This paper has been accepted for publication at the IEEE Conference on
Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, 2023.
©IEEE

EVREAL: Towards a Comprehensive Benchmark and Analysis Suite for Event-based Video Reconstruction

Burak Ercan [1,2] Onur Eker [1,2] Aykut Erdem [3,4] Erkut Erdem [1,4]

[1] Hacettepe University, Computer Engineering Department [2] HAVELSAN Inc.
[3] Koç University, Computer Engineering Department [4] Koç University, KUIS AI Center

Abstract

Event cameras are a new type of vision sensor that incorporates asynchronous and independent pixels, offering advantages over traditional frame-based cameras such as high dynamic range and minimal motion blur. However, their output is not easily understandable by humans, making the reconstruction of intensity images from event streams a fundamental task in event-based vision. While recent deep learning-based methods have shown promise in video reconstruction from events, this problem is not completely solved yet. To facilitate comparison between different approaches, standardized evaluation protocols and diverse test datasets are essential. This paper proposes a unified evaluation methodology and introduces an open-source framework called EVREAL to comprehensively benchmark and analyze various event-based video reconstruction methods from the literature. Using EVREAL, we give a detailed analysis of the state-of-the-art methods for event-based video reconstruction, and provide valuable insights into the performance of these methods under varying settings, challenging scenarios, and downstream tasks.

More results, analyses, and source code available at: https://ercanburak.github.io/evreal.html

1. Introduction

Event cameras are a new type of biologically-inspired vision sensor that have the potential to overcome the limitations of conventional frame-based cameras. Unlike traditional cameras, event cameras have pixels that work independently and asynchronously from one another. Each pixel detects changes in the relative brightness of its local area and generates an event signal when the change exceeds a certain threshold. Hence, the output is a continuous stream of events, with each event containing information about the location, the polarity of the intensity change, and the precise time of the detected change. The rate of event generation varies depending on the scene characteristics, with more events being triggered for scenes showing rapid motion or instant changes in brightness and texture. Due to their unique working principles, event cameras provide many advantages, such as high dynamic range, high temporal resolution, and minimal motion blur [6].

While event streams have many desirable properties, they have one major disadvantage – humans cannot easily understand event streams in the same way as intensity images. Hence, a fundamental task in event-based vision literature is reconstructing intensity images from event streams. Reconstructing high-quality videos from events also allows for employing existing frame-based computer vision methods developed for several downstream tasks to event data in a straightforward manner [20].

While deep learning-based methods have made impressive progress in reconstructing videos from event streams lately (e.g., [21, 24, 30]), this research problem is still not completely solved. This can be primarily attributed to the event representations being used in these state-of-the-art approaches, which can cause some latency issues. Moreover, training these methods relies heavily on synthetically created datasets. Consequently, these methods may produce suboptimal video reconstructions that suffer from issues such as blur, low contrast, or smearing artifacts.

A significant effort has been put forth to find better ways to evaluate event-based video reconstruction methods and assess the visual qualities of reconstructed videos. There are several distinct evaluation methodologies involving different datasets, event representations, post-processing steps, quantitative metrics, and downstream tasks (see Tab. 1 for an overview, and refer to the supplementary materials for a more detailed discussion). However, the lack of a standard evaluation procedure makes it hard to fairly compare the performances of different methods. The details of the evaluation procedures are sometimes not clearly defined, even though each minor detail may significantly alter the results. This also poses challenges for other researchers to reproduce the reported results. This motivates the need for open-source codes and standardized protocols for evaluation.

Comparing different methods requires not only well-defined evaluation protocols but also a diverse set of test datasets that cover various real-world settings. Large-scale benchmarks have been instrumental in advancing many computer vision tasks, as demonstrated by ImageNet [22] for image classification and MS-COCO [10] for object detection, providing results that are generalizable to unseen real-world data. However, since event-based vision is a relatively new field compared to classical frame-based computer vision, the current datasets used for assessing event-based video reconstruction are limited in scale and scope, confined to specific domains, scenes, camera types, and motion patterns. To ensure the generalizability of the results and evaluate the methods' effectiveness in more real-world scenarios, it is essential to assess their performances on a large variety of datasets showing different characteristics.

Event data is handy in scenarios where traditional frame cameras fail, such as scenes captured under low-light conditions or with rapid motion, and underexposed or overexposed regions. Hence, it is of utmost importance to evaluate the effectiveness of event-based video reconstruction models in those challenging situations. However, as traditional frame-based cameras especially struggle in these scenarios, collecting high-quality reference frames is a challenging task on its own. This paradox makes it difficult to quantify the success of event-based video reconstruction methods where they are most needed.

Even in scenarios where it is possible to collect high-quality reference frames with minimal motion blur and optimal exposure, assessing image quality remains a subjective endeavor. Hence, the current studies generally consider a perceptual metric like learned perceptual image patch similarity (LPIPS) [32] along with distortion-aware metrics like PSNR and structural similarity (SSIM) [29]. However, as a full-reference metric, LPIPS is trained on distortions that are not commonly seen in the reconstructed intensity images from event data. Hence, this raises some doubts about the significance of these perceptual comparisons.

Reconstructing images from events is a complex task. It depends on many variables that can affect the performance of the methods. These include sensor noise characteristics, sensor parameters, event generation rate, event grouping scheme, grouping rate, frame reconstruction rate, and temporal regularity. Despite their importance, the literature often overlooks the robustness of the methods to changes in these variables. Therefore, it is crucial to evaluate the sensitivity of the methods to these variables and to consider their performance under changing conditions. A method that performs well under specific settings may not be suitable for general use when these variables are expected to change.

Event cameras are known for their low-latency and non-redundant data flow, making them ideal for scenarios that require real-time and low-power processing. As a result, the computational efficiency of event-based video reconstruction methods is just as important as the visual quality of reconstructions. Neglecting this aspect in a benchmark could lead to choosing a method that provides high-quality reconstructions, but is impractical for real-time processing.

To address these issues and facilitate progress in event-based video reconstruction, in this study, we propose EVREAL, Event-based Video Reconstruction Evaluation and Analysis Library, an open-source framework based on PyTorch [19]. Our framework offers a unified evaluation pipeline to benchmark pre-trained neural networks and a result analysis tool to visualize and compare reconstructions and their scores. We use a large set of real-world test sequences and various full-, and no-reference image quality metrics to perform qualitative and quantitative analysis under diverse conditions, including challenging scenarios such as rapid motion, low light, and high dynamic range, many of which have not been reported before. Moreover, we conduct experiments to assess the performance of each method under variable conditions and analyze their robustness to these varying settings. We also evaluate the quality of video reconstructions via downstream tasks like camera calibration, image classification, and object detection. This extrinsic evaluation can be considered a proxy metric for image quality and a task-specific metric if the goal of event-based video reconstruction is to perform these downstream tasks.

In Table 1, we present an overview of our experimental setup in comparison to prior work. Along with this paper, we build a website to share our results and findings, together with the source code to reproduce them. We also intend to update this webpage on a regular basis as new event-based video reconstruction methods are proposed. Our contributions in this paper can be summarized as follows:

- We propose a unified evaluation methodology and an open-source framework to benchmark and analyze event-based video reconstruction methods from the literature.
- Our benchmark includes additional datasets, metrics, and analysis settings that have not been reported before. We present quantitative results on challenging scenarios involving rapid motion, low light, and high dynamic range.
- Moreover, we conduct additional experiments to analyze the robustness of methods under varying settings such as event rate, event tensor sparsity, reconstruction rate, and temporal irregularity.
- To further examine the quality of the reconstructions, we provide quantitative analysis on several downstream tasks, including camera calibration, image classification, and object detection.

2. Methodology of Evaluation and Analysis

2.1. Task Description

Suppose we have a stream of events $\{e_i\}$ containing N_E events and spanning T sec. Each event $e_i = (x_i, y_i, t_i, p_i)$ in the stream represents a change in brightness perceived

Evaluation Setup in	Test Datasets	# of Frames	Compared Methods	Metrics (FR/NR)	Comp. Eff.	Chlng. Scnrs.	Downst. Tasks	Robust. Exp.	Open Source
[27]	[1]	0.2K	[1, 15, 27]	–/B		🖼			
[21]	[13]	1.9K	[1, 15, 21]	MSLT/–	✓	🖼	🖼 🖼 📊 📊	🖼	
[23]	[13]	1.9K	[1, 15, 21, 23]	MSL/–	✓	🖼			
[24]	[13, 24, 34]	28.7K	[21, 23, 24]	MSL/–					
[2]	[13]	3.1K	[2, 21, 23]	MSLT/R	✓		📊		✓
[30]	[13, 24, 34]	28.7K	[21, 23, 24, 30]	MSL/–		🖼			✓
[18]	[13, 24]	17.4K	[18, 21, 23, 24]	MSL/–		🖼			
[35]	[13, 24, 34]	28.7K	[2, 21, 23, 24, 35]	MSL/–	✓			📊	
[33]	[13, 16]	1.9K	[18, 21, 24, 33]	MSL/–	✓	🖼		📊	
Ours	[13, 21, 24, 25, 34]	47.7K	[2, 18, 21, 23, 24, 30]	MSL/BNM	✓	📊	📊 📊 📊	📊	✓

Table 1. **A summary of experimental setups considered in earlier work.** We provide a comparison of our proposed EVREAL framework to the experimental evaluation setups reported in the existing work in terms of datasets being used, methods compared, number of reconstructed frames used in quantitative analysis, and metrics being utilized. We also indicate whether each evaluation setup includes analysis of computational efficiency, challenging scenarios (fast motion, low light, or high-dynamic range), downstream tasks, and robustness. Finally, we mark whether the implementation of this evaluation setup is open-sourced or not. In the metrics column, FR and NR stand for full-reference and no-reference metrics, respectively. M:MSE, S:SSIM [29], L:LPIPS [32], and T:Temporal Consistency [9] are the full-reference metrics, while R:RMS contrast, B:BRISQUE [11], N:NIQE [12], and M:MANIQA [31] are the no-reference metrics. In the Challenging Scenarios, the Downstream Tasks and Robustness Experiments columns, each 🖼 symbol denotes a reported qualitative analysis and a 📊 symbol represents a quantitative analysis being performed along with a qualitative comparison.

by the sensor, and contains information about the location (x_i, y_i), the timestamp t_i, and the polarity p_i of this intensity change. Here, $t_i \in [0, T]$, $p_i \in \{+1, -1\}$, $x_i \in \{0, \ldots, W-1\}$, and $y_i \in \{0, \ldots, H-1\}$ for all $i \in \{0, \ldots, N_E - 1\}$, with W and H denoting the width and height of the sensor array, respectively. Given these events, the task aims to generate a stream of N_I images $\{\hat{I}_k\}$, corresponding to the same T sec. period as the events. Each image $\hat{I}_k$ represents the absolute brightness of the scene, as if it were captured by a standard frame-based camera at a particular time s_k within the time period of T seconds, where k ranges from 1 to N_I, and $\hat{I}_k$ ranges from 0 to 1.

2.2. Evaluation Framework and Pipeline

EVREAL implements several standardized components crucial for deep event-based video reconstruction models, including *event pre-processing*, *event grouping*, *event representation*, *representation processing*, and *image post-processing* (see Fig. 1). We have included components to evaluate the visual quality of each frame in the generated videos, which are split into *full-reference* metrics and *no-reference* metrics. The former is utilized when high-quality, distortion-free ground truth frames are available, whereas the latter are used when ground truth frames are of low quality or not available at all (refer to Sec. 2.4). EVREAL also includes an analysis tool. Given a set of reconstructions generated by one or more methods, it collects ground truth frames, event visualizations, event rate statistics, and instantaneous values for a set of quantitative metrics. It then generates an output video that displays this data in a time-synchronized manner, including plots of quantitative metrics (see the supplementary for a sample output video). Our tool is particularly valuable in pinpointing specific limitations and failure cases of methods. For instance, it can reveal situations where noisy reconstructions significantly impact future reconstructions due to the sequential nature of the method. Such scenarios can be visually identified from the plots of quantitative metrics. To assess the practical use of a given method, our framework allows for evaluating it on several downstream tasks. Specifically, we analyze the performance of tested models on three downstream tasks, *object detection*, *image classification*, and *camera calibration* (Sec. 2.7). We want to emphasize that our objective with this work is to conduct a comprehensive evaluation and analysis of existing pre-trained models from the literature in order to characterize them rather than to provide a ranking of them. Developing a new model is also beyond the scope of this work. In the following, we provide detailed descriptions of the components of our evaluation framework.

Event pre-processing. This component can be employed to process raw events before grouping them. Possible pre-processing operations include event temporal downsampling and adding artificial event noise, to perform robustness experiments under these conditions.

Event grouping. Each event in isolation contains limited information about the scene, so a common practice is to group a number of events together and process them as a whole. We consider the following grouping options:

- **Fixed-number:** We group every N_G number of events such that the kth event group can be defined as:

$$G_k \doteq \{e_i \mid kN_G \le i < (k+1)N_G\} \tag{1}$$

  Here, the rate at which the groups are formed varies according to the incoming event rate.

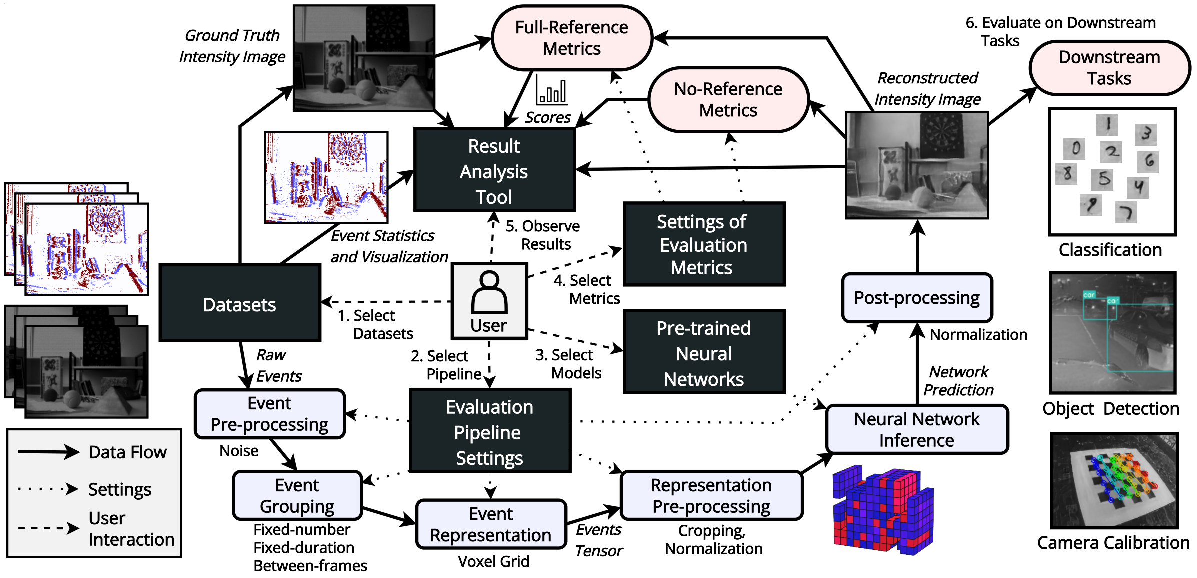


Figure 1. **An overall look at our proposed EVREAL (Event-based Video Reconstruction – Evaluation and Analysis Library) toolkit.**

- **Fixed-duration:** We group events according to non-overlapping time windows with a fixed duration of T_G secs. The kth event group contains all events with timestamps t_i falling within the kth time window, defined as:

$$G_k \doteq \{e_i \mid kT_G \leq t_i < (k+1)T_G\} \qquad (2)$$

In this scheme, the number of events in each group varies according to the incoming event rate.

- **Between-frames:** If we also have ground truth intensity frames together with the incoming event stream, we can group events such that every event between consecutive frames belongs to the same group. Therefore, the set of events in the kth event group can be defined as follows:

$$G_k \doteq \{e_i \mid s_k \leq t_i < s_{k+1}\} \qquad (3)$$

If the ground truth frames arrive at a fixed rate, then this option is a special case of the fixed temporal window grouping. Note that, however, the time difference between consecutive frames may not be fixed all the time, *e.g.*, due to changing exposure times of the camera in real-world datasets, or due to adaptive rendering schemes of simulators in synthetic datasets.

Event representation. To utilize deep CNN architectures for event-based data, a common choice is to accumulate grouped events into a grid-structured representation such as a voxel grid [36]. We also follow this approach in our evaluation procedure. The details of this event representation are given in the supplementary material.

Representation pre-processing. After forming a representation from grouped events, it is possible to pre-process this representation before feeding it to the neural network, such as cropping or applying normalization. While our framework supports this, we did not include such a pre-processing in the experimental analysis reported in this paper.

Neural network inference. This module is used for predicting intensity frames given the event representation by employing the pre-trained neural network model chosen by the user. As mentioned earlier, we use PyTorch here.

Post-processing. It is also possible to post-process the intensity frame that the network predicts, by utilizing procedures like robust min/max normalization, as done in [21]. While EVREAL supports this, we did not employ any post-processing operations in our experiments in this paper.

2.3. Tested Approaches

We compare seven methods from the literature that have PyTorch based open-source model codes and pre-trained models. These methods include E2VID [21], FireNet [23], FireNet+ and E2VID+ [24], SPADE-E2VID [2], SSL-E2VID [18], and ET-Net [30]. Note that E2VID+ and SSL-E2VID share the same deep network architecture as E2VID, while FireNet+ employs the same architecture as FireNet. Here we utilize the pre-trained models shared publicly by the authors and evaluate them on the same datasets under a common experimental evaluation setup.

2.4. Quantitative Image Quality Metrics

To quantitatively assess the quality of videos reconstructed from events, we use both full-reference and no-reference metrics. Full-reference metrics, as their name implies, provide a quality score for an image in regard to a given reference image. In contrast, no-reference metrics do not require any ground truth image and give perceptual quality scores by directly processing input images.

We utilize three full-reference evaluation metrics: MSE, SSIM [29], and LPIPS [32]. These metrics are employed only when high-quality, distortion-free ground truth frames are available. While the MSE and SSIM are better suited for capturing distortions, LPIPS measures the perceptual similarity by a deep network trained to conform with human visual perception. Furthermore, we utilize three no-reference metrics: BRISQUE [11], NIQE [12], and MANIQA [31]. These metrics are used when the ground truth frames are of low quality or when they are not available at all. BRISQUE and NIQE are traditional metrics that employ hand-crafted features and measure conformity to natural scene statistics, considering various synthetic and authentic distortions such as blur, noise, and compression. On the other hand, MANIQA is a deep-learning based method employing vision transformer architecture [3], which is trained in an end-to-end manner to assess perceptual image quality while specifically focusing on distortions seen in the outputs of neural network based image restoration algorithms.

The results of the aforementioned metrics can be influenced by specific settings. Hence, to ensure consistency, we provide the detailed settings that we use in the supplementary material. Although performing histogram equalization before calculating image quality metrics (as in [18, 20, 23]) is supported by EVREAL, we do not perform this operation in our experiments reported in this paper.

2.5. Datasets

We adopt commonly used sequences from three datasets in our evaluation framework, namely the Event Camera Dataset (ECD) [13], the Multi Vehicle Stereo Event Camera (MVSEC) dataset [34], the High-Quality Frames (HQF) dataset. In addition, we use handheld sequences from the Beam Splitter Event and RGB (BS-ERGB) Dataset [25]. We use full-reference metrics to evaluate the performance of the models on these sequences. To assess the performance of the models in challenging scenarios, we also use no-reference metrics and include sequences with fast motion, low light, and high dynamic range scenes. Specifically, we use the latter parts of ECD sequences where camera movements increase to evaluate fast motion (denoted as ECD-FAST), night driving sequences from the MVSEC to evaluate low light (denoted as MVSEC-NIGHT), and HDR sequences from [21] to evaluate high dynamic range scenes. Please refer to the supplementary material for detailed information about these datasets.

2.6. Robustness Analysis

To analyze the factors that affect the performance of reconstructing images from events, several variables need to be considered. In this paper, we investigate the impact of four critical ones: *event rate*, *event tensor sparsity*, *image reconstruction rate*, and *temporal irregularity*. To evaluate the results, we utilize the LPIPS metric and employ commonly used sequences from the ECD, MVSEC, and HQF datasets, as mentioned earlier. In the following sections, we provide detailed descriptions of these experiments.

Event rate. To evaluate the robustness of the methods to varying event rates, we employ between-frames event grouping and collect statistics on event rates, measured in events per second, for each group. We then reconstruct intensity images using each method based on the event groups, and calculate LPIPS scores for each time step. We divide the event rate spectrum into ten equally spaced bins and compute the mean LPIPS scores for each bin and method. This enables us to assess the performance of each method under different event rate conditions and determine which methods are most robust to changes in event rate.

Tensor sparsity. To analyze how the sparsity of event tensors affect the performance of each method, we carry out experiments utilizing fixed-number grouping and a tolerance of 1 ms to match the reconstructions with ground truth frames. With this grouping approach, each group contains the same number of events, resulting in event tensors with the same sparsity level. Specifically, we conduct 9 different experiment runs, with event numbers ranging from 5K to 45K. We then compute the mean LPIPS scores for each experiment run and for each method.

Note that if there exists slow motion or little texture in the scene, using fixed-number grouping can result in event groups that span a large temporal window when the event rate is small. Furthermore, the motion or texture captured by the event camera might be contained in a small region of pixels rather than being homogeneously distributed to all of the sensor area. In that case, the temporal discretization performed in the event representation (to a fixed number of temporal bins) means more compression of the temporal information, and this might result in reconstruction artifacts such as saturation or blur in these regions. The tensor sparsity experiments aid us in assessing each method's robustness to these situations.

Reconstruction rate. To evaluate the impact of changing frame reconstruction rates on each method's performance, we conduct experiments using fixed-duration grouping, which generates a fixed number of frames per second. We perform ten experiment runs, each with a different event grouping duration ranging from 10 ms to 100 ms, which correspond to frame reconstruction rates from 10 FPS to 100 FPS. We use a tolerance of 1 ms to match the reconstructions with ground truth frames. We then compute the average LPIPS values for each experiment run and method to determine their performance under different frame rates.

Temporal irregularity. To evaluate the robustness of each method in generating frames at irregular intervals, we conduct experiments by removing a certain percentage of ground truth frames from each sequence and by using between-frames event grouping to group events between the remaining frames. In particular, we conduct ten experiment runs with different discarding ratios ranging from 0.0 (standard case) to 0.9. We then calculate the mean LPIPS scores obtained for each experiment run for each method.

2.7. Analysis on Downstream Tasks

Event cameras, due to their unique characteristics, can provide a viable alternative to traditional frame-based cameras in challenging conditions. As a result, using images reconstructed from event streams for downstream tasks when standard cameras fail can be beneficial. To assess the effectiveness of each method in an extrinsic manner, we leverage downstream computer vision tasks including object detection, image classification, and camera calibration.

Object detection. Object detection is a crucial area of research in computer vision, with numerous applications ranging from autonomous navigation to medical imaging. However, traditional frame-based cameras often fail to capture satisfactory images under low-light conditions, affecting the performance of object detection methods. Since event cameras have a high dynamic range compared to traditional cameras, we evaluated the performance of object detection on low-light images captured by event cameras. For this purpose, we used the MVSEC-NIGHTL21 car detection dataset [8], derived from the *outdoor_night1_data* sequence of MVSEC, captured under night driving conditions. The dataset contains 2,000 labeled intensity images, with 1,600 frames for training and 400 frames for validation. We reconstructed images from the provided event sequence for each method and extracted frames corresponding to those in the MVSEC-NIGHTL21 dataset. We then used YOLOv7 [26] object detector to detect cars in the reconstructed images and intensity images of the frame camera. We used a model trained on the COCO dataset [10] for car detection in the images and evaluated the results using the PASCAL VOC metric [4], providing the AP score for each method on the dataset. See the supplementary material for sample detection visualizations.

Image classification. We evaluated the performance of our image reconstruction methods using two image classification datasets: Neuromorphic-Caltech101 (N-Caltech101) [16] and Caltech101 [5]. N-Caltech101 is a spiking version of the original Caltech101 dataset, containing 100 object classes plus a background class (excluding the "Faces" class). We trained a ResNet50 [7] classification model on Caltech101, excluding the "Faces" class to ensure consistency between the datasets. For each method, we reconstructed images on event streams from N-Caltech101, and ran the ResNet50 model on the reconstructions to evaluate their accuracy on the dataset.

Camera calibration. It is a critical component of computer vision systems, but traditional calibration techniques for standard frame-based cameras cannot be applied to event cameras due to their asynchronous pixel output. Recently, Muglikar *et al.* [14] demonstrated that image reconstruction can be used to apply conventional calibration techniques for accurate event-camera calibration. In this study, we compare the performance of various image reconstruction methods for camera calibration using the *calibration* sequence from the ECD dataset. This sequence consists of an event camera moving in front of a calibration target, and the intrinsic calibration parameters of the DAVIS240C, provided by ECD, serve as the ground truth. We reconstruct image sequences using each method and accordingly obtain intrinsic calibration parameters using the reconstructed images and the `kalibr` toolbox [17]. We then measure the mean absolute percentage error (MAPE) of the intrinsic calibration parameters to determine the most effective method.

3. Evaluation Results and Discussion

Table 2 presents the quantitative results of image reconstruction methods on four datasets (ECD, MVSEC, HQF, and BS-ERGB) and using three evaluation metrics (MSE, SSIM, and LPIPS), while Fig. 2 displays qualitative results from sample scenes. The table highlights that the methods ET-Net and E2VID+ are the top performers across all datasets, with ET-Net being overall the most accurate. E2VID+ performs the best on the BS-ERGB dataset, obtaining the lowest LPIPS score and highest SSIM. The self-supervised method SSL-E2VID obtains the best MSE scores on the ECD and MVSEC datasets, while also obtaining the worst SSIM and LPIPS scores on the ECD, HQF, and BS-ERGB datasets. This demonstrates the importance of the metrics used for image quality assessment and emphasizes the need for prudence while interpreting the results of these metrics. The ground truth images from the ECD and MVSEC datasets are often underexposed, and the reconstructions of SSL-E2VID are quite dark compared to other methods (Fig. 2). This brings an advantage to the SSL-E2VID method in terms of MSE scores on the ECD and MVSEC datasets, but one should note that MSE scores are not always in line with human perception of image fidelity [28]. While the other methods achieve lower performance than these three, some of them still obtain relatively good results on specific datasets. These results demonstrate that the choice of the method can depend on the dataset, highlighting the importance of evaluating methods on multiple datasets to assess their generalization ability.

Table 3 presents the results of the quantitative analysis on challenging scenarios involving fast motion, low light, and high-dynamic range, assessed by using no-reference metrics BRISQUE, NIQE, and MANIQA. Among the methods compared in the table, in general, FireNet+ and E2VID achieve the best results. SPADE-E2VID and SSL-E2VID achieve the lowest scores in all three metrics compared

	ECD [13]			MVSEC [34]			HQF [24]			BS-ERGB [25]		
	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓
E2VID [21]	0.212	0.424	0.350	0.337	0.206	0.705	0.127	0.540	0.382	0.121	0.446	0.548
FireNet [23]	0.131	0.502	0.320	0.292	0.261	0.700	0.094	0.533	0.441	0.096	0.449	0.532
E2VID+ [24]	0.070	0.560	0.236	0.132	0.345	0.514	0.036	0.643	**0.252**	0.075	**0.493**	**0.432**
FireNet+ [24]	0.063	0.555	0.290	0.218	0.297	0.570	0.040	0.614	0.314	0.090	0.453	0.480
SPADE-E2VID [2]	0.091	0.517	0.337	0.138	0.342	0.589	0.077	0.521	0.502	0.090	0.464	0.641
SSL-E2VID [18]	**0.046**	0.364	0.425	**0.062**	0.345	0.593	0.126	0.295	0.498	0.192	0.196	0.676
ET-Net [30]	0.047	**0.617**	**0.224**	0.107	**0.380**	**0.489**	**0.032**	**0.658**	0.260	**0.071**	0.491	0.442

Table 2. **Full-reference quantitative results on the ECD, MVSEC, HQF, and BS-ERGB datasets**. In here, we use between-frames event grouping. No pre-processing or post-processing is applied. The best and second best scores are given in **bold** and underlined.

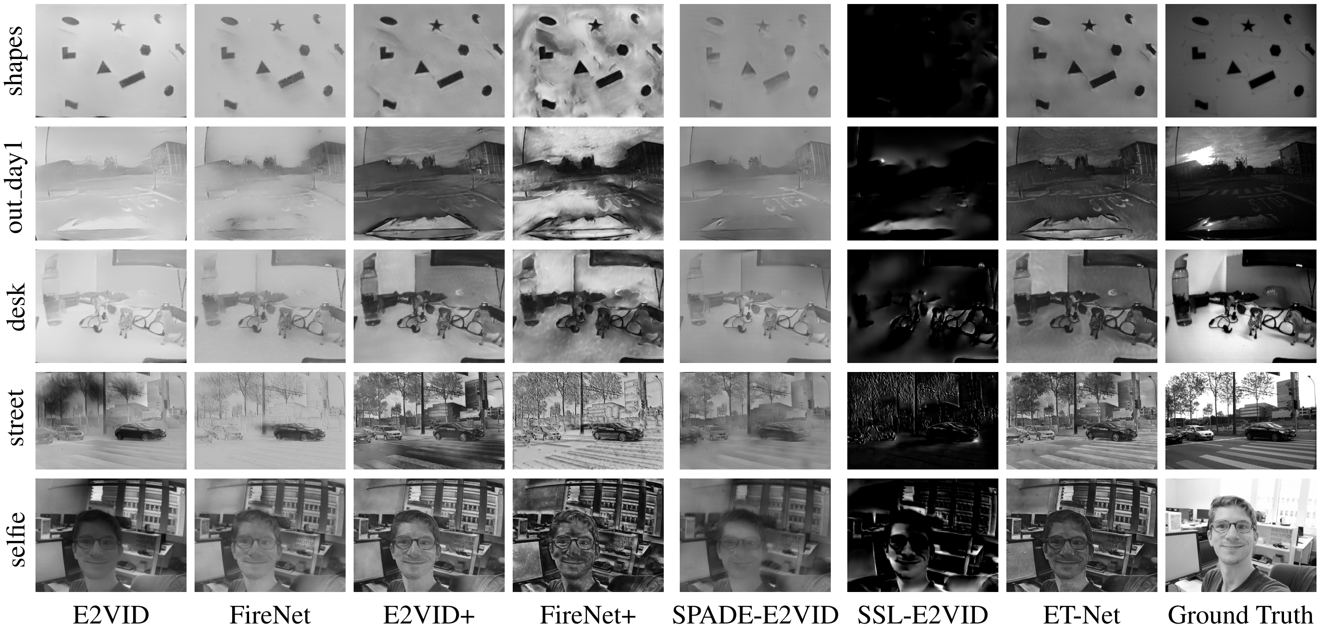


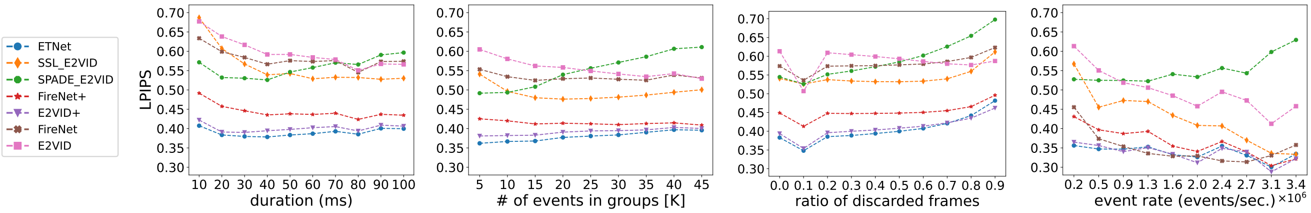
Figure 2. **Qualitative comparisons**. A sample scene from the ECD, MVSEC, HQF, BS-ERGB, and HDR datasets is given at each row, from top to bottom, respectively (Note that the rightmost image in the last row is a reference frame rather than the groundtruth.)

	ECD-FAST [13]			MVSEC-NIGHT [34]			HDR [21]		
	BRISQUE↓	NIQE↓	MANIQA↑	BRISQUE↓	NIQE↓	MANIQA↑	BRISQUE↓	NIQE↓	MANIQA↑
E2VID [21]	**14.729**	7.991	**0.407**	**8.793**	5.191	0.431	18.330	4.209	**0.4255**
FireNet [23]	19.957	8.095	0.393	21.311	6.071	0.413	21.540	4.085	0.389
E2VID+ [24]	22.627	6.734	0.372	12.285	4.313	0.379	21.340	3.903	0.3751
FireNet+ [24]	18.399	**5.460**	0.395	10.019	**4.306**	**0.439**	**15.680**	**3.236**	0.3813
SPADE-E2VID [2]	18.925	10.008	0.094	24.011	8.485	0.434	25.838	5.567	0.4069
SSL-E2VID [18]	60.523	19.201	0.350	63.847	14.133	0.373	69.471	9.358	0.3525
ET-Net [30]	19.698	7.530	0.381	15.533	5.229	0.416	23.526	3.643	0.3791

Table 3. **No-reference quantitative results on challenging sequences involving fast motion, low light, and high-dynamic range**. Here, we use between-frames event grouping for ECD-FAST and MVSEC-NIGHT, and fixed-duration event grouping for HDR with a duration of 40 ms. No pre-processing or post-processing is applied. The best and second best results are given in **bold** and underlined.

to other methods. Interestingly, ET-Net, the model that achieves the best scores on standard benchmark datasets in terms of full-reference metrics (*cf.* Table 2), performs poorly in these challenging situations. These results suggest that to assess the overall effectiveness of the image reconstruction methods from events, standard benchmark sequences are not enough and further analysis is needed.

Table 4 shows the quantitative results of image reconstruction methods on three downstream tasks, including results using ground truth intensity frames as a baseline for comparison. The evaluation metrics employed are AP (Average Precision) for object detection, accuracy for image



(a) (b) (c) (d)

Figure 3. **Robustness analysis.** We investigate how factors including (a) image reconstruction rate, (b) event tensor sparsity, (c) temporal irregularity (c), and (d) event rate affect the performance of each method.

Methods	Obj. Det. AP (%)	Img. Class. Accuracy (%)	Img. Cal. MAPE (%)
E2VID [21]	**47.51**	**71.34**	1.89
FireNet [23]	<u>45.78</u>	67.91	<u>1.63</u>
E2VID+ [24]	30.58	<u>70.49</u>	10.24
FireNet+ [24]	9.83	47.18	3.90
SPADE-E2VID [2]	19.90	19.53	**0.70**
SSL-E2VID [18]	3.35	33.90	14.89
ET-Net [30]	21.24	26.61	3.46
Ground Truth Frames	66.84	–	5.06

Table 4. **Quantitative results on downstream tasks.** The best and second best results are highlighted in **bold** and <u>underlined</u>.

classification, and MAPE (Mean Absolute Percentage Error) for camera calibration. E2VID achieves the highest scores on object detection and image classification tasks, with FireNet being the second and third best, respectively. E2VID+ obtains lower scores on object detection but still performs well on image classification. Its performance on camera calibration is, however, significantly worse than the first two methods. Conversely, SPADE-E2VID has the lowest score on object detection but the highest score on camera calibration. However, its performance on image classification is substantially lower than the other methods. Even though ET-Net achieves state-of-the-art results in full-reference image quality metrics, its downstream task performance is relatively low compared to other methods.

The object detector achieves the highest score when run on the original intensity images. Interestingly, on the camera calibration task, using intensity sequence does not give the minimum MAPE scores. For the image classification task, since the N-Caltech101 dataset does not include intensity images, we leave the accuracy of the intensity images blank in the last row. In conclusion, choosing a method depends on the specific downstream task: E2VID is superior for night-time vehicle detection and image classification, while SPADE-E2VID is the best performer for camera calibration. However, it is important to consider other factors such as model complexity, training time, and dataset size when choosing a method. Intensity images provide a strong baseline for the object detection task, and further research is needed to improve object detection performance.

Fig. 3 shows plots of mean LPIPS scores for robustness analysis. As the event grouping duration increases (3a), some of the worse-performing methods start to improve while the best-performing methods maintain their performance. When the number of events in groups increases (3b), the performance of SPADE-E2VID decreases significantly, while a decrease in the number of events reduces the performance of SSL-E2VID and E2VID. The other methods remain fairly robust to changes in this setting. Interestingly, as we discard 10% of the ground truth frames (3c), all the performances improve, which may be an indication of a sub-optimal event grouping in the original setting. As the discard ratio is increased above 0.1, the performances decrease significantly, except for E2VID. SPADE-E2VID is susceptible to event rate increase (3d), while this change is beneficial for some other methods.

We also analyzed the computational complexity of each method, which is presented in the supplementary material.

4. Conclusion

This paper presents a framework called EVREAL, which provides a unified evaluation scheme for event-based video reconstruction methods. EVREAL can serve as a valuable resource for researchers and practitioners working in event-based vision. In this study, we utilized EVREAL to analyze state-of-the-art models and yielded insightful observations on their performance under varying settings, challenging scenarios, and downstream tasks. These models, however, require certain event representations as their inputs, making evaluating them with different event representations impractical. This could be considered a limitation of the current work. Future work will include incorporating a temporal consistency metric, expanding our test datasets, exploring additional downstream tasks, and developing color image reconstruction capabilities in conjunction with model training. Overall, we believe that our work will contribute to the development of more effective and robust event-based video reconstruction models.

5. Acknowledgments

This work was supported in part by KUIS AI Research Award, TUBITAK-1001 Program Award No. 121E454, and BAGEP 2021 Award of the Science Academy to A. Erdem.

References

[1] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 884–892, 2016. 3

[2] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. 3, 4, 7, 8

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88:303–308, 2009. 6

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[6] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[8] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic DVS events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1312–1321, 2021. 6

[9] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 170–185, 2018. 3

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 6

[11] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 3, 5

[12] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3, 5

[13] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 3, 5, 7

[14] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. How to calibrate your event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1403–1409, 2021. 6

[15] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 3

[16] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 3, 6

[17] Luc Oth, Paul Furgale, Laurent Kneip, and Roland Siegwart. Rolling shutter camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1360–1367, 2013. 6

[18] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 3, 4, 5, 7, 8

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 2

[20] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 1, 5

[21] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 3, 4, 5, 7, 8

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[23] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 156–163, 2020. 3, 4, 5, 7, 8

[24] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conf. Comput. Vis.(ECCV)*, 2020. 1, 3, 4, 7, 8

[25] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 3, 5, 7

[26] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 6

[27] Lin Wang, S. Mohammad Mostafavi, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 3

[28] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009. 6

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 3, 5

[30] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 1, 3, 4, 7, 8

[31] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 3, 5

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 3, 5

[33] Zelin Zhang, Anthony Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, early access, Dec. 20, 2022. 3

[34] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 3, 5, 7

[35] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3604, 2022. 3

[36] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 4

EVREAL: Towards a Comprehensive Benchmark and Analysis Suite for Event-based Video Reconstruction

Supplementary Material

Burak Ercan [1,2] Onur Eker [1,2] Aykut Erdem [3,4] Erkut Erdem [1,4]

[1] Hacettepe University, Computer Engineering Department [2] HAVELSAN Inc.
[3] Koç University, Computer Engineering Department [4] Koç University, KUIS AI Center

In this supplementary document, we provide additional material to complement the main paper. First, we present recent related work on event-based video reconstruction, especially focusing on their evaluation details (Sec. 1). Second, we share the details of the event representation that we have employed (Sec. 2). Third, we share implementation details of the evaluation metrics we considered in our analysis (Sec. 3). Next, we provide the overview of the datasets being used in our proposed EVREAL framework (Sec. 4). Then, we present the details and results from the computational complexity analysis that we performed (Sec. 5). Finally, we share additional qualitative results from several datasets (Sec. 6).

1. Related Work

In recent years, there has been a surge of methods aiming to reconstruct intensity images from events, each taking into account different assumptions and employing distinct processing methodologies. Early approaches were limited, often relying on basic assumptions such as known or restricted camera movement, static scenes, or brightness constancy. More recent methods utilize deep neural networks and incorporate natural image priors in their models to achieve better results. Here, we limit our discussion to these recent methods and especially focus on their evaluation details.

Wang *et al*. [33] proposed a conditional GAN based model, in which input events are represented by means of spatio-temporal voxel grids. Their evaluation setup includes a small amount of data containing 1000 intensity frames taken from both real and simulated datasets, including the sequences from [1]. They compared their method against [1, 17] for sequences without any ground truth outputs, by utilizing the no-reference metric BRISQUE [14]. The authors do not share their evaluation code.

Rebecq *et al*. [23, 24] introduced a recurrent fully convolutional network, named E2VID. The authors used a selection of seven sequences from the ECD [16] dataset, using a fixed number of events to form event voxel grids and a tolerance of 1 ms to match the reconstructions with ground truth frames. To improve the output quality, they applied robust normalization as a post-processing step and then performed local histogram equalization before computing scores for MSE, SSIM and LPIPS [37]. They compared their approach against [1] and [17]. They also reported a temporal consistency score that requires a ground truth optical flow map between each frame. To obtain this, they used an off-the-shelf frame-based optical flow network [11], which has its own prediction errors. The researchers conducted experiments on challenging scenarios involving rapid motion, low-light conditions and high dynamic range, without providing any quantitative scores. Additionally, they reported color image reconstruction results from the event data available in CED dataset [27], without providing any quantitative analysis.

Rebecq *et al*. also evaluated their method on four downstream tasks, including image classification, visual-inertial odometry, object detection, and monocular depth estimation [23, 24]. To perform these tasks, they fed reconstructed frames as inputs to task-specific frame-based methods and reported either qualitative or quantitative results. For instance, for object classification, they used events from N-MNIST [18], N-Caltech101 [18], and N-Cars [28] datasets, and provided accuracy scores achieved by a ResNet-18 [10] network. Similarly, for visual-inertial odometry, they employed events from the ECD dataset, and investigated mean translation errors obtained via VINS-Mono [21]. For object detection and monocular depth estimation, they used YOLOv3 [25] and MegaDepth [13], respectively, and only shared qualitative results in a supplementary video. Additionally, they analyzed the computational efficiency of their approach by reporting the frame synthesis time. The authors do not release their evaluation code publicly.

Scheerlinck *et al*. [26] proposed FireNet, a lightweight recurrent network, as a replacement for E2VID, and demonstrated that it can attain similar performance with much less memory consumption and faster inference. In their evaluation setup, they followed the methodology in [24], and performed experiments on the selected frames from the sequences in the ECD dataset. They utilized a fixed number of events to form event voxel grids, and applied local histogram equalization to reconstructions and ground truth frames before estimating quantitative metrics such as MSE, SSIM, and LPIPS. Additionally, they performed qualitative analysis on color image reconstruction and challenging scenarios involving high-dynamic range and fast motion. They focused on evaluating computational efficiency and compared several resolutions on GPU and CPU by examining the number of model parameters, memory consumption, FLOPs, and inference times. However, they did not conduct any downstream task experiments, and their evaluation codes are not made publicly available.

Stoffregen *et al*. [30] proposed an enhanced version of E2VID, named E2VID+, by retraining it on synthetic training data exhibiting similar statistics with real-world test data. They also employed the same strategy for improving the FireNet architecture, resulting in FireNet+. They evaluated their methods on a larger set of real-world sequences from three datasets, namely ECD and MVSEC [39] datasets, and their proposed HQF dataset. For ECD and MVSEC, they used the sequences commonly used in earlier work, and reported MSE, SSIM, and LPIPS scores. They always had a matching ground truth frame for each reconstruction, as they used events between each consecutive ground truth frame to form voxel grids. It is not clear whether they applied normalization or histogram equalization before calculating these scores. Moreover, they did not perform any experiments on challenging scenarios or downstream tasks, nor did they perform computational efficiency analysis. The evaluation code is not publicly available.

Cadena *et al*. [4] proposed SPADE-E2VID, which integrates spatially-adaptive denormalization (SPADE) [20] layers into the E2VID architecture to enhance the quality of the reconstructed videos. The authors evaluated their approach using seven sequences from the ECD dataset, starting from the very first frames of each sequence, and reported MSE, SSIM, and LPIPS scores for quantitative comparison with E2VID and FireNet. They also introduced an RMS contrast metric to demonstrate that their method produces higher contrast reconstructions. To assess temporal consistency, they used a different off-the-shelf frame-based optical flow network [22] and reported the corresponding scores. In addition, they performed object detection analysis on a single sequence of the ECD dataset, using events and YOLOv4 [2] to process reconstructed frames. They estimated ground truth object labels for two object classes by applying the same object detection network to ground truth intensity images and shared average precision scores for this downstream task accordingly. They analyzed the computational efficiency of their approach by reporting reconstruction time for inputs with various resolutions. While they released an evaluation code, we were unable to reproduce their results with it.

Weng *et al*. [35] improved the E2VID architecture by adding a Transformer-based module to better exploit the global context of event tensors, thus naming their model as ET-Net. Their experiments were conducted using the ECD, MVSEC, and HQF datasets, with the same sequence cuts as in [30]. In their experiments, events between consecutive ground truth frames were used to form voxel grids. To evaluate their approach, they calculated MSE, SSIM, and LPIPS scores, without any normalization or histogram equalization applied to the reconstructed images. They compared their method with E2VID, E2VID+, FireNet, and FireNet+, and shared qualitative results on challenging scenarios involving high-dynamic-range and rapid motion in their supplementary material. However, they did not perform a computational efficiency analysis or an experiment on a downstream task. The authors provided an open-source evaluation code, and we are able to use to reproduce their results.

Paredes-Vallés and de Croon [19] proposed a self-supervised learning method called SSL-E2VID, which employs the event-based photometric constancy assumption [8] to estimate optical flow and intensity images simultaneously. As done in earlier work, events between each consecutive ground truth frame were used to form voxel grids. Their experiments were conducted on ECD and HQF datasets, and they made quantitative comparisons with E2VID, E2VID+, FireNet, and FireNet+. Local histogram equalization was employed before calculating quantitative scores. Since they did not introduce a new architecture, computational efficiency analysis was not performed. Qualitative results were given also for challenging scenarios such as high-dynamic-range and high-speed. No downstream task analysis was performed, and their evaluation code was not made publicly available.

Zhu et al. [40] proposed a spiking neural network architecture that achieves comparable performance to E2VID, E2VID+, FireNet, and SPADE-E2VID with higher computational efficiency. They used the ECD, MVSEC, and HQF datasets in their evaluation and reported quantitative scores using MSE, SSIM, and LPIPS metrics. In reconstructing intensity images, they used the events between each consecutive ground truth frame as input. They applied histogram equalization before calculating these scores. In addition, they provided an analysis of energy consumption. However, they did not release an open-source evaluation code.

Zhang *et al*. [38] presented a novel approach for event-based image reconstruction by formulating it as a linear inverse problem based on optical flow. They conducted a quantitative comparison with E2VID, E2VID+, and SSL-E2VID using MSE, SSIM, and LPIPS metrics. They focused on test sequences with limited camera motion, specifically selected from the ECD dataset, and utilized events from N-Caltech101 [18] dataset. They aligned reconstructions with respective reference frames using Enhanced Correlation Coefficient Maximization [6]. They reported median scores for each sequence instead of mean scores and presented distribution plots of scores of each method on various sequences. They also analyzed the effect of histogram equalization on quantitative scores and emphasized the importance of taking various factors into account while interpreting these scores. They showcased their method's ability to perform color reconstruction and demonstrated temporal consistency on two example frames from the DSEC dataset [9]. They did not conduct experiments on downstream tasks and did not share their evaluation code.

2. Details of Event Representation

Following the common practice in the literature, we form voxel grids from grouped events in order to utilize deep CNN architectures for event-based data. Let G_k be a group of events that span a duration of ΔT seconds, T_k be the starting timestamp of that duration, and B be the number of temporal bins used to discretize the timestamps of continuous-time events in the group. The voxel grid $V_k \in \mathbb{R}^{W \times H \times B}$ for that group is formed by normalizing the timestamps of events from the group to the range $[0, B-1]$. Each event contributes its polarity to the two temporally closest voxels using a linearly weighted accumulation similar to bilinear interpolation. Specifically, the voxel grid is computed as follows:

$$V_k(x, y, t) = \sum_i p_i \max(0, 1 - |t - t_i^*|)\delta(x - x_i, y - y_i) \tag{1}$$

where δ is the Kronecker delta that selects the pixel location, and t_i^* is the normalized timestamp which is calculated as:

$$t_i^* = (B-1)(t_i - T_k)/(\Delta T) \tag{2}$$

In all our experiments, we set the number of temporal bins B to 5.

3. Implementation Details for Quantitative Image Quality Metrics

MSE. The Mean Squared Error is a commonly used metric that does not require any parameters. When comparing two images, the only factor that can impact the MSE result is the range of pixel values that the images possess. We calculate the MSE using floating-point pixel values within the range $[0, 1]$. Lower MSE values indicate better results.

SSIM. We utilize the `scikit-image` image processing library's [32] implementation for structural similarity. We adjust the parameters to use the Gaussian weighting scheme explained in [34]. Like MSE, we compute SSIM using images with floating point pixel values in the range of $[0, 1]$. Higher scores of SSIM indicate better results.

LPIPS. We utilize the official implementation of LPIPS [37][1], v0.1.4, and employ the variant that uses the pre-trained AlexNet [12] network. To comply with the implementation, we normalize the images so that their pixel values fall in the range of $[-1, 1]$. In the LPIPS score calculation, a lower score indicates better quality.

BRISQUE. For BRISQUE [14], we use the implementation in IQA-PyTorch toolbox [5][2], v0.1.5, with default settings. The implementation supports 3-channel RGB images. Therefore, we convert intensity images into RGB images by concatenating three copies of the grayscale image along the third channel before calculating the scores.

NIQE. For NIQE [15], we again use the implementation in IQA-PyTorch toolbox [5], v0.1.5, with default settings. The implementation supports 3-channel RGB images. Therefore, we convert intensity images into RGB images by concatenating three copies of the grayscale image along the third dimension before calculating the scores.

MANIQA. For MANIQA [36], we also use the implementation in IQA-PyTorch toolbox [5], v0.1.5, with default settings. The implementation supports 3-channel RGB images. Therefore, we convert intensity images into RGB images by concatenating three copies of the grayscale image along the third channel before calculating the scores. MANIQA works by taking random crops of size 224×224 pixels from the images, whereas the ECD dataset used in our analysis has a lower resolution. To address this discrepancy, we upscale the images to the desired size before calculating the scores.

4. Dataset Details

Event Camera Dataset (ECD). Following the practice explained in [24], we use seven different sequence with diverse characteristics from the ECD dataset [16]. These sequences are short, taken indoors, and mostly contain simple scenes of office environments with stable objects. The data was captured by a DAVIS240C sensor [3], which is mostly moving with 6 degrees of freedom (DOF) and with increasing speed. The camera generates events and frames from the same pixel array, which has a spatial resolution of 240×180

[1] The code is accessible from https://github.com/richzhang/PerceptualSimilarity

[2] The code is accessible from https://github.com/chaofengc/IQA-PyTorch

pixels. The ground truth intensity frames are available at an average rate of 22 Hz.

To allow methods to generate meaningful results, we exclude the initial few seconds of each sequence from quantitative evaluation. Additionally, when using full-reference metrics, as commonly done in earlier work, we do not include the latter parts of the sequences as they may contain motion blur due to the increased speed of camera movement. However, when evaluating with no-reference metrics, we specifically concentrate on these sections that have fast camera movement, to which the corresponding ground truth intensity images are of lower quality.

Multi Vehicle Stereo Event Camera (MVSEC) dataset. The MVSEC dataset [39] contains longer sequences captured by a pair of DAVIS 346B cameras, each having a spatial resolution of 346 × 260 pixels. These sequences depict both indoor and outdoor environments. To evaluate the quality of the videos generated by the methods using full-reference metrics, we followed the approach taken by [30] and considered six commonly used sequences from this dataset. Four of these sequences were captured indoors by a flying hexacopter, while the remaining two were taken outdoors during the daytime from a driving vehicle. The average rate of ground truth intensity frames was approximately 30 Hz for indoor sequences and 45 Hz for outdoor sequences. Additionally, we used three night sequences from this dataset, each captured from a vehicle as well, for our experimental evaluation involving no-reference metrics as the ground truth frames at night-time tend to be underexposed.

High-Quality Frames (HQF) dataset. The HQF dataset [30] contains fourteen sequences that exhibit a wide range of different motion behaviors, including static, slow, and fast camera motion, and cover both indoor and outdoor scenes. Two different DAVIS240C cameras are used to capture the data, providing distinct noise and contrast threshold characteristics. The cameras generate events and intensity frames from the same 240 × 180 pixel array. The scenes and camera parameters are adjusted to ensure that the ground truth frames are well-exposed and have minimal motion-blur. The average rate of ground truth intensity frames is 22.5 Hz. We use the entire sequences from this dataset for evaluation using full-reference quantitative metrics.

Beam Splitter Event and RGB (BS-ERGB) Dataset. The BS-ERGB Dataset [31] is originally collected for the event-based video frame interpolation task. The dataset consists of events recorded by a Prophesee Gen4M event camera [7] having a spatial resolution of 1280×720 pixels, and RGB frames captured by a global shutter RGB Flir camera with a resolution of 4096×2196 pixels. Both of these data are then post-processed to have the same spatial resolution of 970×625 pixels. Most of the sequences are short and captured with a static camera observing fast motions in the scene. Since events are confined to small regions where motion is observed, reconstructing intensity frames for other parts of the scene is not feasible. There are a few sequences recorded with a handheld camera where every pixel generates many events. We evaluate the models on ten of these handheld sequences.

Network Architecture	Number of Params (M)	Inference Time (ms)
E2VID [19, 24, 30]	10.71	5.1
FireNet [26, 30]	0.04	1.6
SPADE-E2VID [4]	11.46	16.1
ET-Net [35]	22.18	32.1

Table 1. **Computational complexity of different network architectures** in terms of the number of model parameters (in millions) and inference time (in milliseconds).

High Speed and HDR Datasets These high-speed and HDR sequences are recorded by Rebecq *et al.* [24], using a Samsung DVS Gen3 event camera [29] with a spatial resolution of 640×480. We use all three HDR sequences from this dataset, namely the hdr_selfie, hdr_sun, and hdr_tunnel sequences.

5. Computational Complexity

We also analyzed the computational complexity of each method by considering two metrics: the number of model parameters and inference time. The former is an essential metric as it indicates the memory requirements, while the latter reflects the real-time performance by determining the maximum FPS that can be achieved. To measure the inference time, we used a workstation equipped with a Quadro RTX 5000 GPU and considered data with a spatial resolution of 240×180. We report the average inference time for each method in ms. Table 1 compares the computational complexity of image reconstruction methods. In this table, we use the same row for the methods that share the same deep architecture. Overall, in terms of the number of parameters and inference times, FireNet is much smaller and faster than E2VID, while SPADE-E2VID is slightly larger and slower. ET-Net has the highest number of parameters which is twice as large as SPADE-E2VID, the second largest model, and its inference time is approximately 6× slower than E2VID and 20× slower than FireNet.

6. Additional Qualitative Results

Here, we provide qualitative comparisons for various sequences from the ECD, MVSEC, HQF, BS-ERGB, ECD-FAST, and MVSEC-NIGHT datasets. We present these results in Figures 1-6.

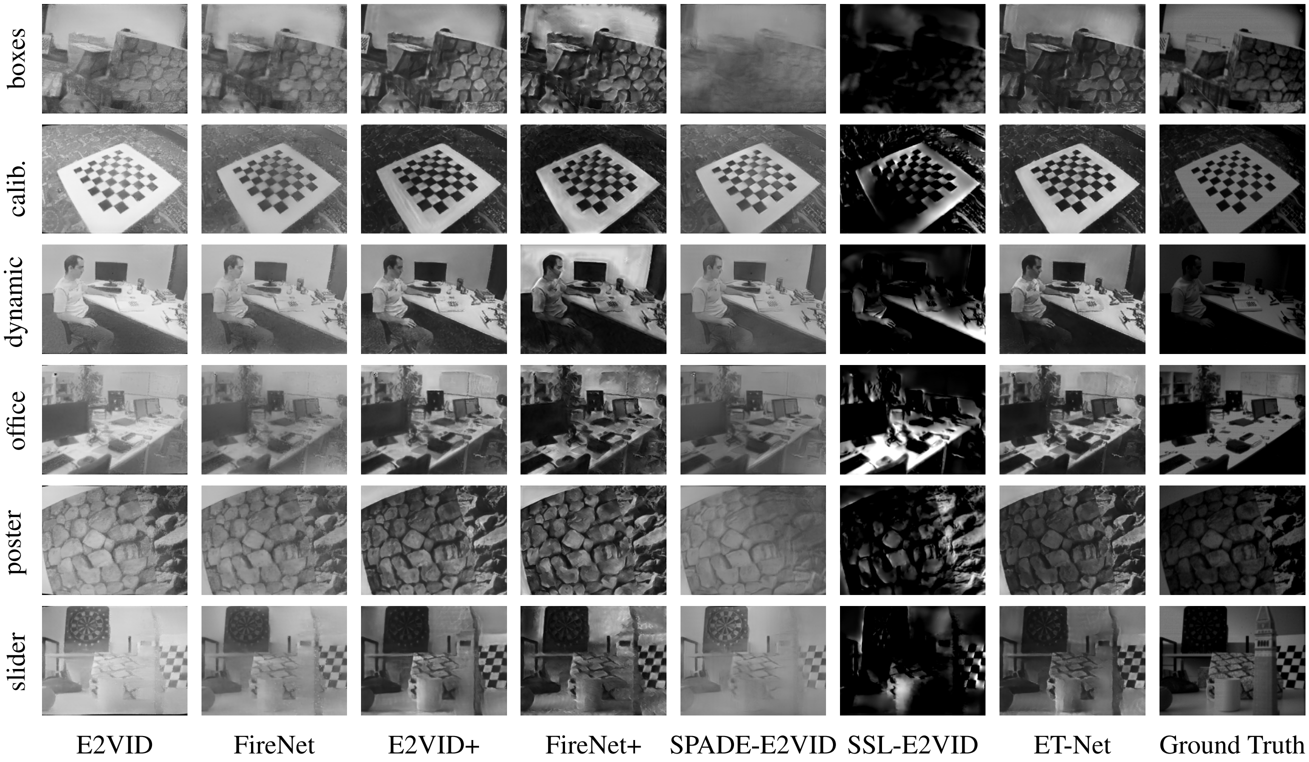


Figure 1. Additional qualitative comparisons on the ECD dataset.

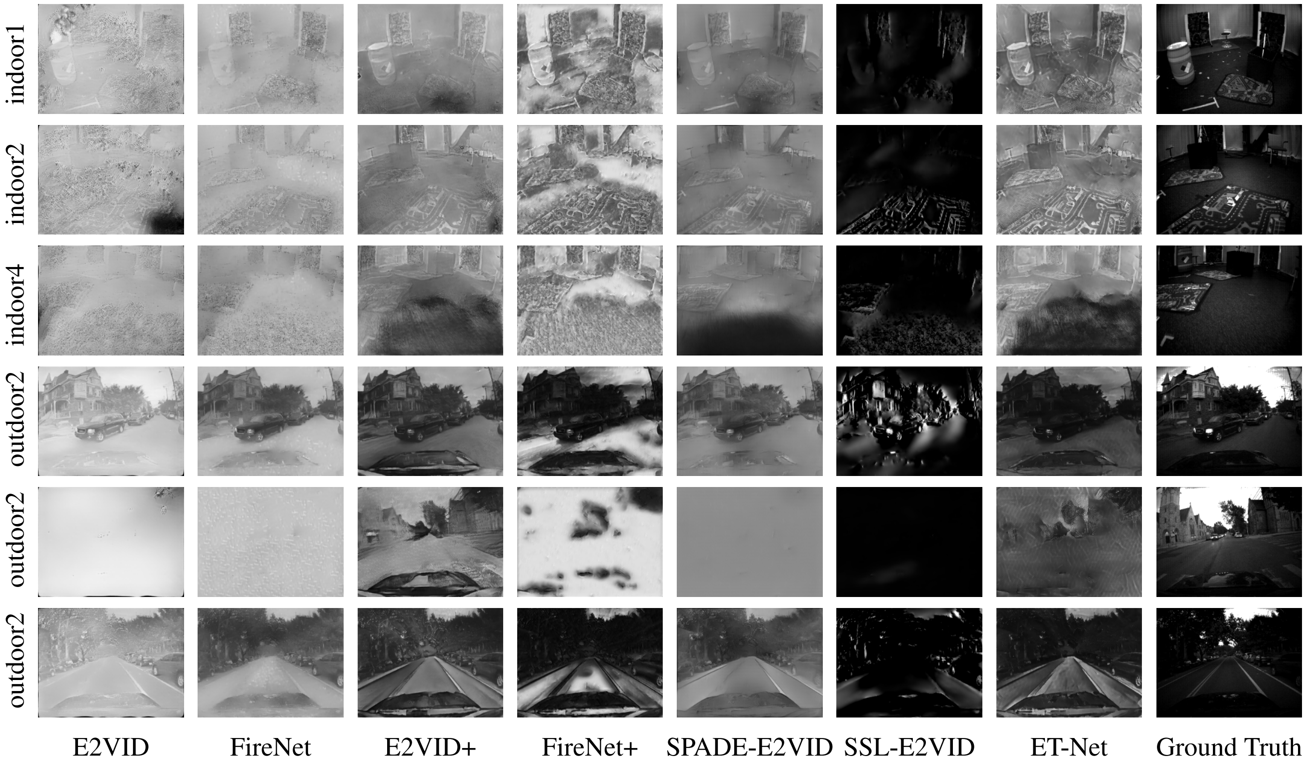


Figure 2. Additional qualitative comparisons on the MVSEC dataset.

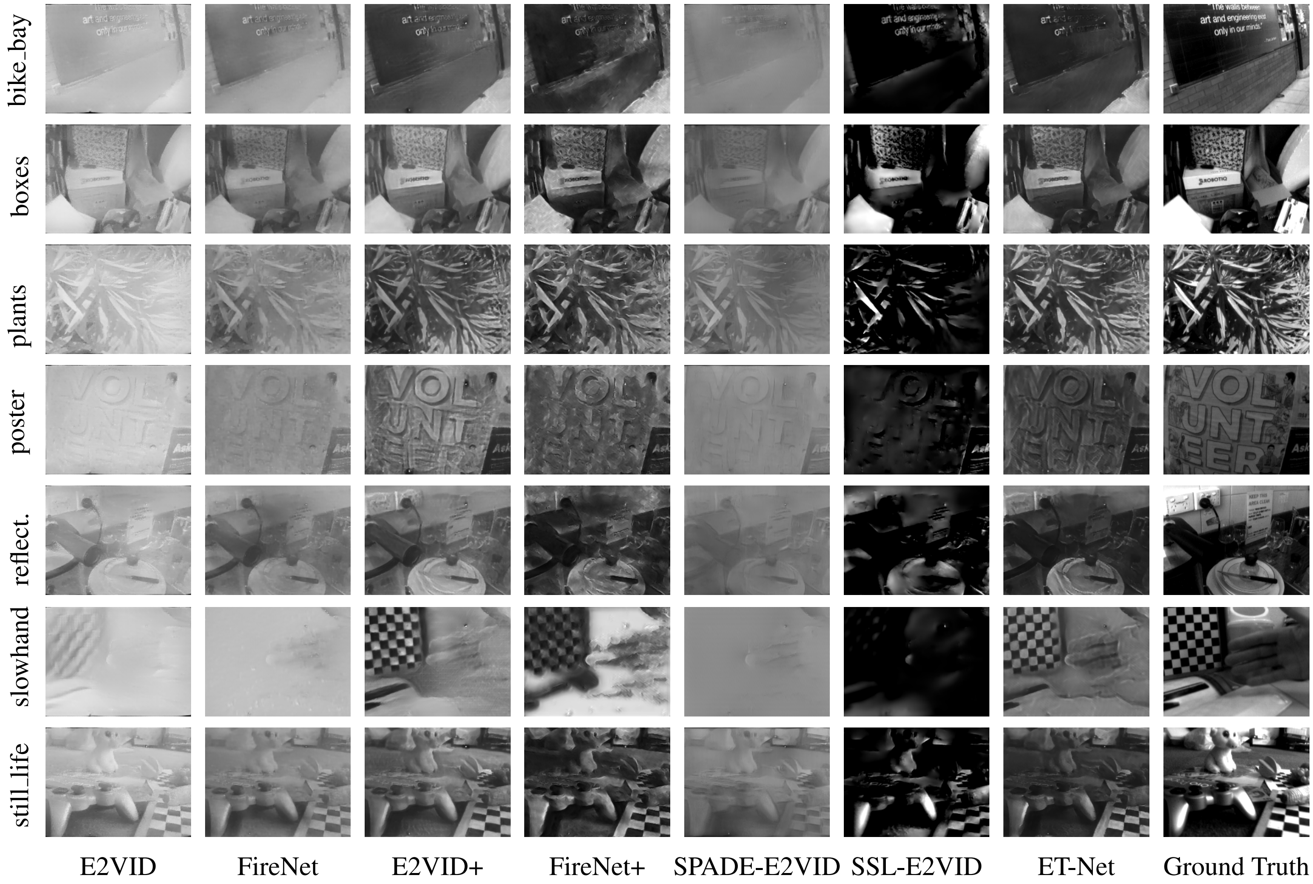


Figure 3. Additional qualitative comparisons on the HQF dataset.

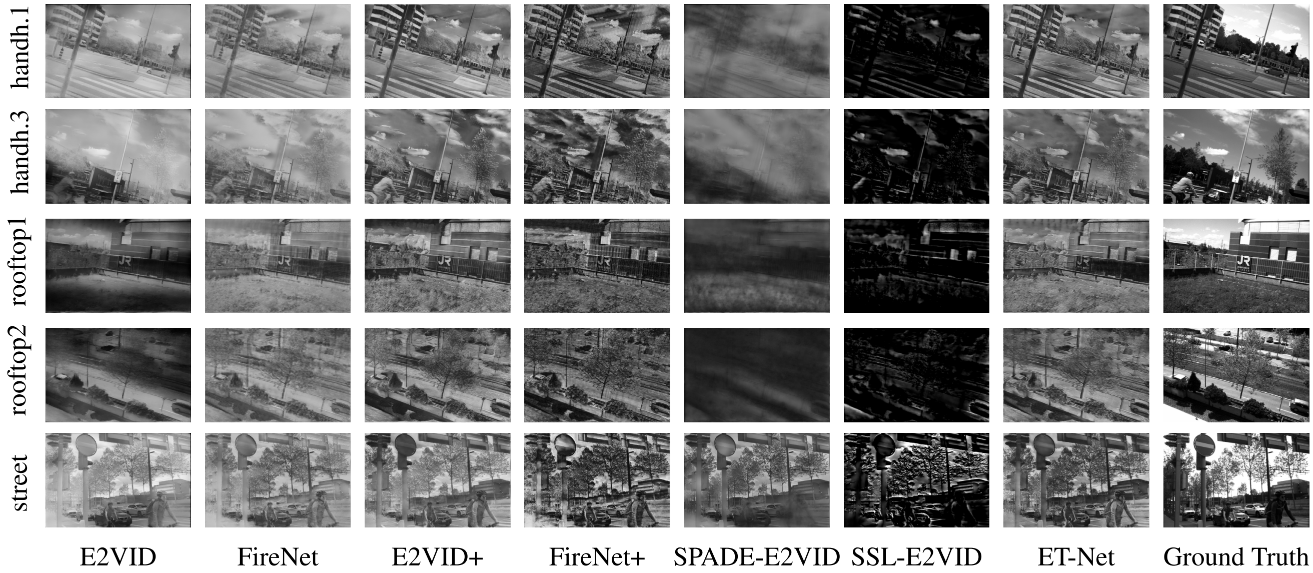


Figure 4. Additional qualitative comparisons on the BS-ERGB dataset.

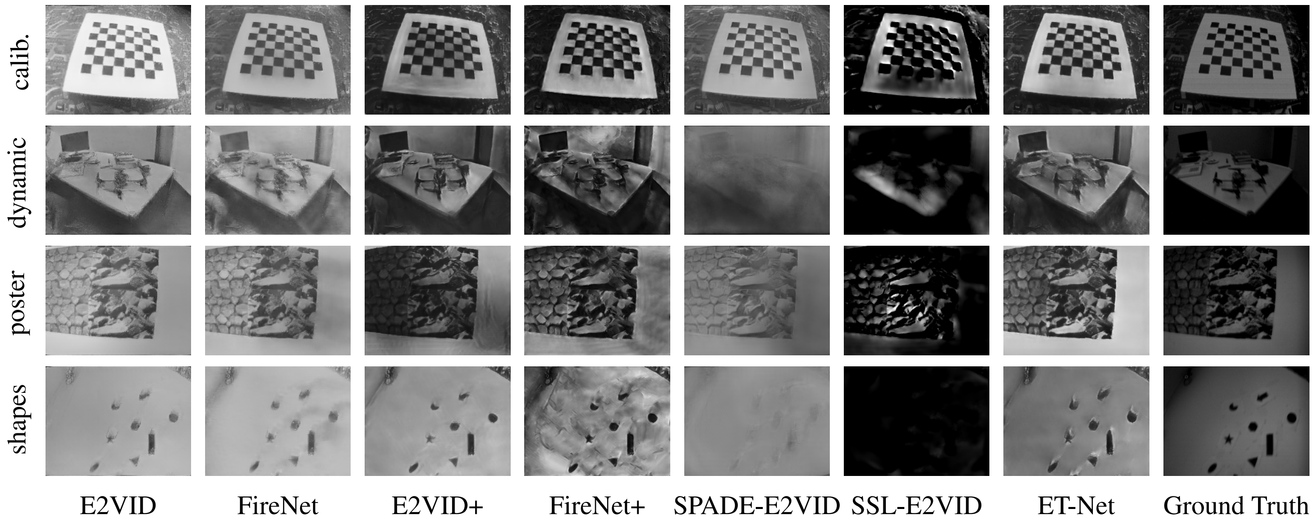


Figure 5. Additional qualitative comparisons on the fast parts of the ECD dataset (ECD-FAST).

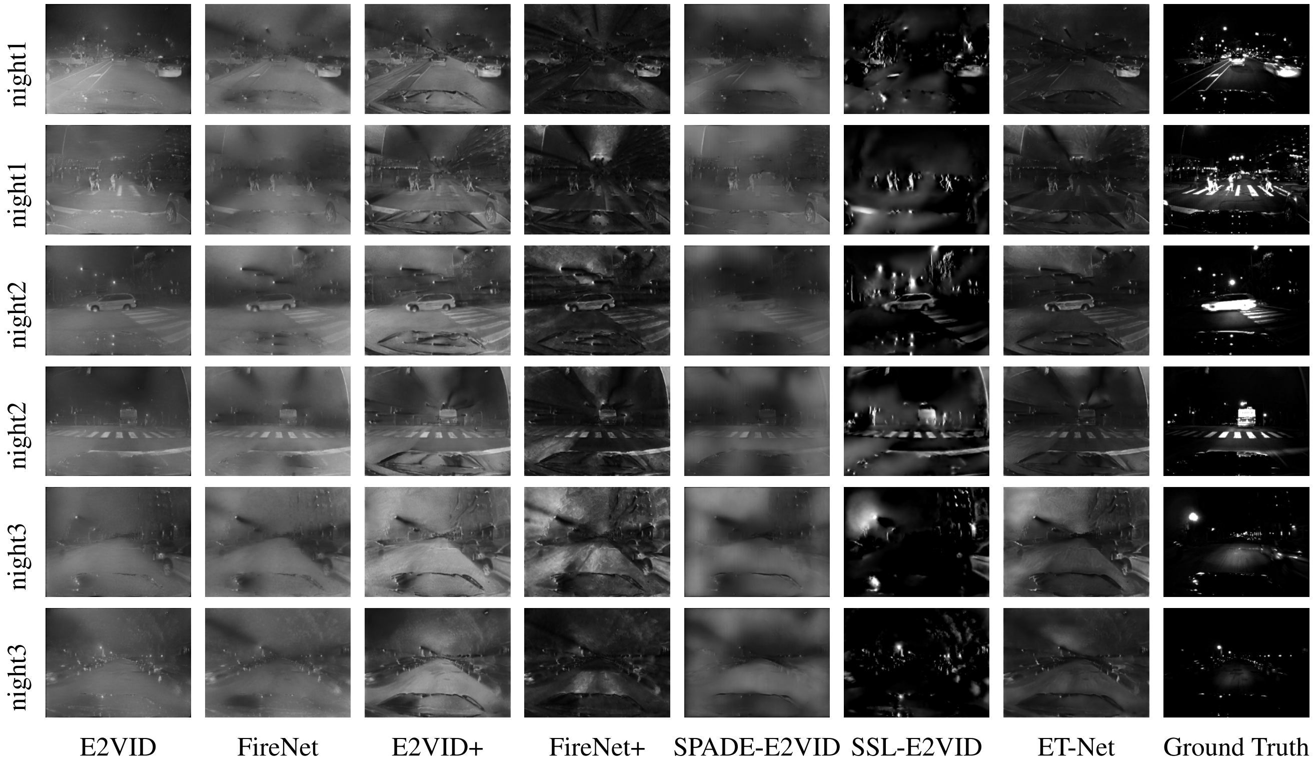


Figure 6. Additional qualitative comparisons on the night sequences of the MVSEC dataset (MVSEC-NIGHT).

References

[1] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 884–892, 2016. 1

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 3

[4] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. 2, 4

[5] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. 3

[6] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1858–1865, 2008. 3

[7] Thomas Finateu, Atsumi Niwa, Daniel Matolin, Koya Tsuchimoto, Andrea Mascheroni, Etienne Reynaud, Pooria Mostafalu, Frederick Brady, Ludovic Chotard, Florian LeGoff, et al. 5.10 a 1280×720 back-illuminated stacked temporal contrast event-based vision sensor with 4.86 μm pixels, 1.066 geps readout, programmable event-rate controller and compressive data-formatting pipeline. In *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*, pages 112–114. IEEE, 2020. 4

[8] Guillermo Gallego, Christian Forster, Elias Mueggler, and Davide Scaramuzza. Event-based camera pose tracking using a generative event model. *arXiv preprint arXiv:1510.01972*, 2015. 2

[9] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 3

[13] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[14] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 1, 3

[15] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3

[16] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 1, 3

[17] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 1

[18] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 1, 3

[19] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 2, 4

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 2

[21] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 1

[22] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 2

[23] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 1

[24] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 3, 4

[25] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[26] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 156–163, 2020. 1, 4

[27] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. CED:

color event camera dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[28] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018. 1

[29] Bongki Son, Yunjae Suh, Sungho Kim, Heejae Jung, Jun-Seok Kim, Changwoo Shin, Keunju Park, Kyoobin Lee, Jinman Park, Jooyeon Woo, et al. 4.1 a 640×480 dynamic vision sensor with a 9μm pixel and 300meps address-event representation. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 66–67. IEEE, 2017. 4

[30] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conf. Comput. Vis.(ECCV)*, 2020. 2, 4

[31] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 4

[32] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014. 3

[33] Lin Wang, S. Mohammad Mostafavi, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 1

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[35] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 2, 4

[36] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 3

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 3

[38] Zelin Zhang, Anthony Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, early access, Dec. 20, 2022. 2

[39] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multi-vehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 2, 4

[40] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3604, 2022. 2