over time, whereas the degradation is caused exclusively by a *dynamic* reflection layer moving across the glass surface, we can formally define the problem as follows: let the continuous mixture image $\mathbf{M}(x, y, t)$ received by the event sensor be the sum of the background layer $\mathbf{T}(x, y, t)$ and the moving reflection layer $\mathbf{R}(x, y, t)$ such that:

$$\mathbf{M}(x, y, t) = \mathbf{T}(x, y, t) + \mathbf{R}(x, y, t) \tag{3}$$

With the static background constraint, the transmission becomes time-invariant, meaning that $\mathbf{T}(x, y, t) = \mathbf{T}(x, y, t - \delta t) = \mathbf{T}(x, y)$ does not generate events. By substituting Eq. (3) into the general neuromorphic event-triggering condition defined in Eq. (2), we can approximate the relationship as:

$$\mathbf{e}_{\mathbf{M}}(x, y, t, \sigma) \approx \log\left(\mathbf{R}(x, y, t)\right) - \log\left(\mathbf{R}(x, y, t - \delta t)\right) \geq \Theta \tag{4}$$

As a result, the asynchronous event stream isolates the high-frequency moving edges of the reflection layer.

4 Methodology

Building on the formulation described in Sec. 3, the starting point of our work is a dual-camera system consisting of one RGB sensor and an event camera. The RGB camera captures a color mixture image $\mathbf{M} \in \mathbb{R}^{3\times H\times W}$ at a specific timestamp t_0, where H and W denote the height and width of the image, respectively; simultaneously, the event sensor records a continuous spatio-temporal stream of events $\mathcal{E}$. To pair the RGB frame with temporally synchronized event data, we isolate a subset of the event stream restricted to a symmetric temporal neighborhood centered around the frame acquisition time t_0, spanning both the immediate past and future: $\mathcal{E}_{\delta t} = \{e_k \in \mathcal{E} \mid t_k \in [t_0 - \delta t, t_0 + \delta t]\}$. Then, we convert this localized stream $\mathcal{E}_{\delta t}$ of N events into a fixed-size voxel grid, a well-established event representation choice in image reconstruction tasks [9, 32, 59]. Let t_{start} and t_{end} denote the timestamps of the first and last events in $\mathcal{E}_{\delta t}$, respectively. The temporal duration $\Delta t = t_{end} - t_{start}$ is partitioned into B temporal bins, and Eq. 5 defines the mapping of events to the corresponding bins:

$$\begin{gathered} \mathbf{V} = \sum_{n=0}^{N-1} \sigma_n \, \max\!\big(0,\, 1 - \big|\, t - \hat{t}_n \,\big|\big) \\ \textit{with } t = 0, \ldots, B-1, \quad \hat{t}_n = \frac{t_n - t_{start}}{\Delta t}(B-1) \end{gathered} \tag{5}$$

We tackle the problem of reflection removal as a two-stage estimation problem: given a single corrupted RGB image $\mathbf{M} \in \mathbb{R}^{3\times H\times W}$ and a synchronized segment of event data $\mathbf{V} \in \mathbb{R}^{B\times H\times W}$, our objective is to first estimate the reflection layer $\hat{\mathbf{R}} \in \mathbb{R}^{3\times H\times W}$ through a Reflection Estimation Network ($\mathcal{R}$), and then restore a reflection-free background image $\hat{\mathbf{T}} \in \mathbb{R}^{3\times H\times W}$ through a Background Restoration Network ($\mathcal{T}$):

$$\hat{\mathbf{R}} = \mathcal{R}\left(\mathbf{M}, \mathbf{V}\right), \qquad \hat{\mathbf{T}} = \mathcal{T}(\mathbf{M}, \mathbf{V}, \hat{\mathbf{R}}) \tag{6}$$

Figure 2 illustrates the complete pipeline of the proposed framework, which we describe in the following sub-sections.

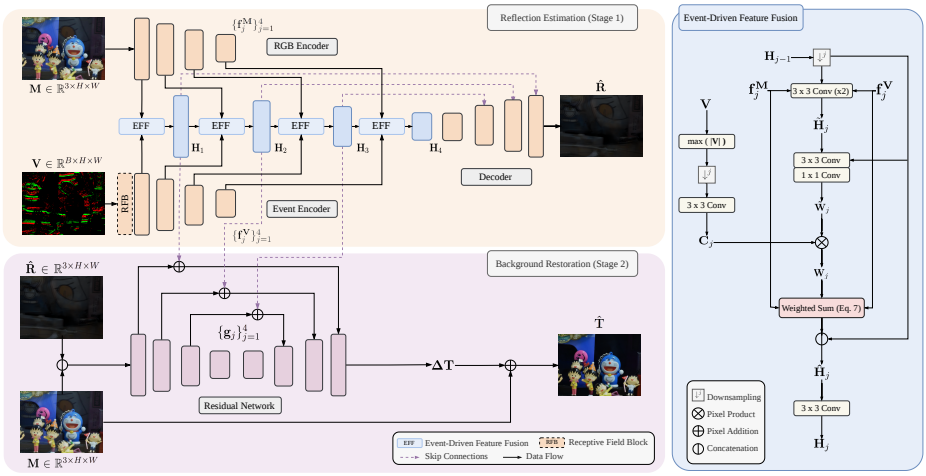


Fig. 2: Overview of the proposed framework. In the first stage, a dual-branch encoder processes the mixture image $\mathbf{M}$ and the event voxel grid $\mathbf{V}$, fusing their features through the EFF module to estimate the reflection layer $\hat{\mathbf{R}}$. In the second stage, $\mathbf{M}$ and $\hat{\mathbf{R}}$ are used to reconstruct the final clean background transmission $\hat{\mathbf{T}}$, leveraging cross-stage connections ($\mathbf{H}_j$).

4.1 Reflection Estimation Network

Dual-Branch Encoding The mixture image $\mathbf{M}$ and the synchronized voxel grid $\mathbf{V}$ are processed by two independent encoders on 4 scales. Each encoder scale is built upon standard Residual Blocks [12] combined with Group Normalization [48]. This encoding process creates a multi-scale set of features denoted as $\{\mathbf{f}_j^{\mathbf{M}}\}_{j=1}^4$ for the frame branch and $\{\mathbf{f}_j^{\mathbf{V}}\}_{j=1}^4$ for the event branch.

Before entering the first residual block, the raw voxel grid $\mathbf{V}$ is processed by one single Receptive Field Block (RFB) [31] that we adapt from [47] and consists of three parallel dilated convolution paths with dilation rates of 1, 2, and 3, respectively. Their outputs are concatenated along the channel dimension and fused via a 1×1 convolution layer. By acting as the first operation on $\mathbf{V}$, the RFB has been shown effective in expanding the receptive field prior to any downsampling operations.

Feature Fusion Module We implement an Event-Driven Feature Fusion (EFF) module based on Zou *et al.* [61] that dynamically weights the event stream. First, a preliminary cross-modal tensor $\tilde{\mathbf{H}}_j$ is computed via two consecutive 3×3 convolutional layers (each followed by normalization and ReLU) over the concatenated features ($\mathbf{f}_j^{\mathbf{M}}$, $\mathbf{f}_j^{\mathbf{V}}$), and combined with the tensor from the previous scale ($\mathbf{H}_{j-1}$) to estimate a raw spatial gating mask $\tilde{\mathbf{W}}_j$.

In addition, we introduce an Event Confidence Map ($\mathbf{C}_j$) calculated directly from $\mathbf{V}$ by taking the maximum absolute intensity across the temporal bins and

processing it via a shallow 3×3 convolutional layer. For each scale j, $\mathbf{C}_j$ is downsampled to match the current resolution and modulate the spatial gate via $\mathbf{W}_j = \tilde{\mathbf{W}}_j \odot \mathbf{C}_j$. The gated modality features are then mixed and concatenated with the cross-scale context $\mathbf{H}_{j-1}$ as:

$$\hat{\mathbf{H}}_j = \Big[(1 - \mathbf{W}_j) \odot \mathbf{f}_j^{\mathbf{M}} + \mathbf{W}_j \odot \mathbf{f}_j^{\mathbf{V}} \,;\, \mathbf{H}_{j-1}\Big], \tag{7}$$

where $[\cdot\,;\,\cdot]$ denotes channel-wise concatenation. Finally, $\hat{\mathbf{H}}_j$ is projected through a 3×3 convolutional layer (followed by normalization and ReLU) to adjust the channel dimensions and obtain the fused feature map $\mathbf{H}_j$. This gating mechanism enables the network to discard the event contribution whenever $\mathbf{V}$ is not sufficiently informative.

Decoder and Reflection Estimation The multi-scale fused features $\{\mathbf{H}_j\}_{j=1}^4$ are passed to a U-Net style decoder with skip connections to perform the explicit reflection estimation, obtaining the intermediate output $\hat{\mathbf{R}}$.

4.2 Background Restoration Network

Joint Encoding and Cross-Stage Propagation In the second stage, $\mathbf{M}$ and the predicted reflection $\hat{\mathbf{R}}$ are concatenated into a 6-channel tensor and processed by a secondary residual encoder to produce features $\{\mathbf{g}_j\}_{j=1}^4$. To prevent high-frequency details loss in the Stage 1 bottleneck, we decided to introduce cross-stage connections that re-inject the intermediate fused features $\{\mathbf{H}_j\}_{j=1}^3$ from Stage 1 into Stage 2, by adding them to the extracted features $\{\mathbf{g}_j\}_{j=1}^3$, after a 1×1 convolution layer initialized to zero weights, ensuring that the cross-stage path activates dynamically if beneficial.

Decoder and Transmission Estimation Finally, the enhanced feature maps are mapped through a U-Net decoder to predict first a background residual mapping $\mathbf{\Delta T}$, that is later used to recover the final clean transmission image $\hat{\mathbf{T}}$ by adding it to the mixture $\mathbf{M}$.

4.3 Objective Function

To optimize the network end-to-end, we define a composite loss $\mathcal{L}_{\text{total}}$ that balances pixel reconstruction, structural perception, and layer separation. Given the estimated reflection $\hat{\mathbf{R}}$ from $\mathcal{R}$ and the restored transmission $\hat{\mathbf{T}}$ from $\mathcal{T}$, the global training loss $\mathcal{L}_{\text{total}}$ is formalized as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{R}}(\hat{\mathbf{R}}, \mathbf{R}) + \mathcal{L}_{\mathcal{T}}(\hat{\mathbf{T}}, \mathbf{T}) + \omega_e \mathcal{L}_{\text{excl}}(\hat{\mathbf{T}}, \hat{\mathbf{R}}), \tag{8}$$

where $\mathbf{R}$ and $\mathbf{T}$ are the ground-truth reflection and transmission layers respectively; $\mathcal{L}_{\mathcal{R}}$ and $\mathcal{L}_{\mathcal{T}}$ are composite losses separately designed for the two stages, while $\mathcal{L}_{\text{excl}}$ represents the cross-stage exclusion loss.

Stage-Wise Reconstruction Losses. For both estimation ($\mathbf{X} = \mathbf{R}$) and restoration ($\mathbf{X} = \mathbf{T}$) stages, the loss terms are defined as:

$$\mathcal{L}_{\mathbf{X}}(\hat{\mathbf{X}}, \mathbf{X}) = \omega_\ell \mathcal{L}_{pix}(\hat{\mathbf{X}}, \mathbf{X}) + \omega_\phi \mathcal{L}_{\text{perc}}(\hat{\mathbf{X}}, \mathbf{X}) \tag{9}$$

The pixel-level loss $\mathcal{L}_{pix}$ ensures photometric fidelity:

$$\mathcal{L}_{pix}(\hat{\mathbf{X}}, \mathbf{X}_{\text{gt}}) = \|\hat{\mathbf{X}} - \mathbf{X}_{\text{gt}}\|_1 \tag{10}$$

To preserve semantic content, the perceptual loss $\mathcal{L}_{\text{perc}}$ computes the MSE over the activation maps Φ_i from the relu1_2 and relu2_2 layers of a pretrained VGG-19 network [19, 37]:

$$\mathcal{L}_{\text{perc}}(\hat{\mathbf{X}}, \mathbf{X}) = \sum_{i \in \{1,2\}} \|\Phi_i(\hat{\mathbf{X}}) - \Phi_i(\mathbf{X}_{\text{gt}})\|_2^2 \tag{11}$$

We set $\omega_\ell = 1.0$ and $\omega_\phi = 0.1$.

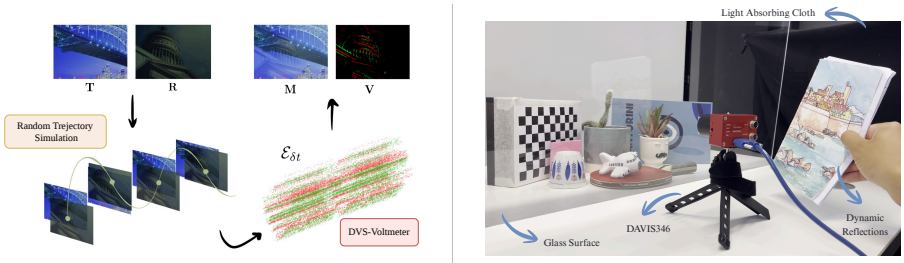
Cross-Stage Exclusion Loss. To ensure structural separation, we make use of a multi-scale exclusion loss as defined in [26, 52, 55], which is evaluated across $L = 3$ pyramid levels with a weight ω_e that we set to 0.02:

$$\mathcal{L}_{\text{excl}}(\hat{\mathbf{T}}, \hat{\mathbf{R}}) = \frac{1}{L} \sum_{l=0}^{L-1} \sqrt{\|\Psi(\hat{\mathbf{T}} \downarrow^l, \hat{\mathbf{R}} \downarrow^l)\|_F}, \tag{12}$$

where $\downarrow^l$ denotes downsampling via average pooling, $\|\cdot\|_F$ is the Frobenius norm, and $\Psi(\cdot)$ computes the normalized spatial correlation between the gradient maps.

5 Dataset

Synthetic Dataset To train our multimodal network, we construct a large-scale synthetic dataset featuring, for each sample, the corrupted mixture frame $\mathbf{M}$, the event voxel grid $\mathbf{V}$, and the background and reflection ground truths $\mathbf{T}$, $\mathbf{R}$. Both the background scene and the reflection layer are sampled from the COCO dataset [30], and at each iteration we combine two random images to form the mixture image and the event data. Following [40, 44], to emulate the real-world reflection behavior, we apply a random Gaussian blur to the original reflection image $\mathbf{R}_{\mathcal{G}} = \mathcal{G}(\mathbf{R}; \sigma)$, where $\mathcal{G}(\cdot)$ denotes the Gaussian kernel and σ regulates the blur intensity. At this point, inspired by Han *et al.* [9], we simulate continuous motion across the static background, animating *only* $\mathbf{R}_{\mathcal{G}}$ over a sequence of K frames using a smooth 2D cubic spline trajectory generated via random keypoints. At each frame i, the blurred reflection $\mathbf{R}_{\mathcal{G}}^{(i)}$ is blended with the static background $\mathbf{T}$ to synthesize the corrupted mixture frame $\mathbf{M}^{(i)}$. To construct the mixture image, we follow standard approach from the literature [15, 57] and implement a non-linear screen blending formulation such that



(a) Synthetic Data Generation. **(b)** Real-world Data Collection.

Fig. 3: Data collection overview. (a) A random trajectory simulation pipeline generates a sequence of mixture frames that are fed into the DVS-Voltmeter simulator to obtain synchronized event streams, then converted in voxels. (b) A real DAVIS346 event sensor captures real-world mixture frames and native event streams, with the help of a light-absorbing black cloth to cut out any static reflection.

$\mathbf{M}^{(i)} = \alpha\mathbf{T} + \beta\mathbf{R}_{\mathcal{G}}(i) - \alpha\mathbf{T} \cdot \beta\mathbf{R}_{\mathcal{G}}^{(i)}$, where α and β are scale factors. Finally, the generated frame sequence is fed directly into the DVS-Voltmeter simulator [29], to produce the corresponding synchronized stream of events centered on the precise midpoint frame $\mathbf{M}^{(K/2)}$. To match a realistic imaging pipeline, we accumulate the event stream over a symmetric window of ± 16 ms centered on this target frame timestamp (spanning a total duration of approximately 33 ms, equivalent to a standard 30 fps inter-frame period). This sub-stream of events is then converted into voxel grid as described in Eq. (5). Figure 3a summarizes the described synthetic protocol.

Real-World Dataset To evaluate our model in the real-world, we collect a multimodal frame-event dataset named *ER-100*. Ideally, a multimodal setup requires a high-resolution RGB camera and an event camera perfectly synchronized. However, as a preliminary benchmark, we record the data using a single commercial DAVIS346 sensor that naturally captures both asynchronous events and standard APS grayscale images on the same pixel grid at a resolution of 346×260 pixels. While we acknowledge that this represents a hardware limitation, the dataset still helps to evaluate the generalization capacity and the sim-to-real gap of our model trained on synthetic data. In the setup, we place a real glass panel between the camera and the background scene. To block any static ambient reflections, a light-absorbing black cloth is placed behind the camera and kept there for the entire recording session. Following a video-based recording protocol from [60], at the beginning of each video the scene is completely static with no reflections in order to save the clean grayscale frame as the transmission ground truth. Right after, we create dynamic reflections on the glass by moving physical objects, body parts, or playing videos on a display. Finally, we randomly sample frames to extract a diverse subset of images. Our dataset includes 5 different scenes for a total of 138 multimodal testing samples. Figure 3b illustrates the setup we use for the data collection.

6 Experimental Setup

Benchmark & Evaluation Metrics We compare our approach against five state-of-the-art single modality methods, IBCLN [22], LANet [6], DSRNet [15], Zhu *et al*. [60], and RDNet [57]. Note that event-based frameworks we mentioned in Sec. 2.2 ([23, 56, 61]) target different degradation physics (*i.e*. occlusion and flare removal) and cannot be directly included in our comparison. For all compared methods, we utilize their publicly available pre-trained weights. Furthermore, to ensure a fair evaluation, we retrained DSRNet, Zhu *et al*., and RDNet from scratch on our synthetic training dataset using their official open-source code, as they represent the most recent and competitive baselines in the literature. Our evaluation protocol is structurally divided into two distinct phases. First, we conduct a cross-dataset evaluation to investigate the generalization performance of all models on unseen RGB scenes from the widely adopted SIR^{2+} dataset [40,44]. Specifically, we evaluate the performance of our framework on an extended version of SIR^{2+} by including simulated event streams. To achieve this, we select only the subset of samples in the dataset that contain both transmission and reflection ground-truth layers, as the reflection ground truth is strictly required for the simulation process. Following the same pipeline described in Sec. 5, we simulate the event data without applying a Gaussian blurring filter this time, since the reflection images are already naturally blurred. While this cross-dataset configuration does not constitute a full zero-shot real-world test due to the simulated nature of the events, it evaluates the robustness of the proposed method on different RGB data. Finally, we perform an authentic real-world evaluation using coupled frames and event streams captured with a real event sensor.

Following literature conventions, we use PSNR (↑) and SSIM (↑) as metrics. Additionally, we incorporate the LPIPS (↓) [54] metric to assess the structural preservation and perceptual fidelity.

Implementation Details The model is implemented in PyTorch and trained on `NVIDIA L40S GPU` for 200 epochs with batch size of 4, learning rate set to 10^{-4}, `AdamW` optimiser, and `Cosine Annealing` decay to 10^{-6}. For the generation of the mixture image as described in Sec. 5, we set $\alpha \in [0.70, 0.85]$, $\beta \in [0.35, 0.60]$, $\sigma \in [1.1, 3.0]$; we use $B = 5$ for the voxel bins. RGB and voxel inputs are cropped to 320×320 and we perform data augmentation consisting of random flips applied to both modalities. In total, the training dataset consists of 9418 samples, with 80:20 train/validation split.

7 Results

7.1 Quantitative Evaluation on Synthetic Data

The quantitative results on the extended SIR^{2+} dataset are summarized in Table 1. In the evaluation process, we assign each testing sample into three categories based on event density (*i.e*. sparse, medium, and dense), allowing us to

Table 1: Quantitative comparison on the SIR^{2+} dataset across different event densities. Best and second-best results are highlighted in red and green, respectively. † and ‡ indicate pretrained models and models retrained on our synthetic dataset. *Base* refers to the original input mixture evaluated directly against the ground truth.

Methods	Event Density											
	Sparse (315)			Medium (153)			High (16)			Average (484)		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Base	25.35	0.941	0.069	22.80	0.879	0.123	22.21	0.810	0.205	24.44	0.914	0.090
LANet†	23.48	0.940	0.058	22.37	0.903	0.083	22.67	0.849	0.151	23.10	0.926	0.069
IBCLN†	26.25	0.941	0.070	24.78	0.900	0.112	23.97	0.843	0.184	25.71	0.925	0.087
DSRNet†	27.90	0.955	0.045	26.51	0.930	0.067	25.98	0.878	0.135	27.40	0.944	0.055
Zhu *et al.* †	28.97	0.950	0.062	26.86	0.921	0.088	27.57	0.892	0.123	28.26	0.939	0.072
RDNet†	29.06	0.958	0.045	28.55	0.943	0.053	29.37	0.909	0.091	28.91	0.952	0.049
DSRNet‡	26.04	0.929	0.046	28.19	0.918	0.054	28.70	0.897	0.084	26.81	0.925	0.050
Zhu *et al.* ‡	25.21	0.925	0.053	27.76	0.918	0.061	28.67	0.901	0.079	26.13	0.922	0.057
RDNet‡	26.58	0.933	0.048	29.28	0.925	0.061	29.19	0.892	0.095	27.52	0.929	0.054
E-SIRR‡	28.27	0.942	0.044	29.66	0.927	0.043	31.01	0.898	0.057	28.80	0.936	0.044

analyze the impact of the events on the reflection removal process. When events are sparse and miss reflection regions the network shows a performance degradation. In contrast, the network has a competitive advantage in situations where event density is medium/high. In general, our model achieves state-of-the-art performance, with superior results particularly in perceptual fidelity (LPIPS).

Ablation Study We conduct an ablation study to verify the importance of each design component. Table 2 summarizes the results for five different configurations. We first evaluate the network under an event dropout scenario by passing empty voxel grid at inference, demonstrating performance drop across all metrics compared to our full model; nevertheless, this configuration is competitive with some single-modality baselines, confirming that the event stream contributes to boost performance rather than compensating for a weak RGB backbone. Then, replacing the RFB head in the event encoder with a standard 3×3 convolution decreases performance. Without the confidence map in EFF the results are degraded (especially on sparse event grids), showing that the network in these cases should rely more on RGB than event data. Disabling the cross-stage connection modality underscores the importance of their role in guiding the background restoration process; finally, training without exclusion loss reduces the separation ability of the model.

7.2 Quantitative Evaluation on Real-World Data

We evaluate all models on our real-world dataset *ER-100* (Table 3). Since these images are in grayscale, we replicate the luminance across three channels to

Table 2: Ablation study of our design components on the synthetic SIR^{2+} dataset. Best results are highlighted in red.

Experiment		PSNR	SSIM	LPIPS
(a)	w/o Event Data (Dropout)	27.24	0.916	0.081
(b)	w/o RFB (3×3 Conv)	27.94	0.927	0.045
(c)	w/o Confidence Map (Zou *et al.* [61])	28.16	0.928	0.045
(d)	w/o Cross-Stage	27.69	0.921	0.046
(e)	w/o $\mathcal{L}_{\text{excl}}$	27.73	0.923	0.047
(f)	Full Model	28.80	0.936	0.044

Table 3: Quantitative comparison on *ER-100* dataset. Best and second-best results are highlighted in red and green, respectively. †, ‡ and ♯ indicate pretrained models, models retrained on our synthetic dataset, and those trained on its grayscale version. *Base* refers to the original input mixture evaluated directly against the ground truth.

Metric	Base	Methods										
		LANet†	IBCLN†	DSRNet		Zhu *et al.*		RDNet			**E-SIRR**	
				†	‡	†	‡	†	‡	♯	‡	♯
PSNR	24.47	25.01	23.95	25.20	27.31	26.50	26.49	27.37	26.67	27.33	27.91	28.10
SSIM	0.812	0.817	0.807	0.790	0.844	0.843	0.858	0.846	0.842	0.851	0.863	0.872
LPIPS	0.084	0.089	0.091	0.106	0.063	0.073	0.063	0.069	0.068	0.064	0.054	0.051

match the expected input shape of the networks without changing their architectures. For our model, we test the version trained on synthetic data and also the one trained on the grayscale version of the same synthetic dataset. DSRNet and Zhu *et al.* are evaluated in both pre-trained and retrained configuration, as well as RDNet which is also retrained on the grayscale synthetic data, as it represents the strongest competitor; for LANet and IBCLN, we consider their pretrained models. To construct the event voxel input from the real stream, we use the exact same synchronization protocol as for the synthetic dataset. Our model in both training configurations outperforms the competitors on average across all metrics, highlighting its ability to handle complex real-world dynamics.

7.3 Visual Comparison

We present a visual comparison of the reconstructed reflection-free images in Fig. 4 and 5. Due to space constraints, we focus our comparison on DSRNet, Zhu *et al.*, and RDNet. For each sample, we select the specific configuration of each competitor (pre-trained or retrained) that achieves the highest score in metrics. This creates a challenging baseline comparison, as we always show their best inference results. When evaluated on the synthetic SIR^{2+} benchmark (Fig. 4), models like DSRNet and Zhu *et al.* struggle in removing high-contrast reflections, leaving visible text residues and structured patterns overlapping with the target background. We also notice that while RDNet alleviate these artifacts,

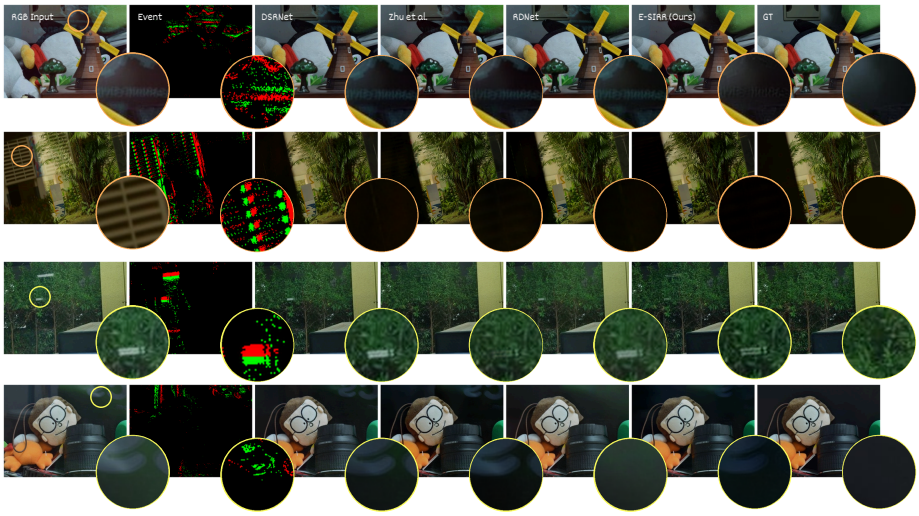


Fig. 4: Qualitative comparison on synthetic data from SIR^{2+}. From left to right: Input, Event Voxel, DSRNet, Zhu *et al.*, RDNet, E-SIRR (Ours), and Ground Truth. State-of-the-art models results are shown in their best-performing configuration († or ‡).

it also over-smooths structural textures, ending up sometimes to remove sharp background details. Our framework can isolate the reflection layer, provided that the event voxel contains sufficient information. This layer separation capability translates to the grayscale real-world dataset *ER-100* (Fig. 5). RGB baselines often alter background structures, instead our model shows a more controlled and explainable behavior as it does not change large pixel areas to obtain artificially smooth images, with the drawback that sometimes it does not completely eliminate all artifacts (*e.g.* particularly in regions where sparse event sampling does not provide sufficient information).

8 Limitations and Future Work

We acknowledge that there are two main limitations in our approach. First, the static background hypothesis with only moving reflections fails when both layers are stationary (*e.g.* fixed objects/lights). A possible improvement could be to consider camera motion that would trigger events for both layers but with different spatial patterns and apparent velocities, since reflections and the background lie at different depths. Second, our real-world dataset is limited to low-resolution and grayscale images. Building a true RGB-Event dataset is a crucial next step, using either a color event sensor or a beam-splitter stereo rig to perfectly align a standard camera with an event sensor.

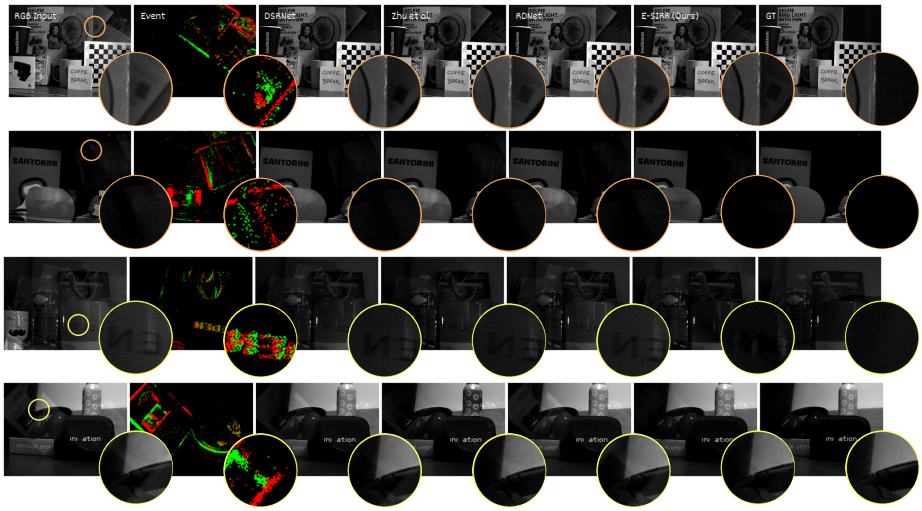


Fig. 5: Qualitative comparison on real-world data from *ER-100*. From left to right: Input, Event Voxel, DSRNet, Zhu *et al.*, RDNet, E-SIRR (Ours), and Ground Truth. State-of-the-art models and our results are shown in their best-performing configuration (†, ‡ or ♯).

9 Conclusion

In this paper we present the first event-guided framework to address the task of reflection removal which exploits the precise spatiotemporal cues of an event camera to resolve structural layer ambiguity. Our network integrates the multimodal inputs through cross-modal feature fusion to explicitly estimate the reflection and the background layers. To support this architecture, we also introduce a synthetic dataset and a preliminary real-world multimodal benchmark, with the conducted experiments that demonstrate the soundness of the proposed method, opening interesting directions for future research.

Acknowledgements

This work was partly supported by the European Union's Horizon Europe research and innovation program under Grant Agreement No 101094831 for the Converge-Telecommunications and Computer Vision Convergence Tools for Research Infrastructures project.

References

1. Adra, M., Melcarne, S., Mirabet-Herranz, N., Dugelay, J.L.: Event-based solutions for human-centered applications: a comprehensive review. Frontiers in Signal Processing **5**, 1585242 (2025)

2. Arvanitopoulos, N., Achanta, R., Susstrunk, S.: Single image reflection suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4498–4506 (2017)
3. Chen, H., Teng, M., Shi, B., Wang, Y., Huang, T.: A residual learning approach to deblur and generate high frame rate video with an event camera. IEEE Transactions on Multimedia **25**, 5826–5839 (2023)
4. Chen, K., Yu, L.Y.: Motion deblur by learning residual from events. IEEE Transactions on Multimedia **26**, 6632–6647 (2024)
5. Cui, M., Wang, Z., Wang, D., Zhao, B., Li, X.: Color event enhanced single-exposure hdr imaging. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1399–1407 (2024). https://doi.org/10.1609/aaai.v38i2.27904, https://doi.org/10.1609/aaai.v38i2.27904
6. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5017–5026 (2021)
7. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 154–180 (2022). https://doi.org/10.1109/TPAMI.2020.3008413
8. Guo, G., Feng, Y., Lv, H., Zhao, Y., Liu, H., Bi, G.: Event-guided image super-resolution reconstruction. Sensors **23**(4), 2155 (February 2023), https://www.mdpi.com/1424-8220/23/4/2155
9. Han, J., Yang, Y., Duan, P., Zhou, C., Ma, L., Xu, C., Huang, T., Sato, I., Shi, B.: Hybrid high dynamic range imaging fusing neuromorphic and conventional images. IEEE Transactions on pattern analysis and machine intelligence **45**(7), 8553–8565 (2023)
10. Han, J., Yang, Y.t.l., Zhou, C., Xu, C., Shi, B.: Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4862–4871 (October 2021)
11. Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xia, C., Shi, B.: Neuromorphic camera guided high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1730–1739 (June 2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hong, Y., Lyu, Y., Li, S., Cao, G., Shi, B.: Reflection removal with nir and rgb image feature fusion. IEEE Transactions on Multimedia **25**, 7101–7112 (2023). https://doi.org/10.1109/TMM.2022.3217446
14. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation (2021), https://arxiv.org/abs/2110.10546
15. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13138–13147 (2023)
16. Jiang, Y., Wang, Y., Li, S., Zhang, Y., Zhao, M., Gao, Y.: Event-based low-illumination image enhancement. IEEE Transactions on Multimedia **26**, 1920–1931 (2024)
17. Jiang, Z., Zhang, Y., Zou, D.Z., Ren, J.S., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3320–3329 (June 2020)

18. Jing, Y., Yang, Y., Wang, X., Song, M., Tao, D.: Turning frequency to resolution: Video super-resolution via event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7768–7777 (June 2021)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution (2016), https://arxiv.org/abs/1603.08155
20. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14811–14820 (2021)
21. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(9), 1647–1654 (2007)
22. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3565–3574 (2020)
23. Li, S.Q., Gao, Y., Dai, Q.H.: Image de-occlusion via event-enhanced multi-modal fusion hybrid network. Machine Intelligence Research **19**(4), 307–318 (2022)
24. Li, Y., Brown, M.S.: Single image layer separation using relative smoothness. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2752–2759 (2014)
25. Li, Y., Liu, M., Yi, Y., Li, Q., Ren, D., Zuo, W.: Two-stage single image reflection removal with reflection-aware guidance. Applied Intelligence **53**(16), 19433–19448 (Mar 2023). https://doi.org/10.1007/s10489-022-04391-6, https://doi.org/10.1007/s10489-022-04391-6
26. Li, Y., Liu, M., Yi, Y., Li, Q., Ren, D., Zuo, W.: Two-stage single image reflection removal with reflection-aware guidance. Applied Intelligence **53**(16), 19433–19448 (2023)
27. Li, Y., Yan, Q., Zhang, K., Xu, H.: Image reflection removal via contextual feature fusion pyramid and task-driven regularization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 553–565 (2022). https://doi.org/10.1109/TCSVT.2021.3063308
28. Lin, S., Zhang, J., Yu, J., Zhou, D., Cui, D., Jia, W., Wang, X., Xu, L., Zhang, Y.J.: Learning event-driven video deblurring and interpolation. In: Proceedings of the European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 12363, pp. 695–710. Springer (August 2020)
29. Lin, S., Ma, Y., Guo, Z., Wen, B.: Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In: European Conference on Computer Vision. pp. 578–593. Springer (2022)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
31. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 385–400 (2018)
32. Melcarne, S., Dugelay, J.L.: Fusing thermal and event data for visible spectrum image reconstruction. In: VISAPP 2026, 21st International Conference on Computer Vision Theory and Applications. pp. 661–669. SCITEPRESS-Science and Technology Publications (2026)
33. Mostafavi, M., Nam, Y., Choi, J., Yoon, K.J.: E2sri: Learning to super-resolve intensity images from events. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **44**(10), 6890–6909 (October 2022)

34. Shang, W., Ren, D., Zou, D.Z., Ren, J.S., Luo, P., Zuo, W.: Bringing events into video deblurring with non-consecutively blurry frames. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4511–4520 (October 2021)
35. Shaw, R., Catley-Chandar, S., Leonardis, A., Perez-Pellitero, E.: Hdr reconstruction from bracketed exposures and events (2022), https://arxiv.org/abs/2203.14825
36. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3193–3201 (2015)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), https://arxiv.org/abs/1409.1556
38. Song, B., Zhou, J., Chen, X., Zhang, S.: Real-scene reflection removal with raw-rgb image pairs. IEEE Transactions on Circuits and Systems for Video Technology **33**(8), 3759–3773 (2023). https://doi.org/10.1109/TCSVT.2023.3241319
39. Sun, L., Bao, Y., Zhai, J., Liang, J., Zhang, Y., Wang, K., Paudel, D.P., Gool, L.V.: Low-light image enhancement using event-based illumination estimation (2025), https://arxiv.org/abs/2504.09379
40. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: Proceedings of the IEEE international conference on computer vision. pp. 3922–3930 (2017)
41. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Crrn: Multi-scale guided concurrent reflection removal network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4777–4785 (2018)
42. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Reflection scene separation from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
43. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Kot, A.C.: Corrn: Cooperative reflection removal network. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(12), 2969–2982 (2020). https://doi.org/10.1109/TPAMI.2019.2921574
44. Wan, R., Shi, B., Li, H., Hong, Y., Duan, L.Y., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: IEEE Transacations on Pattern Analysis and Machine Intelligence (2022)
45. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8178–8187 (2019)
46. Wen, Q., Tan, Y., Qin, J., Liu, W., Han, G., He, S.: Single image reflection removal beyond linearity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3771–3779 (2019)
47. Weng, J., Li, B., Huang, K.: Event-based image enhancement under high dynamic range scenarios. In: Proceedings of the Asian Conference on Computer Vision. pp. 2456–2470 (2024)
48. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
49. Xu, S., Sun, Z., Liu, K., Lu, X., Jiang, R., Fu, X., Zha, Z.J.: Event-illumination collaborative low-light image enhancement with a high-resolution real-world dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22270–22280 (2026)

50. Yang, J., Li, H., Dai, Y., Tan, R.T.: Robust optical flow estimation of double-layer images under transparency or reflection (2016), https://arxiv.org/abs/1605.01825
51. Yang, J., Gong, D., Liu, L., Shi, Q.: Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In: Proceedings of the european conference on computer vision (ECCV). pp. 654–669 (2018)
52. Yang, K., Sun, H., Cai, J., Fu, L., Ding, J., Li, J., Meng, Z.: Survey on single-image reflection removal using deep learning techniques. In: 2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 20–26. IEEE (2025)
53. Yang, Y., Han, J., Liang, J., Sato, I.t.l., Shi, B.: Learning event guided high dynamic range video reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13924–13934 (June 2023)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
55. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4786–4794 (2018)
56. Zhang, Y., Chen, Y., Zhou, J., Peng, W., Liu, X., Ma, Z., Guo, Y.: Nighttime reflective flare removal from images with event camera. IEEE Transactions on Instrumentation and Measurement **PP**, 1–1 (01 2026). https://doi.org/10.1109/TIM.2026.3687977
57. Zhao, H., Li, M., Hu, Q., Guo, X.: Reversible decoupling network for single image reflection removal. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26430–26439 (2025)
58. Zhou, C., Zhang, Y., Zhang, J., Lin, S., Yu, J., Xu, L., Wang, X., Huang, T., Shi, B.: Deblurring low-light images with events. International Journal of Computer Vision (IJCV) **131**(5), 1284–1298 (May 2023)
59. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 989–997 (2019)
60. Zhu, Y., Fu, X., Jiang, P.T., Zhang, H., Sun, Q., Chen, J., Zha, Z.J., Li, B.: Revisiting single image reflection removal in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25468–25478 (2024)
61. Zou, R., Muglikar, M., Messikommer, N., Scaramuzza, D.: Seeing behind dynamic occlusions with event cameras (2023), https://arxiv.org/abs/2307.15829