

DELUX: Cross-Modal Local Artifact Restoration in Video Using Neuromorphic Data

Bartosz Stachowiak¹ and Dariusz Brzezinski¹

Institute of Computing Science, Poznan University of Technology,
ul. Piotrowo 2, 60-965 Poznan, Poland
{bartosz.stachowiak,dariusz.brzezinski}@cs.put.poznan.pl

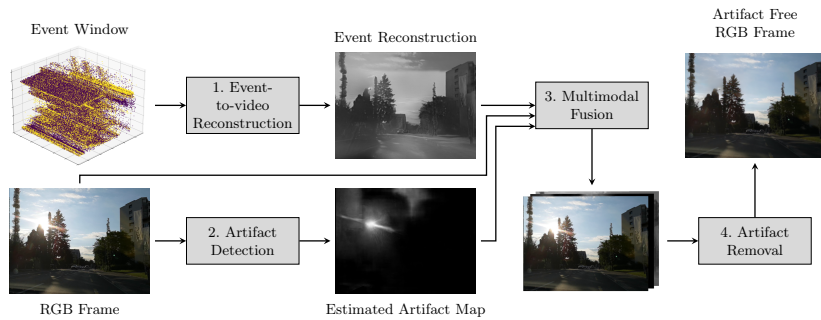


Fig. 1: Overview of the proposed DELUX framework for cross-modal local artifact restoration. To address the challenge of localized image information loss, our modular pipeline explicitly decouples artifact detection from multimodal fusion and image inpainting. As illustrated, an RGB frame corrupted by glare is processed alongside a corresponding neuromorphic event window. By using event-to-video reconstruction as a prior and explicitly isolating the degraded region via an estimated artifact map, the multimodal fusion stage guides the targeted inpainting of corrupted RGB pixels.

Abstract. Conventional RGB cameras suffer from lighting artifacts such as flare, glare, flicker, and overexposure, leading to irrecoverable information loss that necessitates computational restoration. However, existing approaches treat these problems in isolation, failing to recover structural details completely obscured by complex spatially discrete image degradations. In this paper, we propose a novel cross-modal restoration paradigm and present DELUX, a modular proof-of-concept pipeline that leverages neuromorphic event streams as a structural prior to guide the targeted detection and inpainting of lighting artifacts in RGB video. Validation on synthetic benchmarks and real-world automotive footage demonstrates that DELUX effectively suppresses local artifacts and restores affected regions. The proposed approach outperforms existing RGB-only baselines and event-guided HDR models, achieving an average MS-SSIM of over 0.99 across all artifact types and demonstrating up to an 88% reduction

in artifact severity in real-world automotive footage. The synthetic artifact generation tools and curated real-world evaluation datasets are made publicly available to foster future research on cross-modal restoration.

Keywords: cross-modal restoration, lighting artifact removal, event-based vision, multimodal fusion, video inpainting

1 Introduction

Outdoor scenes with extreme dynamic range present a persistent challenge for conventional cameras. When exposed to intense or variable illumination, RGB sensors often produce *lighting artifacts* such as lens flare [47], glare [34], overexposure [19], and flicker [11] (Fig. 2). These artifacts obscure critical visual details, distort local color and contrast, and, in extreme cases, completely saturate the image, making recovery through RGB-only post-processing impossible. Such degradations can severely limit the reliability of visual perception systems in applications such as autonomous driving and robotics [4, 23].

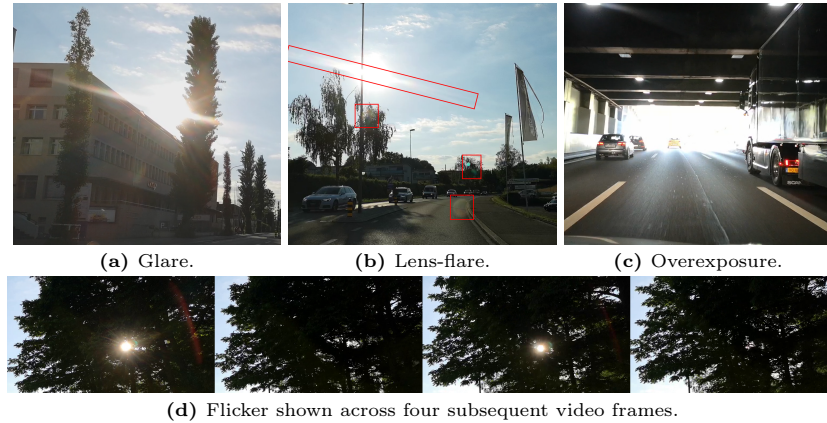


Fig. 2: Examples of lighting artifacts.

However, existing image and video restoration methods address these artifacts only *in isolation*, for example, by mitigating only glare [21, 34, 40], flare [7, 8, 47], global overexposure [10, 24, 27, 43, 51–53] or flicker [11, 20, 22]. Moreover, there is a difference in how these lighting problems are tackled. High dynamic range (HDR) reconstruction methods perform global tone mapping by compressing the dynamic range but inherently preserve the structure of optical artifacts, such as flare rays or glare halos. Conversely, targeted artifact removal methods rely solely on frame-based RGB information, which lacks the dynamic range needed to reconstruct regions where pixel intensities are fully clipped. Therefore, a main challenge in the literature is a highly fragmented approach to image

degradation. Because localized optical anomalies and extreme dynamic range deficiencies are treated as disjoint sub-fields, the literature lacks a generalized paradigm capable of addressing diverse lighting artifacts simultaneously.

In this work, we propose a new research direction that combines the challenges of all the mentioned image degradations: *cross-modal local artifact restoration*. In this problem setting, localized lighting artifacts represent a modality-specific information loss that can be mitigated by integrating a secondary sensing modality. To validate this concept, we use neuromorphic event cameras, which asynchronously capture changes in brightness with microsecond precision and an effective dynamic range exceeding 140 dB [55], while remaining largely unaffected by glare and saturation even when RGB sensors saturate (Fig. 3).

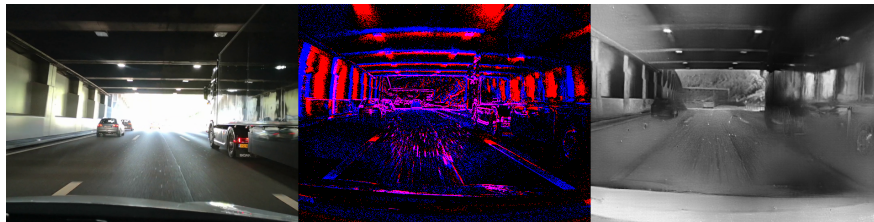


Fig. 3: Comparison showing that event-based data can recover details obscured by lighting artifacts. The overexposed RGB image (left) loses all structure at the tunnel exit, while the corresponding raw event data (center) retain brightness changes. The event-based reconstruction (right) restores the scene details.

To further validate the feasibility of treating various lighting artifacts as instances of the proposed cross-modal local artifact restoration problem, we present DELUX, an end-to-end, modular proof-of-concept pipeline for lighting artifact mitigation that jointly leverages RGB video and neuromorphic event data (Fig. 1). DELUX explicitly isolates the detection of lighting artifacts from the reconstruction of missing image details, using event data to guide the inpainting of corrupted RGB regions. To the best of our knowledge, DELUX is the first system designed to detect and remove multiple types of lighting artifacts within a single deep learning pipeline.

The key contributions of our work are as follows:

- We introduce a novel research direction for targeted removal of lighting artifacts that leverages complementary sensor modalities to guide RGB inpainting, moving beyond global tone-mapping approaches and single modality image restoration.
- We propose DELUX, a modular pipeline that demonstrates the feasibility of this paradigm by fusing standard video with neuromorphic event data to jointly detect and remove diverse lighting artifacts.
- We establish a rigorous benchmark comprising synthetically generated and real-world lighting artifacts. We demonstrate the efficacy of our approach

against state-of-the-art restoration and event-guided HDR baselines, providing an evaluation protocol for future cross-modal artifact restoration models.

2 Related Works

Glare Removal. Glare, caused by direct light sources such as the sun or headlights, manifests itself as haze or rays that obscure the details of the scene [34] (Fig. 2a). Early methods focused on reflective glare in medical imaging [5, 21]. For outdoor veiling glare, Shoshin et al. [40] trained a UNet with dual branches on synthetic data to reduce haze and restore intensity, although prominent rays persisted. CR-GAN [34] employs a GAN with a Cascaded Pyramid Neck and Glare Attention Detector, trained on synthetic and real paired data (e.g., via impure glass), achieving better perceptual quality and precise reconstruction. Furthermore, general-purpose restoration frameworks such as Deep Adversarial Decomposition (DAD) [57] and Uformer [45] offer flexible architectures that can be adapted to remove glare. DAD formulates image separation as a generative adversarial task, using a learned decomposition prior to cross-road loss to disentangle superimposed layers, which could be extended to isolate glare components. Uformer, a Transformer-based U-Net variant, leverages window-based attention and multi-scale modulation to restore fine details across diverse degradation types, making it suitable for training on glare-specific artifacts. These approaches (excluding Uformer) use artifact detectors, but focus solely on glare and rely on RGB, which limits recovery in fully saturated areas. Our work extends this by incorporating neuromorphic data for cross-modal restoration.

Flare Removal. Lens flare arises from internal reflections within the camera lens, creating ghosts or streaks (Fig. 2b). Physical solutions such as anti-reflective coatings [3] are impractical for post-capture correction. Computational methods include Wu et al.’s [47] UNet pipeline, which removes flares in a semi-supervised fashion by using synthetic masks that blend backlight sources and introduce flares to the training data. Matta et al. [31] use Generalizable NeRFs with a transformer to model flares in 3D scenes, supported by a multi-flare dataset. For nighttime flares, Dai et al.’s Flare7K [7] enables training of MFDNet [16], a Laplacian pyramid-based model for low-light distortions. We have incorporated flares from Flare7K++ [8] into our training pipeline, but contrary to the works mentioned above, we additionally use event streams and restore the underlying images rather than simply performing RGB-only inpainting.

Overexposure. Overexposure in high dynamic range (HDR) scenes saturates bright areas, hiding image details (Fig. 2c). Multi-image HDR combines LDR exposures [19, 32, 46], often with CNNs for alignment. Single-image methods map LDR to HDR using UNets [10, 30, 49, 53] or reverse pipelines [28]. SHDR [27] explicitly models the inverse camera pipeline—dequantization, linearization, and hallucination—using dedicated CNNs, improving HDR reconstruction fidelity across diverse scenes. Related works enhance low-light [48], combine

over/underexposed frames [54], or inpaint overexposed regions with transformers [50] or GANs [35]. BacklitNet [29] targets overexposure with a dual-UNet. Recent methods, such as HDRev [51] and HDRev-Diff [52], use event streams to guide HDR reconstruction. However, like all other approaches that tackle overexposure, they are designed to perform global tone mapping and dynamic range expansion. Because they do not separate the underlying scene from superimposed degradations, they inherently preserve and reproduce lighting artifacts, such as lens flare and veiling glare, within the output image. In contrast, our study addresses the problem of explicit suppression of local artifacts.

Flicker Removal. Flicker is usually defined as periodic intensity variations from lights or water refraction (Fig. 2d). For underwater sunflicker, Shihavuddin et al. [39] predict patterns with an open-loop linear dynamic system from prior frames. In high-speed video, Kanj et al. [20] use SLIC superpixels and color correction matrices. These require tuning and are computationally heavy, focusing on compensation rather than artifact removal. Deflicker [22] is a blind deflickering framework that leverages a flawed neural atlas and a filtering strategy to enforce temporal consistency without requiring prior knowledge.

Neuromorphic Vision. Neuromorphic cameras capture asynchronous events with high temporal resolution and a wide dynamic range [12]. Event representations such as voxel grids [56] enable integration with CNNs/Transformers for tasks such as deblurring [17, 41], superresolution [33], depth estimation [26], and HDR reconstruction [37, 43, 52]. Datasets like DSEC [13] and E2VID [36] provide paired RGB event streams, while simulators like CARLA [9] generate synthetic data. We use voxel grids with E2VID++ [36, 42] for artifact-free grayscale reconstruction, combining it with RGB via a ResUNet [14]. Previous event-HDR works [32, 37] inspire our approach, but they focus on global enhancement rather than local artifact removal. The proposed DELUX pipeline represents a departure from global enhancement, serving instead as a cross-modal local artifact restoration framework for lighting anomalies.

3 DeLux Pipeline

To demonstrate the utility of the cross-modal local artifact restoration paradigm and to address the challenge of information loss in RGB frames corrupted by lighting artifacts, we introduce DELUX, a pipeline that leverages event camera data to recover scene details. Unlike prior methods that rely solely on RGB input or global tone mapping, our approach uses cross-modal cues to localize and suppress artifacts at their source.

DELUX is designed around four core stages (Fig. 1): (1) *event-to-video reconstruction*, (2) *artifact detection*, (3) *multimodal fusion*, and (4) *artifact removal*. Each stage is implemented as a standalone module, allowing flexible substitution or extension. This modularity enables future integration of alternative modalities, such as thermal imagery or manual annotations, thereby making DELUX a template for cross-modal local artifact removal.

3.1 Event-to-video Reconstruction

The first stage of the DELUX pipeline transforms neuromorphic event streams into dense grayscale video frames that serve as a complement to RGB input. These reconstructions are critical for restoring regions affected by lighting artifacts, particularly those that are saturated or flickering, where RGB data alone fail to preserve scene details.

We utilize pre-trained E2VID [36] and E2VID++ [42] models that convert asynchronous event data into intensity frames. The input to this model is a voxel grid representation of events, constructed by binning spatio-temporal event tuples into a fixed number of temporal slices. Each voxel encodes the polarity and timestamp of brightness changes at a given pixel, capturing fine-grained motion and illumination dynamics. The E2VID models synthesize grayscale frames from these voxel inputs. The output frames are temporally aligned with the RGB stream and serve as a secondary modality for both detection and fusion. Additional details of the event reconstruction and spatiotemporal alignment can be found in the Supplementary, Section A.

Importantly, this stage is modular—the event-to-video reconstructor can be replaced with alternative modalities (e.g., thermal imagery [1]), provided that they offer complementary information in artifact-prone regions. This design choice ensures that DELUX remains adaptable to diverse sensor configurations and application domains.

3.2 Artifact Detection

The second stage of the DELUX pipeline identifies regions affected by lighting artifacts such as glare, flare, flicker, and overexposure. These artifacts manifest locally and vary in spatial characteristics, making their detection a non-trivial task. To address this, DELUX uses a dedicated artifact detector implemented as a UNet-based segmentation network.

The detector is trained on synthetically generated artifact masks, allowing it to generalize across diverse lighting conditions and artifact types. Synthetic training data includes controlled examples of saturation, flicker patterns, glare streaks, and lens flare overlays. The input to the detector is a single RGB frame, and the output is a pixel-wise confidence map that indicates the likelihood of artifact presence. These maps are used in subsequent stages to guide the image restoration, ensuring that corrections are applied only where necessary.

Unlike prior approaches that embed artifact detection implicitly within a GAN or restoration network, DELUX treats detection as an explicit and interpretable step. This separation improves transparency, facilitates debugging, and enables a targeted evaluation of detection performance against synthetic ground-truth artifact masks.

3.3 Multimodal Fusion

The third stage of the DELUX pipeline integrates information from three modalities—RGB frames, secondary modality (in this case event-based reconstructions), and

artifact masks—into a unified representation that guides artifact-aware restoration. This fusion is implemented as a convolutional module designed to selectively leverage the strengths of each input: the spatial richness of RGB, the temporal and dynamic range resilience of event reconstructions, and the localization precision of the artifact detector.

The three inputs are concatenated along the channel dimension and passed through a series of convolutional layers that learn to blend them into a coherent intermediate representation. This fusion strategy is fully differentiable and is trained together with the artifact removal module (Section 3.4). Rather than relying on learned attention, the module blends the three channel-concatenated representations through stacked convolutions, while the soft artifact mask remains exposed as a manual-override interface. This makes fusion behavior configurable at inference *without retraining*, where supplying an edited or empty mask predictably changes which regions are modified (Supplementary Section E). By explicitly conditioning the fusion on artifact localization, DELUX avoids overcorrecting clean regions. This targeted blending is critical for preserving fine details and maintaining perceptual consistency (Section 4.4, Supplementary Section B).

3.4 Artifact Removal

The final stage of DELUX performs targeted restoration of RGB frames by removing lighting artifacts identified in earlier stages. This module receives the fused representation, comprising the original RGB frame, the event-based reconstruction, and the artifact mask, and outputs a clean, artifact-free RGB image.

Artifact removal is implemented using a UNet-based architecture trained to inpaint corrupted regions while preserving unaffected content. To encourage perceptual consistency and structural fidelity, the model is trained using a combination of pixel-wise losses, structural, and perceptual similarity metrics. In addition, the pipeline supports optional use of a non-mask loss term to regularize global appearance and prevent overfitting to masked regions.

The modular design of this stage enables adaptation to various artifact types or restoration goals. For example, the removal network can be retrained to target specific distortions such as flare streaks or flicker bands, or extended to operate on alternative input modalities. This flexibility makes DELUX suitable for a wider range of configurations.

3.5 Training Procedure

The DELUX pipeline is trained to jointly optimize the artifact detector, the multimodal fusion module, and the artifact remover (Fig. 4). Only the event-to-video reconstructor (E2VID or E2VID++) is frozen throughout the training, providing a stable grayscale representation of events from the neuromorphic camera. Joint optimization encourages the detector and reconstructor to co-adapt, ensuring that artifact localization and restoration are mutually consistent.

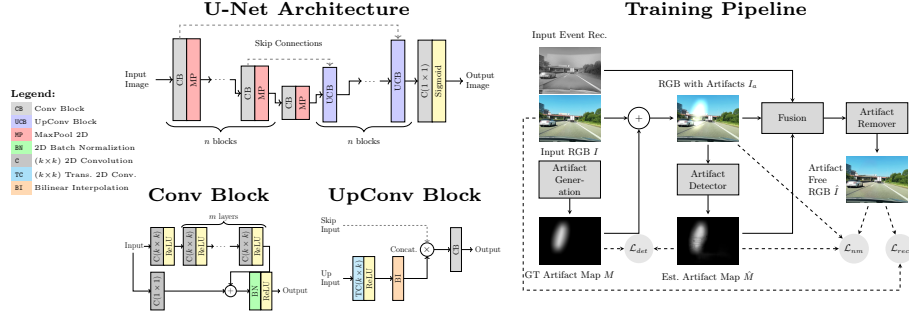


Fig. 4: DELUX training pipeline. The architecture diagram (left) illustrates the shared U-Net backbone and its building blocks, parameterized by the number of convolutional blocks n , block depth m , and kernel size k . The training pipeline (right) depicts the end-to-end joint optimization setup with synthetic artifact generation, multimodal inputs, and the composite loss functions defined in Section 3.5.

Synthetic Artifact Generation. To provide diverse and controlled supervision, we employ a custom artifact generation pipeline that simulates lighting distortions directly on clean RGB frames. The generators produce glare, flicker, overexposure, and saturation artifacts through intensity modulation, procedural glare streak synthesis, and localized dynamic brightness fluctuations. We additionally use high-quality, physically plausible flare patterns sourced from the Flare7K++ [8] and Google-Flare [47] datasets, ensuring a broad range of artifact shapes, scales, and illumination dynamics. Corresponding artifact masks are recorded along with the corrupted images to serve as ground truth for artifact detection. The event-based reconstructions are obtained from real neuromorphic datasets (e.g., DSEC [13], E2VID [36]) or generated using the CARLA simulator [9], providing authentic multimodal conditions. This hybrid data strategy ensures that DELUX generalizes effectively to real-world lighting artifacts.

Optimization Objective. The detector, fusion, and reconstructor modules are trained jointly under a unified objective that balances artifact localization and image reconstruction quality. The total loss ($\mathcal{L}_{total}$) combines three complementary components, i.e., detection ($\mathcal{L}_{det}$), reconstruction ($\mathcal{L}_{recon}$), and a non-mask consistency term ($\mathcal{L}_{nm}$):

$$\mathcal{L}_{total}(I, I_a, \hat{I}, M, \hat{M}) = \mathcal{L}_{det}(\hat{M}, M) + \mathcal{L}_{recon}(\hat{I}, I) + \mathcal{L}_{nm}(\hat{I}, I_a, \hat{M}), \quad (1)$$

where I is the clean ground-truth frame, I_a the synthetically corrupted input frame, $\hat{I}$ the reconstructed artifact-free RGB frame, M the ground-truth artifact mask, and $\hat{M}$ the predicted artifact confidence map.

The detection loss $\mathcal{L}_{det}$ is a Huber loss [15] between $\hat{M}$ and M , using $\delta=0.1$ to maintain sensitivity to small prediction errors. The reconstruction loss $\mathcal{L}_{recon}$

combines pixel-wise, perceptual, and structural similarity terms:

$$\mathcal{L}_{\text{recon}}(\hat{I}, I) = \lambda_1 \text{MAE}(\hat{I}, I) + \lambda_2 (1 - \text{MS-SSIM}(\hat{I}, I)) + \lambda_3 \text{VGG}(\hat{I}, I) + \lambda_4 \text{TV}(\hat{I}), \quad (2)$$

where MAE encourages accurate pixel reconstruction, MS-SSIM promotes perceptual and structural fidelity [44], VGG denotes the perceptual feature loss [18], and TV regularizes spatial smoothness [38]. In this study, we manually set the coefficients to $\lambda_1=1.0$, $\lambda_2=0.5$, $\lambda_3=0.1$, $\lambda_4=0.05$.

The non-mask loss $\mathcal{L}_{\text{nm}}$ penalizes undesired edits outside artifact regions and promotes consistency in clean areas:

$$\mathcal{L}_{\text{nm}}(\hat{I}, I_a, \hat{M}) = \text{MAE}(\hat{I}, I_a) \odot (1 - \hat{M}) + 0.5 \cdot (1 - \text{MS-SSIM}((1 - \hat{M}) \odot \hat{I}, (1 - \hat{M}) \odot I_a)), \quad (3)$$

where $\odot$ denotes element-wise multiplication. This term guides the reconstructor to modify only artifact-corrupted regions, preserving color and texture in unaffected areas, at the cost of less accurate artifact detection maps.

Training Details. The network was trained using the Adam optimizer ($\beta_1=0.9$, $\beta_2=0.999$) with an initial learning rate of 10^{-4} . Training was performed for 200 epochs on an NVIDIA A100 GPU with a batch size of 8. Each batch contained temporally aligned RGB frames, event-based reconstructions, and artifact masks. Data augmentation included random cropping, horizontal flipping, exposure jitter, and brightness perturbation to improve robustness to illumination changes. A hold-out validation split was used to monitor convergence, and the best model was selected based on the minimum reconstruction loss $\mathcal{L}_{\text{recon}}$. The hyperparameters of the UNet were optimized with Optuna [2], resulting in $n=3$, $m=2$, $k=7$ for the detector and $n=2$, $m=3$, $k=7$ for the reconstructor (Fig. 4).

Training used a composite dataset combining real and synthetic event-based data for robustness across domains. Only artifact-free recordings from **E2VID** [36] and **DSEC** [13] were used, with synthetic lighting artifacts injected during training. Additional data were generated using the **CARLA** autonomous driving simulator [9] to provide perfectly aligned RGB-event pairs, and the **Cityscapes** [6] dataset contributed close to 5,000 RGB frames to which we created synthetic grayscale reconstructions that simulate event data. All datasets were standardized to 640×480 resolution, with recordings downsampled to every third frame (102 ms for E2VID, 150 ms for DSEC and CARLA) to ensure diversity in terms of motion, illumination, and scene content. The entire training set consisted of 9,540 RGB-event reconstruction frame pairs.¹

¹ The code and data for training DeLux are available at: <https://github.com/Tremirre/event-sun-effects-remover>.

4 Experiments

4.1 Experimental Setup

Datasets. We used two complementary groups of datasets, both distinct from all training data. The first consisted of 1,724 artifact-free frames from separate recordings, to which synthetic artifacts were injected using the same generation procedure as in training. The second comprised 11,056 frames from eight real-world recordings with lighting artifacts: six from the **E2VID** dataset [36] and two from **DSEC** [13]. Importantly, these recordings represent real-world driving scenarios exhibiting unconstrained motion and dynamic lighting. We isolated specific temporal windows in which severe lighting artifacts occur natively, establishing the first targeted evaluation subset for localized artifact removal in real-world automotive contexts.

Baseline Methods. To evaluate our DELUX pipeline, we compared it with a diverse set of representative baselines tailored to different artifact type. For lens flare and glare removal, we selected Flare7K++ [8] and a model developed by Wu et al. [47]. For overexposure correction, we employed SHDR [27], with HDR to PNG conversion done using the algorithm outlined by Liang et al. [25]. For global HDR reconstruction using event data, we used HDRRev-Diff [52]. Due to the scarcity of pretrained open-source models specifically designed for the analyzed artifacts, we additionally trained a variant of the universal DAD framework [57] (denoted DAD[†]) to separate artifacts from the underlying clean image based on the RGB and grayscale event reconstructions. Notably, the DAD variant was trained on the same data as DELUX and the proposed training pipeline. Therefore, DAD provides a direct comparison between an existing single neural network and the proposed modular approach.

Evaluation Metrics. The tested models were separately evaluated in terms of artifact *detection* and *removal*.

For evaluating artifact detection, we used synthetic datasets where ground-truth artifact-free images were available. The detection performance was assessed using the accuracy and F1 score of the predicted artifact maps.

The efficacy of artifact removal was quantified using both synthetic and real-world data. For the synthetic data, we computed the Multi-Scale Structural Similarity Index Measure (MS-SSIM) [44], Peak Signal-to-Noise Ratio (PSNR), and Mean Absolute Percentage Error (MAPE) to compare the restored final images against the corresponding clean reference frames. In contrast, real-world data lack corresponding artifact-free ground truth, necessitating the use of alternative quality assessment strategies. As such, we introduce a custom metrics: Strong Artifact Suppression (SAS, Eq. 4). SAS depends on a reliable artifact detector, as it compares detection maps before ($\hat{M}_{input}$) and after ($\hat{M}_{output}$) processing by the model.

$$\text{SAS} = P(\hat{M}_{input} > 0.5) - P(\hat{M}_{output} > 0.5) \quad (4)$$

Table 1: Detection results on the synthetic test set across different artifact types. Best values are highlighted in **bold**, second-best values are underlined.

Metric	Category	DAD [†] [57]	F7K [8]	Wu et al. [47]	DeLux	DeLux-D
Accuracy $\uparrow$	Glare	<u>0.969</u>	0.914	0.830	0.964	0.985
	HQ Flares	<u>0.938</u>	0.920	0.779	0.891	0.959
	Overexposure	<u>0.972</u>	0.862	0.819	0.943	0.982
	Simple Flares	0.973	0.991	0.880	0.977	<u>0.989</u>
	Flares+Glare	0.936	0.835	0.799	<u>0.939</u>	0.973
	Overall	<u>0.949</u>	0.915	0.809	0.925	0.971
F1-score $\uparrow$	Glare	0.805	0.145	0.339	<u>0.815</u>	0.900
	HQ Flares	<u>0.601</u>	0.292	0.301	0.530	0.739
	Overexposure	<u>0.793</u>	0.188	0.409	0.719	0.841
	Simple Flares	0.484	0.065	0.045	<u>0.574</u>	0.749
	Flares+Glare	0.744	0.191	0.403	<u>0.745</u>	0.871
	Overall	<u>0.634</u>	0.212	0.281	0.613	0.785

As the artifact detector, we used the DeLUX with the non-mask loss $\mathcal{L}_{nm}$ turned off (denoted as DeLUX-D), as it produced the most accurate detection maps in our ablation studies (see Section 4.4). Because SAS is computed from a learned detector, we treat it as a relative artifact-suppression indicator suited to ranking methods rather than as an absolute score. To confirm that the ranking is not an artifact of using our own detector, we also computed SAS with the independently trained DAD[†] detector, obtaining similar results (Supplementary Section B.2).

4.2 Results for Artifact Detection

Table 1 reports detection accuracy and F1 scores on synthetic artifacts. DeLUX-D achieves the highest performance in nearly all categories, surpassing the other multimodal models, i.e., DAD[†] and the full DeLUX variant. Its advantage is most pronounced for glare, overexposure, and complex flares. Disabling the non-mask loss term $\mathcal{L}_{nm}$ in DeLUX-D improves artifact map details because the artifact detection is ‘disconnected’ from image reconstruction. However, as we will be shown in Section 4.4, the stronger and slightly larger estimated artifact maps of the full DeLUX variant offer better reconstruction quality. The flare-specific models (Flare7K++ and Wu et al.) fail to generalize beyond their training domains. A comparison of example artifact maps predicted by the analyzed models is presented in Fig. 5 and in the Supplementary, Section C.

4.3 Results for Artifact Removal

Table 2 summarizes the results of artifact removal. On synthetic data with available ground truth, DeLUX achieves the highest MS-SSIM, PSNR, and MAPE across all artifact types, confirming the feasibility and strong restoration fidelity of the proposed cross-modal inpainting approach.

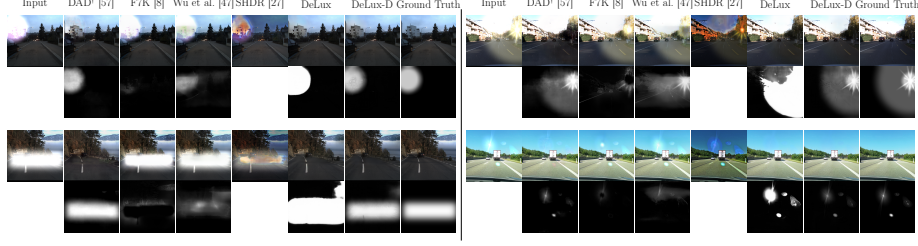


Fig. 5: Qualitative comparison of artifact removal (top row) and artifact detection (bottom row) on different kinds of synthetic artifacts.

Table 2: Results of artifact removal evaluation. Best values are highlighted in **bold**, second-best are underlined. Rows labeled ‘Overall’ denote the score calculated using combined data from all categories. The Δ symbol signifies the relative metric change compared to the input image. HDRRev-Diff was not evaluated on reconstruction metrics as the examples from Cityscapes used synthetic grayscale reconstructions instead of event streams.

	Metric	Category	DAD [†] [57]	F7K [8]	Wu et al. [47]	SHDR [27]	HDRRev-Diff [52]	DeLux (Ours)
Ground Truth Reconstruction	MS-SSIM $\uparrow$	Glare	<u>0.969</u>	0.938	0.938	0.859	-	0.990
		High-Quality Flares	<u>0.968</u>	0.957	0.950	0.867	-	0.990
		Overexposures	<u>0.959</u>	0.891	0.872	0.813	-	0.981
		Simple Flares	<u>0.984</u>	0.979	0.978	0.894	-	0.997
		Sun Flares+Glare	<u>0.967</u>	0.949	0.947	0.868	-	0.990
		Overall	<u>0.970</u>	0.954	0.949	0.868	-	0.991
	PSNR $\uparrow$	Glare	<u>28.279</u>	20.753	21.195	13.522	-	35.287
		High-Quality Flares	<u>29.567</u>	27.888	26.767	14.688	-	35.858
		Overexposure	<u>29.208</u>	21.002	19.877	12.864	-	34.501
		Simple Flares	32.128	<u>32.231</u>	31.032	14.796	-	41.402
		Sun Flares+Glare	<u>27.384</u>	21.840	22.178	14.249	-	34.122
		Overall	<u>29.525</u>	26.583	25.848	14.418	-	36.461
	MAPE $\downarrow$	Glare	<u>0.059</u>	0.097	0.099	0.576	-	0.025
		High-Quality Flares	<u>0.063</u>	0.064	0.072	0.553	-	0.032
		Overexposure	<u>0.061</u>	0.135	0.153	0.531	-	0.031
		Simple Flares	0.039	<u>0.025</u>	0.031	0.442	-	0.015
		Sun Flares+Glare	<u>0.076</u>	0.107	0.111	0.576	-	0.034
		Overall	<u>0.060</u>	0.071	0.078	0.537	-	0.028
Artifact Removal	Δ SAS $\uparrow$	Sun 02	87.64%	78.78%	81.22%	55.46%	84.78%	87.36%
		Sun 05	87.50%	85.06%	85.61%	-90.44%	87.79%	87.90%
		Sun 06	-130.65%	33.62%	26.24%	-847.04%	34.75%	-5.64%
		Sun 09	88.22%	83.81%	87.62%	61.36%	89.18%	88.32%
		Sun 10	87.74%	81.62%	83.40%	51.76%	26.22%	87.79%
		Sun 11	-173.61%	43.53%	40.43%	-761.38%	-5.76%	51.57%
		Zurich City 01-e	-1953.29%	45.60%	-123.38%	-216.95%	-7986.77%	66.17%
		Zurich City 12-a	30.41%	26.00%	-109.16%	51.18%	-15.34%	63.83%
		Overall	-352.50%	58.87%	18.54%	-255.65%	-1402.87%	63.22%

In real-world automotive scenarios, DELUX achieves the highest average artifact suppression (SAS), outperforming the single-network DAD[†] approach in removing both mild and severe distortions. Moreover, the event-guided HDR baseline, HDRRev-Diff, performed mostly global tone mapping, which, for certain

scenarios, was not enough to fully suppress localized artifacts (Fig. 6 rows 2–4), resulting in negative SAS scores. This empirical result confirms our hypothesis: while global tone-mapping models successfully recover dynamic range, they often preserve local optical artifacts. In contrast, the explicit detection and fusion stages of DELUX successfully isolated and removed these degradations by using the event stream to reconstruct the underlying image structure.

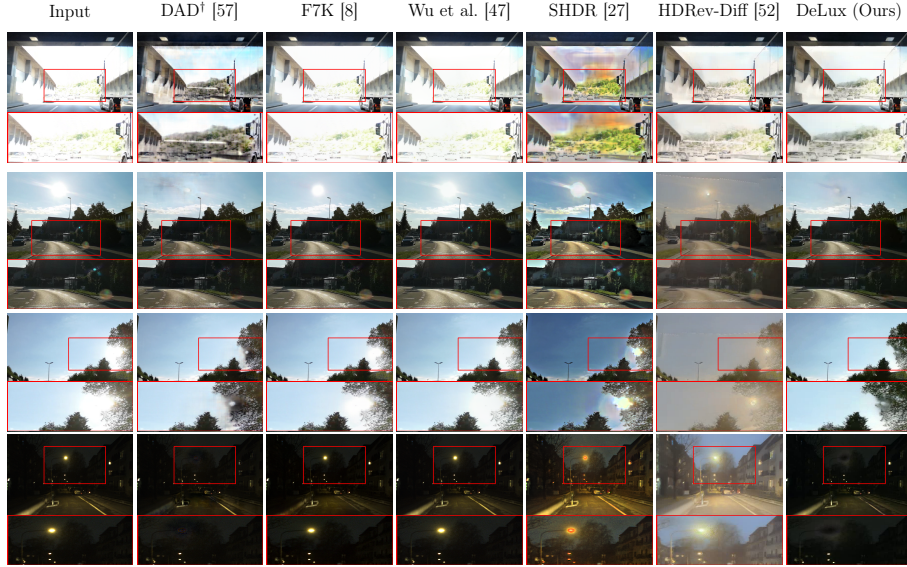


Fig. 6: Comparison of artifact removal on real-world recordings. Zoom-ins (red border) provided for better visual assessment.

Qualitative visual assessments of synthetic and real-world artifacts (Fig. 5 and Fig. 6, and in Sections C and D of the Supplementary) further demonstrate that DELUX robustly handles dynamic light variations and does the best job of restoring lost image details. Overall, DELUX offers the most consistent balance between artifact removal and image quality.

In terms of efficiency, DELUX runs at 16.5 FPS with 90.2M parameters on an NVIDIA A100—an order of magnitude lighter and faster than the diffusion-based HDRev-Diff baseline (1.6B parameters, 1.1 FPS); full model sizes and timings are reported in the Supplementary, Section F.

4.4 Ablation Studies

To evaluate the impact of each architectural component of the default DELUX configuration, we conducted an ablation study with four variants: DELUX-NE (no events), DELUX-ND (no detector), and DELUX-D (disabled non-mask loss). The results are summarized in Table 3.

Table 3: Ablation study results on DeLux and its variants. Best values are **bolded**, and second-best values are underlined.

	Accuracy $\uparrow$	MS-SSIM $\uparrow$	PSNR $\uparrow$	MAPE $\downarrow$	Δ SAS [%] $\uparrow$
DeLux-NE	<u>0.964</u>	<u>0.986</u>	33.757	0.039	67.06%
DeLux-ND	-	0.982	30.822	0.059	-195.34%
DeLux-D	0.971	0.991	<u>35.839</u>	<u>0.029</u>	-1703.38%
DeLux	0.925	0.991	36.461	0.028	<u>63.22%</u>

Disabling the non-mask loss (DeLux-D) yields the highest artifact detection accuracy (0.971), but at the cost of artifact removal (Δ SAS of -1703.38%). Similarly, removing the detector (DeLux-ND) causes the model to fail to reliably suppress artifacts (Δ SAS of -195.34%), underscoring the need for explicit artifact localization. On the other hand, removing the event modality (DeLux-NE) yields the strongest apparent artifact suppression (Δ SAS of 67.06%), but it also results in the loss of structural cues, leading to inaccurate reconstructions (PSNR=33.757).

For general-purpose scenarios requiring a balance between artifact detection and removal, we recommend the default DeLux variant. For strictly artifact detection tasks, we recommend the DeLux-D variant, which provides a detailed map of lighting artifacts. This contrast reflects a deliberate detection-removal trade-off. The non-mask loss $\mathcal{L}_{nm}$ couples the detector and inpainter, forcing edits to remain inside the predicted mask. This markedly improves removal but pushes the detector toward larger, smoother masks at the cost of pixel-sharp localization. Disabling non-mask loss (DeLux-D) decouples the two, yielding the sharpest masks but an unconstrained inpainter that over-modifies clean regions. The two configurations therefore target different uses—DeLux-D for explicit artifact maps (analysis, manual editing) and DeLux for end-to-end restoration.

5 Discussion

In this work, we introduced DeLux, a proof-of-concept framework that puts forward a novel research direction: cross-modal local artifact restoration. By focusing on targeted, modality-guided inpainting, we demonstrated that secondary sensing modalities, such as neuromorphic event cameras, can successfully guide the restoration of degraded RGB pixels. Our core insight is that explicitly decoupling artifact localization from multimodal fusion prevents the unwarranted alteration of clean image regions, allowing the network to selectively use the secondary modality only where the RGB sensor has failed.

However, it is also important to acknowledge certain limitations of the current study. While event cameras offer high dynamic range, they are inherently limited in stationary scenes where no events are triggered. Nevertheless, the modular architecture of DeLux is designed to be modality-agnostic. A possible solution to this static-scene limitation could be to combine multiple sensing

modalities, for instance, by simultaneously fusing RGB, neuromorphic events, and thermal imaging. A second limitation of our current study is the use of synthetic artifact generation for evaluation against a ground-truth image. While we use high-quality flare and glare overlays (such as those from the Flare7K++ dataset), synthetic overlays cannot fully reproduce the complex physical interplay of light within a real lens assembly. The real-world evaluations, on the other hand, lacked an artifact-free equivalent. Therefore, to advance this field, future research must focus on capturing paired, real-world datasets of lighting artifacts. Furthermore, relying on a frozen E2VID reconstructor currently limits full end-to-end optimization, suggesting that future pipelines could directly ingest raw event representations to enable more efficient, task-specific feature learning. Finally, because DELUX restores each frame independently, its outputs are not guaranteed to be temporally consistent, and residual artifacts can vary across adjacent frames, resulting in visible flicker in the reconstructed video. Enforcing temporal consistency, e.g., through multi-frame modeling, is an important direction for future work.

Ultimately, DELUX serves as a template for addressing severe localized image degradations using a secondary sensing modality. By demonstrating the usefulness of cross-modal local artifact restoration, we hope to inspire research that moves beyond single-sensor constraints and paves the way for highly reliable image processing systems capable of operating under the most adverse lighting conditions.

Data and Code Availability. The curated real-world evaluation data, synthetic artifact generation tools, and training code are publicly available at: <https://github.com/Tremirre/event-sun-effects-remover>.

References

1. Visual odometry and mapping under poor visibility conditions using a stereo infrared thermal imaging system. *Electronic Imaging* **35**(5), 325–1–325–1 (2023)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019)
3. Blahnik, V., Voelker, B.: About the reduction of reflections for camera lenses: How t*-coating made glass invisible. Tech. rep., Carl Zeiss AG (March 2016), <https://lenspire.zeiss.com/photo/app/uploads/2022/02/technical-article-about-the-reduction-of-reflections-for-camera-lenses.pdf>
4. Ceccarelli, A., Secci, F.: Rgb cameras failures and their effects in autonomous driving applications. *IEEE Transactions on Dependable and Secure Computing* **20**(4), 2731–2745 (2023)
5. Chiang, C.C., Yang, Y.Y., Liu, W.L., Lin, Y.C.: Improving automatic identification of medications in transparent packaging by glare removal and color correction. *IEEE Access* **11**, 118812–118829 (2023)

6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Dai, Y., Li, C., Zhou, S., Feng, R., Loy, C.C.: Flare7k: A phenomenological nighttime flare removal dataset (2022)
8. Dai, Y., Li, C., Zhou, S., Feng, R., Luo, Y., Loy, C.C.: Flare7k++: Mixing synthetic and real datasets for nighttime flare removal and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(11), 7041–7055 (2024)
9. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning. pp. 1–16 (2017)
10. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics* **36**(6), 1–15 (Nov 2017)
11. Elvidge, C.D., Zhizhin, M., Pipkin, A., Anderson, S., Kowalik, W.S., Bazilian, M.: Exploring the hidden world of lighting flicker with a high-speed camera. *Atmosphere* **15**(4) (2024)
12. Gehrig, D., Loquercio, A., Derpanis, K., Scaramuzza, D.: End-to-End Learning of Representations for Asynchronous Event-Based Data . In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5632–5642. IEEE Computer Society, Los Alamitos, CA, USA (Nov 2019)
13. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* **PP**, 1–1 (03 2021)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015)
15. Huber, P.J.: Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* **35**(1), 73 – 101 (1964)
16. Jiang, Y., Chen, X., Pun, C.M., Wang, S., Feng, W.: Mfdnet: Multi-frequency deflare network for efficient nighttime flare removal. *ArXiv* **abs/2406.18079** (2024)
17. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3317–3326 (2020)
18. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. *ArXiv* **abs/1603.08155** (2016)
19. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* **36**(4) (Jul 2017)
20. Kanj, A., Talbot, H., Luparello, R.: Flicker removal and superpixel-based motion tracking for high speed videos. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 245–249 (09 2017)
21. Lamprinou, N., Psarakis, E.: Fast detection and removal of glare in gray scale laparoscopic images. In: International Conference on Computer Vision Theory and Applications. pp. 206–212 (01 2018)
22. Lei, C., Ren, X., Zhang, Z., Chen, Q.: Blind video deflickering by neural filtering with a flawed atlas. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
23. Li, Y., Moreau, J., Ibanez-Guzman, J.: Emergent visual sensors for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems* **24**(5), 4716–4737 (2023)

24. Li *et al.*, X.: Generalizing event-based hdr imaging to various exposures. *Neuro-computing* **600**, 128132 (2024)
25. Liang, Z., Xu, J., Zhang, D., Cao, Z., Zhang, L.: A hybrid l1-l0 layer decomposition model for tone mapping. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4758–4766 (2018)
26. Liu, X., Li, J., Shi, J., Fan, X., Tian, Y., Zhao, D.: Event-based monocular depth estimation with recurrent transformers. *IEEE Transactions on Circuits and Systems for Video Technology* **PP**, 1–1 (08 2024)
27. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
28. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
29. Lv, X., Zhang, S., Liu, Q., Xie, H., Zhong, B., Zhou, H.: Backlitnet: A dataset and network for backlit image enhancement. *Computer Vision and Image Understanding* **218**, 103403 (2022)
30. Marnerides, D., Bashford-Rogers, T., Hatchett, J., Debattista, K.: Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *Computer Graphics Forum* **37**(2), 37–49 (2018)
31. Matta, G.R., Siddhartha, R., Girish, R.S.V., Sharma, S., Mitra, K.: Gn-fr:generalizable neural radiance fields for flare removal. *ArXiv abs/2412.08200* (2024)
32. Messikommer, N., Georgoulis, S., Gehrig, D., Tulyakov, S., Erbach, J., Boichichio, A., Li, Y., Scaramuzza, D.: Multi-bracket high dynamic range imaging with event cameras. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), New Orleans, 2022 (2022)
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* **44**(10), 6890–6909 (2022)
34. Niu, C., Li, K., Wang, D., Zhu, W., Xu, H., Dong, J.: Gr-gan: A unified adversarial framework for single image glare removal and denoising. *Pattern Recognition* **157**, 110816 (2025)
35. Qiao, J., Wang, X., Chen, J., Jian, M.: Low-light image enhancement with an anti-attention block-based generative adversarial network. *Electronics* **11**(10) (2022)
36. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3852–3861 (2019)
37. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(6), 1964–1980 (2021)
38. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* **60**(1), 259–268 (1992)
39. Shihavuddin, A.S.M., Gracias, N., Garcia, R.: Online sunflicker removal using dynamic texture prediction. In: Proceedings of the International Conference on Computer Vision Theory and Applications - Volume 1: VISAPP, (VISIGRAPP 2012). pp. 161–167. INSTICC, SciTePress (2012)
40. Shoshin, A., Shvets, E.: Veiling glare removal: synthetic dataset generation, metrics and neural network architecture. *Computer Optics* **45**, 615–626 (07 2021)

41. Song, C., Huang, Q., Bajaj, C.: E-cir: Event-enhanced continuous intensity recovery. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7793–7802 (2022)
42. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Klee-man, L., Mahony, R.: Reducing the sim-to-real gap for event cameras (2020)
43. Wang, L., Mostafavi, I.M., Ho, Y.S., Yoon, K.J.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 10073–10082. IEEE (Jun 2019)
44. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003)
45. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17683–17693 (June 2022)
46. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part II. p. 120–135. Springer-Verlag, Berlin, Heidelberg (2018)
47. Wu, Y., He, Q., Xue, T., Garg, R., Chen, J., Veeraraghavan, A., Barron, J.T.: How to train neural networks for flare removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2239–2247 (2021)
48. Xi, Y., Zhang, Z., Wang, W.: Low-light image enhancement method for electric power operation sites considering strong light suppression. *Applied Sciences* **13**(17) (2023)
49. Yang, X., Xu, K., Song, Y., Zhang, Q., Wei, X., Lau, R.W.: Image correction via deep reciprocating hdr transformation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1798–1807 (2018)
50. Yang, Z., Yang, X., Wu, L., Hu, J., Zou, B., Zhang, Y., Zhang, J.: Pre-inpainting convolutional skip triple attention segmentation network for agv lane detection in overexposure environment. *Applied Sciences* **12**(20) (2022)
51. Yang *et al.*, Y.: Learning event guided high dynamic range video reconstruction. In: CVPR. pp. 13924–13934 (2023)
52. Yang *et al.*, Y.: Event-guided hdr reconstruction with diffusion priors. In: ICCV. pp. 11787–11796 (2025)
53. Zhang, J., Lalonde, J.F.: Learning high dynamic range from outdoor panoramas. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4529–4538 (2017)
54. Zheng, Q., Zhao, Y., Zhang, X., Zhu, P., Ma, W.: A multi-view image fusion algorithm for industrial weld. *IET Image Processing* **17**(1), 193–203 (2023)
55. Zheng, X., Liu, Y., Lu, Y., Hua, T., Pan, T., Zhang, W., Tao, D., Wang, L.: Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv.2302.08890* (2023)
56. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Live demonstration: Unsupervised event-based learning of optical flow, depth and egomotion. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1694–1694 (2019)
57. Zou, Z., Lei, S., Shi, T., Shi, Z., Ye, J.: Deep adversarial decomposition: A unified framework for separating superimposed images. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12803–12813 (2020)