Unified Reconstruction of Static and Dynamic Scenes from Events

Qiyao Gao[1,2,3#] Peiqi Duan[1,2#] Hanyue Lou[1,2] Minggui Teng[1,2]
Ziqi Cai[1,2] Xu Chen[3] Boxin Shi[1,2 †]

[1] State Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] Mechatronics, Automation, and Control Systems Laboratory, University of Washington

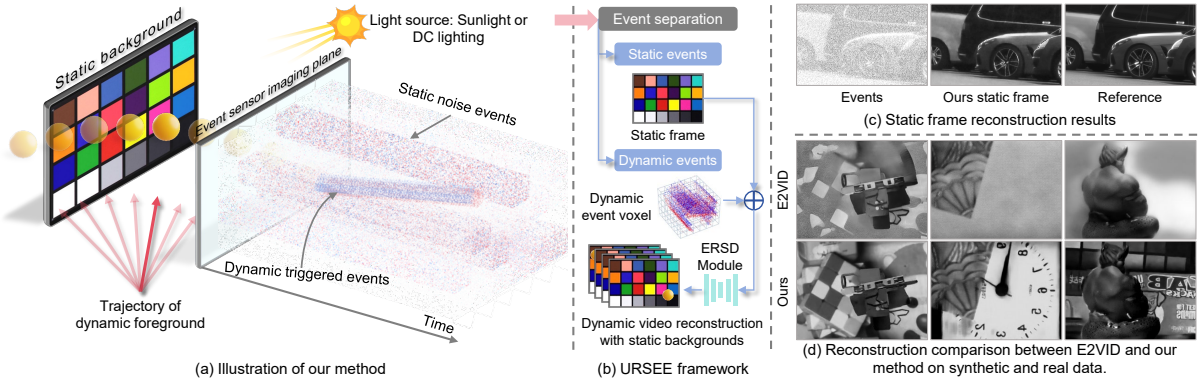

Figure 1. Illustration of our method (a) and pipeline (b), result examples of static frame reconstruction (c) and static-dynamic fusion (d).

Abstract

This paper addresses the challenge that current event-based video reconstruction methods cannot produce static background information. Recent research has uncovered the potential of event cameras in capturing static scenes. Nonetheless, image quality deteriorates due to noise interference and detail loss, failing to provide reliable background information. We propose a two-stage reconstruction strategy to address these challenges and reconstruct static scene images comparable to frame cameras. Building on this, we introduce the URSEE framework designed for reconstructing motion videos with static backgrounds. This framework includes a parallel channel that can simultaneously process static and dynamic events, and a network module designed to reconstruct videos encompassing both static and dynamic scenes in an end-to-end manner. We also collect a real-captured dataset for static reconstruction, containing both indoor and outdoor scenes. Comparison results indicate that the proposed method achieves state-of-the-art performance on both synthetic and real data.

1. Introduction

Event cameras are a novel type of neuromorphic sensors that introduce a paradigm shift in visual representation by responding to brightness changes rather than capturing absolute light intensity at a fixed rate [16, 33]. Each event pixel operates asynchronously and triggers a stream of events for dynamic scenes with high temporal resolution [6, 9, 37], leading to their widespread application in dynamic vision tasks [12, 19, 36]. To enable event cameras to leverage the mature vision algorithms designed for conventional frame cameras, event-based video reconstruction methods have been developed to bridge the gap stemming from the disparate output formats. However, current event-based video reconstruction methods [2, 8, 20, 22, 26, 39] often overlook restoring static backgrounds, resulting in a lack of background detail in reconstructed videos. While incorporating a frame-based camera may mitigate this issue, it diminishes the benefits associated with processing pure event streams, especially in terms of data efficiency.

Recent work has made progress in reconstructing static scenes from event streams, broadly categorized into two approaches based on whether external devices are required. Han *et al.* [13] control active illumination and EvTemMap

# Equal contribution, † Corresponding author
Project page: https://github.com/gaoqiyao1997/URSEE

[1] adjusts transmittance to trigger sufficient events for static scene reconstruction. Still, the application of these methods is limited by the need for specialized light sources or custom lenses. To eliminate the need for external devices, Cao *et al*. [3, 4] and Gao *et al*. [11] model the statistical relationship between noise events and scene intensity to enable static scene reconstruction under constant lighting. Nevertheless, these methods suffer from significant noise interference and detail loss, resulting in low-quality outputs. Besides, existing methods [11, 26] struggle to reconstruct both backgrounds and motion concurrently when facing scenes containing both static and dynamic elements. While Cao *et al*. [3] propose fusing backgrounds into motion videos reconstructed via E2VID [26] using masks, this approach is capped by the performance of the external network and the fusion process may cause artifacts at mask boundaries.

In this paper, we address the above challenges and propose a method for the Unified Reconstruction of Static and dynamic scEnes from Events, named **URSEE**, which carries the meaning of "universal see" all scenes of event camera capturing. Firstly, we measure the event-triggering rate within static scenes in a wide illumination range and analyze the noise impact under underexposure and overexposure. Based on this, we design URSEE as a two-stage strategy to reconstruct videos. The first stage introduces a convolutional integration method for static scene reconstruction to prevent noise accumulation and event saturation in existing methods. The second stage comprises a parallel channel to reconstruct videos featuring both static and dynamic scenes in an end-to-end manner. Furthermore, we build a real-captured event dataset for static scene reconstruction comprising diverse indoor and outdoor static scenes, and a dedicated synthetic dataset for static-dynamic scene reconstruction comprising scenes with varying scenarios for network training. The experimental comparison results with existing methods demonstrate the superior reconstruction performance of our proposed method, which also generalizes well to real-captured data. Our contributions are summarized below:

1. We analyze the influencing factors of event-based static reconstruction and propose a convolutional integration method and a denoising network, achieving state-of-the-art reconstruction results.
2. We introduce a unified framework URSEE to address the incompatibility between static and dynamic event processing prevalent in current methods. This framework enables end-to-end reconstruction of motion videos with static backgrounds from events for the first time.
3. We establish a real-captured dataset E-Static for event-based static scene reconstruction, comprising diverse indoor and outdoor static scenes, and generate a synthetic dataset E-StaDyn with varying static backgrounds and dynamic foregrounds.

2. Related Work

2.1. Event-based static reconstruction

Event-based static reconstruction methods can be broadly divided into two categories: with and without extra devices. The former requires additional equipment, such as active light sources, to create light changes. The latter exploits naturally occurring ambient light changes and necessitates no specialized hardware.

Shaw *et al*. [30] and Tulyakov *et al*. [34] combined event cameras and traditional cameras to fuse high-quality image information with complementary high-frequency and dynamic range information from events. Han *et al*. [13] proposed a method for recovering scene radiance by analyzing the transient event frequency during the split second of turning on the light. While effective in certain scenarios, these methods are limited by the need for specialized equipment.

Gao *et al*. [11] explored how event triggering rates correlate with object surface grayscale, focusing on factors like surface reflectance affecting event camera responses. Galor *et al*. [10] suggested reconstructing static scene intensity images by pixel-wise integration of event streams. Cao *et al*. [3, 4] extended this work by analyzing the noise characteristics of these responses. However, these methods are vulnerable to random noise and extreme pixel values during long integrations, leading to poor-quality reconstructions. URSEE improves the reconstruction quality of static images by introducing a convolutional integration method.

2.2. Event-based dynamic reconstruction

Benefiting from their high temporal resolution and high dynamic range, event cameras are ideal for dynamic visual tasks such as high-speed motion capture [7, 18, 32]. Rebecq *et al*. [26] introduced a recurrent network, named E2VID, that segments the incoming event stream into sequential spatio-temporal windows of events to reconstruct high-frame-rate videos. Stoffregen *et al*. [32] proposed a strategy to obtain high-quality videos by supplementing diversity datasets. Scheerlinck *et al*. [29] developed a lightweight network for fast video reconstruction from events. Cadena *et al*. [2] further introduced the SPADE-E2VID neural network model, which improves the quality of early frames in reconstructed videos and enhances overall contrast. Wang *et al*. [35] proposed a method based on conditional generative adversarial networks (cGANs), using stacks of space-time coordinates of events as input to reconstruct high-dynamic-range images and high-frame-rate videos. Nevertheless, these methods are inadequate for handling static scenes where only a few events are triggered, resulting in a lack of background information in the reconstructed videos. The proposed URSEE adopts a two-stage framework to reconstruct both static and dynamic scenes from events.

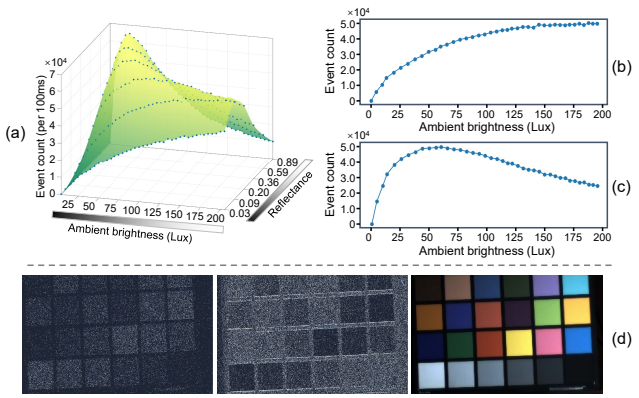

Figure 2. The mapping between ambient brightness, reflectance, and event count. (a) Statistical and fitting results of the event count for capturing the grayscale patches of the Macbeth ColorChecker under various lighting conditions. (b)/(c) The mapping relationship between ambient brightness and event counts with a reflectance of 0.09/0.36. (d) The event count map of real-captured events under different lighting conditions. Left: 20 lux. Middle: 170 lux. Right: image reference.

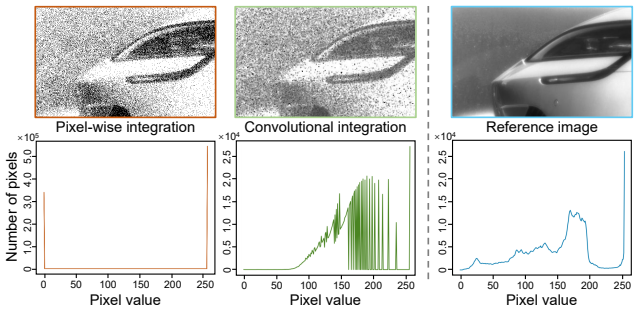

Figure 3. Qualitative and quantitative comparison of pixel-wise integration and convolutional integration. Top: Comparison of reconstruction results. Bottom: Pixel distribution statistics results. For quantitative comparison, we select the same test set of 96 different indoor and outdoor scenes. The scores of metrics for this evaluation are as follows (pixel-wise *vs.* ours): PSNR is 5.755 *vs.* **11.471**. SSIM is 0.022 *vs.* **0.233**. LPIPS is 1.381 *vs.* **0.748**.

3. Static Scene Reconstruction

In this section, we measure the response of event cameras in static scenes over a wide range of illumination and analyze the relationship between ambient brightness, reflectance, and event count under different lighting conditions in Sec. 3.1, present the convolutional integration process in Sec. 3.2 and the denoising module in Sec. 3.3.

3.1. Response of events in static scene

Recent studies [3, 4, 11] demonstrate that event cameras can generate stable event streams containing scene information even when there are no significant changes in light intensity, such as in environments illuminated by constant light sources or natural daylight. The events triggered in these situations are referred to as static events, and the event count of each pixel is closely associated with scene intensity. By integrating the event stream over time, an initial grayscale image can be reconstructed. However, the relationship between the event count and varying illumination conditions has been rarely explored, limiting the generalizability of integration-based reconstruction.

Therefore, we experimentally measure the event triggering rate (*i.e.*, event count per unit time) in static scenes across a wide range of illumination conditions (2.7 lux to 194.7 lux) and scene reflectance, and explicitly characterize the distinct static response behaviors of event cameras under varying lighting conditions. This provides empirical support for event-based static reconstruction methods. Specifically, we use a DC light source with precisely adjustable brightness in an optical darkroom to control scene brightness while minimizing the effects of light flicker, and employ an event camera (Prophesee EVK4) to capture grayscale patches of a standard Macbeth ColorChecker. The ambient brightness values are obtained by averaging multiple measurements taken with a lux meter, and the reflectance values correspond to the fixed parameter of ColorChecker. By measuring the average event triggering rate for capturing each grayscale block under different conditions, we map the response trends of the event camera to the static scene, as shown in Fig. 2 (a).

The experimental results indicate a unique mapping between the event counts and the grayscale values within scenes illuminated by a constant light source. For grayscale patches with lower reflectance, as illustrated in Fig. 2 (b), the response curve is monotonically increasing with ambient brightness. Conversely, for higher reflectance levels, depicted in Fig. 2 (c), the response curve reveals a notable inflection point. This demonstrates that the response of event pixels to static scenes follows a specific pattern: when the illuminance received by the sensor is below a threshold, the event count increases with lighting intensity. However, above the threshold, the event count slowly decreases as the intensity rises, which is validated by real-captured data shown in Fig. 2 (d). We also observe that as scene illumination increases, the event count trend stabilizes, resulting in low distinguishability of event count values. These observations not only reveal the rules for event-based static scene reconstruction but also provide guidance for our reconstruction method.

3.2. Convolutional integration for reconstruction

Traditional methods for reconstructing static scenes [10, 11] suffer from low image quality and loss of detail. By statistically analyzing the pixel value distributions of reconstructed images, we identify two primary causes for these issues. First, noise accumulated during the integration process sig-

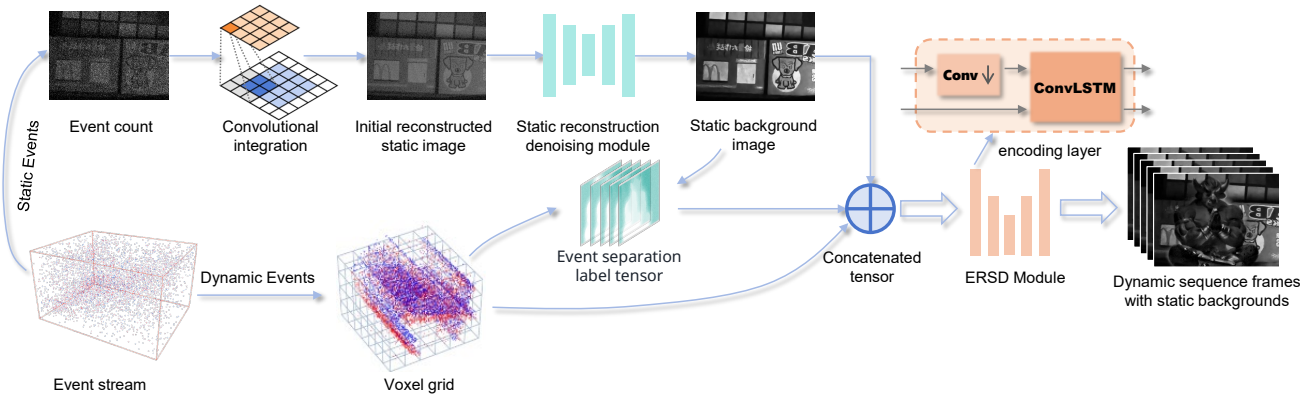

Figure 4. The pipeline of the proposed URSEE framework. The raw event stream, consisting of both motion-triggered dynamic events and background-triggered static events, serves as the input. Dynamic and static events are independently extracted and processed through parallel channels, resulting in the production of a static background frame and a sequence of dynamic voxel grids. By concatenating these with an event separation label tensor, a fused tensor is formed, which is then fed into the ERSD Module to generate a sequence of motion frames characterized by high-quality static backgrounds.

nificantly degrades image quality. Second, prolonged integration leads to event saturation [28], exceeding the sensor's effective response range. This causes pixel values in the reconstructed image to converge to 0 or 255, reducing grayscale images to binary ones, leading to distortion of image contrast and a reduction in information content.

We propose a convolutional integration method to effectively address the two aforementioned challenges. Rather than performing pixel-wise integration like existing methods, we employ a 3×3 convolutional kernel with a mean filter to perform convolutional integration over the pixel plane. Specifically, we first calculate the event count for each pixel within the integration time. Then, after padding the pixel plane, the value of each pixel is defined by the normalized convolution result. Fig. 3 qualitatively and quantitatively compares the existing pixel-wise integration method with the proposed convolution integration method. The statistical analysis of the pixel value distribution in images reconstructed using different methods indicates that our approach demonstrates a higher concentration of mid-range pixel values. This approach mitigates the influence of pixels with zero values and extreme values, effectively alleviating the pixel value polarization and thereby preserving more comprehensive grayscale characteristics.

3.3. Denoising module for quality improvement

To further enhance the reconstructed image quality to a level comparable with frame cameras, we propose a fully convolutional network with a U-Net architecture [27] as a denoising module, termed the "SRD Module" (Static Reconstruction Denoising). Inspired by Chen *et al.* [5], we employ layer normalization within each convolutional block to ensure a more stable training process. Furthermore, we implement a channel attention mechanism to effectively capture and represent the global characteristics of noise, thereby improving the network's robustness and accuracy. The encoder architecture utilizes strided convolutions with a stride of 2, optimizing the downsampling process. Conversely, the decoder blocks employ bilinear upsampling followed by convolutional layers, ensuring high-resolution feature restoration and refined image reconstruction.

Due to the fact that existing event simulators, *e.g.*, V2E [14] and DVS-Voltmeter [17], focus on generating events triggered by dynamic scenes while neglecting events triggered by static backgrounds, the event simulation datasets they generate are inadequate for training models that match the distribution of real-captured static event features. To address this, we use a hybrid imaging system with one traditional RGB camera and one event camera to collect a real-world dataset that includes static event streams along with high-quality frames for model training. A more comprehensive introduction is provided in Sec. 5.1.

4. Static-Dynamic Scene Reconstruction

Previous event-based video reconstruction methods [2, 21, 23, 26] have predominantly focused on moving objects while overlooking static backgrounds. In contrast, recent research on static reconstruction [10, 11] has been dedicated solely to static scenes. Currently, there are no satisfactory approaches for reconstructing videos that encompass both dynamic and static elements. To reconstruct global scenes and thereby expand the applicability of event cameras, the URSEE designs an end-to-end video reconstruction framework capable of reconstructing moving foregrounds while maintaining static background information. Specifically, URSEE demonstrates robust compatibility in the concur-

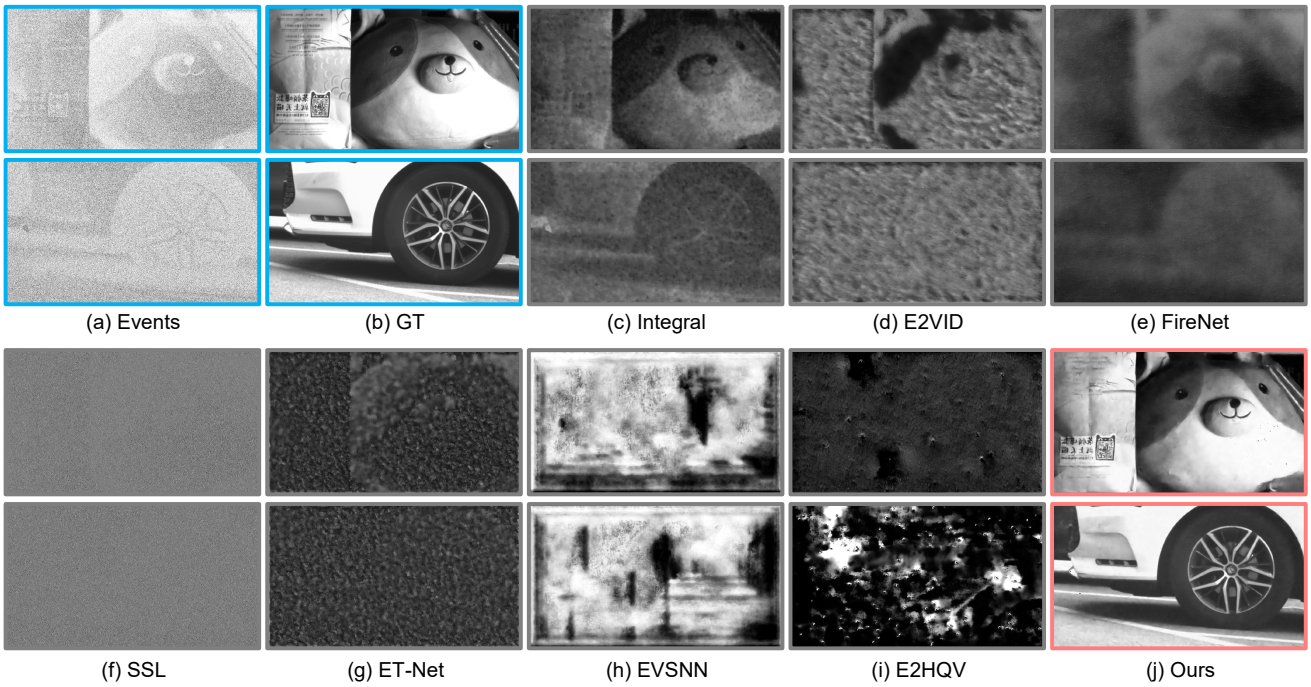

Figure 5. Qualitative comparisons on our E-static dataset for grayscale image reconstruction. (a) Events. (b) Ground Truth. (c)∼(j) Reconstruction images of integral, E2VID [26], FireNet [29], SSL [22], ET-Net [38], EVSNN [41], E2HQV [25], and ours.

Table 1. Quantitative comparison of our method with mainstream approaches for event-based static image reconstruction on our E-static dataset. Arrows ↑ (↓) indicate that higher (lower) values are preferable. Best performance is highlighted in **bold**.

	E2VID [26]	FireNet [29]	SSL [22]	ET-Net [38]	EVSNN [41]	E2HQV [25]	E2VID (retrained) [26]	Ours
PSNR ↑	9.35	7.90	8.78	8.14	7.12	6.49	17.94	**22.43**
SSIM ↑	0.330	0.292	0.039	0.213	0.287	0.130	0.658	**0.860**
LPIPS ↓	0.808	0.935	1.250	0.880	0.797	0.880	0.369	**0.244**

rent processing of both static and dynamic events. It employs distinct and appropriate feature extraction and fusion strategies for static and dynamic event types, ultimately reconstructing videos complete with background details.

Figure 4 presents the pipeline of the URSEE framework. Initially, static and dynamic events are separated from the raw event stream and processed through parallel channels to extract their respective features. For static events, the convolutional integration method and denoising module, as described in Sec. 3.2 and Sec. 3.3, are utilized to reconstruct the static background with high fidelity. Dynamic events are transformed into voxel grids [40] to retain their spatiotemporal information. The reconstructed background and dynamic voxel grids are subsequently concatenated for feature fusion, incorporating an event separation label mechanism to accurately identify the input source for enhanced feature extraction. At the core of the URSEE framework is the proposed ERSD Module (Event-based Reconstruction Network with Static and Dynamic Elements), a fully convolutional network that includes ConvLSTM modules [31]. This network takes the fused feature tensor as input and performs supervised learning to reconstruct the target video from the event data.

Section 4.1 details the proposed spatiotemporal window-based method for separating static and dynamic events, and Sect. 4.2 elaborates on the ERSD network's architecture.

4.1. Event separation

We propose a spatiotemporal window-based method for separating static and dynamic events from a raw event stream. A window, with a spatial scale of 20×20 pixels and a temporal scale of $10ms$, slides non-overlappingly across the pixel plane, recording the spatiotemporal information of each event. Within each window, a global threshold is used to distinguish between static and dynamic events. Events within a window are classified as dynamic if their count ex-

ceeds the threshold, representing a large number of events triggered in this region at the moment. Otherwise, they are classified as static background events.

For static events, we employ the reconstruction strategy outlined in Sec. 3.2 to generate a static background image. For dynamic events, we transform them into voxel grids [42] to maintain their spatiotemporal characteristics. To enable the network to learn both static and dynamic features concurrently, we concatenate these into a unified tensor for input. During this integration, an event separation label tensor is applied to label the origin of each feature, thereby aiding the network in more precise feature extraction.

4.2. ERSD module architecture

The ERSD module is the core of the URSEE framework, designed as a convolutional neural network that takes fused tensors as input and outputs a sequence of frames with static backgrounds and moving foregrounds. Its purpose is to extract dynamic event features from voxel grids and static event features from reconstructed frames through supervised learning, thereby reconstructing both simultaneously.

Our neural network is a recurrent, fully convolutional network employing a U-Net architecture [27], widely adopted for image reconstruction tasks. Its primary structure comprises an introduction layer, followed by N_E recurrent encoding layers, N_M intermediate layers, N_D decoding layers, and a final image prediction layer. As our input is a concatenated tensor encompassing a grayscale image, a voxel grid, and an event separation label tensor, the prediction layer has $B_{in}+2$ input channels, where B_{in} represents the number of channels in the custom voxel grid, set to 5 in this work. Inspired by Chen *et al*. [5], we design a custom convolutional block, incorporating layer normalization and a channel attention mechanism to stabilize training and extract global features. Each encoding layer consists of a custom convolutional block followed by a ConvLSTM [31], while the intermediate and decoding layers consist solely of custom convolutional blocks.

5. Experiment

Since this paper encompasses both event-based static image and dynamic video reconstruction. For the static image reconstruction task, we compare our proposed convolutional integration method and SRD-Module against mainstream methods on our E-Static dataset, including pixel-wise integration methods [11], E2VID [26], FireNet [29], SSL [22], ET-Net [38], EVSNN [41], and E2HQV [25]. For the video reconstruction task, we compare our proposed method with the above methods on our E-StaDyn dataset.

We employ PSNR, SSIM, and LPIPS metrics to evaluate the quality of the reconstructed images. Notably, we retrained E2VID [26] on the proposed datasets for comparative analysis to verify the incompatibility of mainstream reconstruction methods with static events and to further demonstrate the effectiveness of our approach.

In this section, we introduce the training and evaluation datasets in Sec. 5.1, the training process in Sec. 5.2, the comparison for static image reconstruction in Sec. 5.3, and the comparison for video reconstruction in Sec. 5.4.

5.1. Training and evaluation dataset

Event-based static scene reconstruction is an emerging field with limited research and datasets. Therefore, we introduce two new datasets, E-Static and E-StaDyn, for model training and method evaluation.

E-Static is a real-world dataset captured from diverse indoor and outdoor static scenes using a hybrid event-frame camera system, which consists of a traditional RGB camera (Alvium1800 U-240c) and an event camera (Sony IMX636). A beam splitter with a light ratio of 1: 9 (event: frame) is used to align their fields of view, avoiding excessively bright illumination on event pixels that could make it difficult to reconstruct clear images from static events. Based on empirical evaluation, we set the optimal ON and OFF thresholds of the event camera to be -17 and -50, respectively, ensuring the generation of sufficient static events for reconstruction. It comprises 200 sets of raw event streams alongside corresponding high-quality ground truth frames. For training, the original 1280×720 images are divided into six 512×512 sub-images, yielding 1290 training and 96 testing samples.

E-StaDyn is a synthetic dataset comprising 130 distinct scenes, each characterized by a unique static background and dynamic foreground. These scenes are generated by pairing diverse high-quality images with different 3D models exhibiting randomized motion. Each scene is rendered into 600 consecutive frames using Blender software to serve as ground truth data. These frames are then processed through the DVS-Voltmeter [17] simulator to generate synthetic event streams. The dataset is divided into 115 scenes for training and 15 scenes for testing.

5.2. Training procedure

In this study, we train two neural networks, the SRD module, and the ERSD module, for distinct reconstruction tasks. During training, these modules utilize different datasets and data augmentation techniques, which we will describe in detail. Both of them are trained on an NVIDIA RTX 3090 GPU using PyTorch [24] and employ the ADAM optimizer [15] with a learning rate of 0.0001.

The SRD module is designed to leverage supervised learning to extract noise features from reconstructed images, thereby enhancing the quality of images reconstructed through convolution integration. We utilize our E-Static dataset for training and testing the model, augmenting the data with random two-dimensional rotations, horizontal and

Figure 6. Qualitative comparisons on our E-StaDyn synthetic dataset for dynamic video reconstruction. Upper: E2VID [26]. Middle: Our URSEE framework. Bottom: Ground truth. E2VID can reconstruct portions of the static background because the movement of the foreground across the background alters the surface light intensity, thereby triggering key events used for reconstruction.

Table 2. Quantitative comparison of our method with mainstream approaches for event-based dynamic video reconstruction on our E-StaDyn dataset. Arrows ↑ (↓) indicate that higher (lower) values are preferable. Best performance is highlighted in **bold**.

	E2VID [26]	FireNet [29]	SSL [23]	ET-Net [38]	EVSNN [41]	E2HQV [25]	E2VID [26] (retrain)	Ours
PSNR ↑	14.01	9.81	9.69	13.67	8.04	9.71	18.45	**31.97**
SSIM ↑	0.725	0.607	0.599	0.701	0.453	0.532	0.792	**0.959**
LPIPS ↓	0.458	0.574	0.609	0.489	0.673	0.595	0.448	**0.122**

vertical flips, and random cropping (crop size 256×256). The network employs the Mean Squared Error (MSE) as the loss function to ensure the quality of the reconstructed images. After approximately 8 hours of training over nearly 500 epochs, the model converges, significantly improving the quality of the initially reconstructed frames with an enhancement of 11 dB.

The ERSD module employs the MSE as the loss function to ensure the quality of the reconstructed images. During training, each scene from our E-StaDyn dataset is divided into groups of sequence frames with a length of 40. To enhance the data, operations such as adding hot pixels to sequences, adding noise to voxel grids, and introducing random pause mechanisms are applied. Since each encoding layer includes a ConvLSTM module, the low quality of the initial frame can significantly disrupt the reconstruction of subsequent frames. Therefore, we discard the first and the last sequences of each data group for improved training.

5.3. Comparison of static image reconstruction

The qualitative comparison of our proposed two-stage static reconstruction strategy with mainstream reconstruction methods for static image reconstruction is presented in Fig. 5, while the quantitative comparison results are shown in Table 1. The results indicate that our method successfully reconstructs static scenes, achieving results comparable to frame cameras, whereas the other methods fail to extract static background information, demonstrating their incom-

Figure 7. Qualitative comparisons on real-world data for dynamic video reconstruction. Upper: E2VID [26] (pretrained). Middle: E2VID [26] (retrained), retrained on our E-StaDyn dataset for 100 epochs followed by inference. Bottom: Our URSEE framework.

patibility with static events. Additionally, we show the color reconstruction results of static scenes using our method in the supplementary material.

5.4. Comparison of dynamic video reconstruction

The qualitative comparison of our proposed URSEE framework with mainstream reconstruction methods on our E-StaDyn synthetic dataset is depicted in Fig. 6, with the comparison on real data illustrated in Fig. 7. Quantitative comparison results are summarized in Table 2. E2VID can partially reconstruct backgrounds because moving objects alter light intensity as they pass the background, triggering certain feature events that can be used for background reconstruction. However, this is dependent on specific global motion patterns and lighting conditions, which limits its applicability and reconstruction performance, as reflected by detail loss and significant distortion shown in Fig. 6. In contrast, our method achieves independent reconstruction of high-fidelity static backgrounds, providing the proposed framework with robust performance on both synthetic and real data. Even when retrained on the proposed dataset, E2VID fails to reliably reconstruct backgrounds, underscoring the model's incompatibility with static events.

6. Conclusion

In this paper, we propose the URSEE framework to address the challenges of noise impact and event integration saturation in existing event-based static video reconstruction methods. We introduce a convolutional integration method to mitigate the issue of pixel value polarization and propose a fully convolutional neural network to learn static event noise characteristics from the initially reconstructed images. URSEE enables end-to-end reconstruction of motion videos with static backgrounds from events for the first time. The reconstruction results demonstrate that our method effectively captures the key features of both static and dynamic events.

Limitation The proposed URSEE framework performs well under normal lighting conditions, but when the illumination increases, the quality of static scene reconstruction encounters a bottleneck, which is limited by the design of the event sensor. Our next step is to enhance the method to address a broader range of more complex scenes. Additionally, this study employs the Prophesee EVK4 device for scene capturing. For other devices, it is necessary to relearn their response models to static scenes and noise characteristics. This requires further expansion of the existing dataset and improvements in model generalization.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant No. 62402014, 62136001, 62088102), Beijing Natural Science Foundation (Grant No. L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012). Peiqi Duan was also supported by China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076). PKU-affiliated authors thank openbayes.com for providing computing resource.

References

[1] Yuhan Bao, Lei Sun, Yuqin Ma, and Kaiwei Wang. Temporal-mapping photography for event cameras. In *Proc. of European Conference on Computer Vision*, 2024. 2

[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. 1, 2, 4

[3] Ruiming Cao, Dekel Galor, Amit kohli, Jacob L Yates, and Laura Waller. Noise2Image: Noise-enabled static scene recovery for event cameras. *ArXiv*, abs/2404.01298, 2024. 2, 3

[4] Ruiming Cao, Dekel Galor, Amit Kohli, Jacob L Yates, and Laura Waller. Noise2Image: noise-enabled static scene recovery for event cameras. *Optica*, 12(1):46–55, 2025. 2, 3

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proc. of European Conference on Computer Vision*, pages 17–33. Springer, 2022. 4, 6

[6] Peiqi Duan, Zihao Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. EventZoom: Learning to denoise and super resolve neuromorphic events. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 1

[7] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Yi Ma, and Boxin Shi. EventAid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. 2023. 2

[8] Peiqi Duan, Yi Ma, Xinyu Zhou, Xinyu Shi, Zihao W. Wang, Tiejun Huang, and Boxin Shi. NeuroZoom: Denoising and super resolving neuromorphic events and spikes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2023. 1

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. 1

[10] Dekel Galor, Ruiming Cao, Laura Waller, and Jacob Yates. Leveraging noise statistics in event cameras for imaging static scenes. In *International Conference on Computational Photography Posters*, 2023. 2, 3, 4

[11] Qiyao Gao, Xiaoyang Sun, Zhitao Yu, and Xu Chen. Understanding and controlling the sensitivity of event cameras in responding to static objects. In *Proc. of IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 783–786, 2023. 2, 3, 4, 6

[12] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLT: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, pages 1–18, 2019. 1

[13] Jin Han, Yuta Asano, Boxin Shi, Yinqiang Zheng, and Imari Sato. High-fidelity event-radiance recovery via transient event frequency. In *Proc. of Computer Vision and Pattern Recognition*, 2023. 1, 2

[14] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2E: From video frames to realistic dvs events. In *Proc. of Computer Vision and Pattern Recognition Workshops*, 2021. 4

[15] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2), 2008. 1

[17] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Proc. of European Conference on Computer Vision*, pages 578–593. Springer, 2022. 4, 6

[18] Yi Ma, Peiqi Duan, Yuchen Hong, Chu Zhou, Yu Zhang, Jimmy S. Ren, and Boxin Shi. Color4E: Event demosaicing for full-color event guided image deblurring. In *ACM Multimedia*, 2024. 2

[19] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1–9. IEEE/RSJ, 2018. 1

[20] Mohammad Mostafavi, Yeongwoo Nam, Jonghyun Choi, and Kuk-Jin Yoon. E2SRI: Learning to super-resolve intensity images from events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6890–6909, 2022. 1

[21] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2519–2533, 2020. 4

[22] 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 *Proc. of Computer Vision and Pattern Recognition*, pages 3446–3455, 2021. 1, 5, 6

[23] Federico Paredes-Vallés and Guido C. H. E. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 4, 7

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 6

[25] Qiang Qu, Yiran Shen, Xiaoming Chen, Yuk Ying Chung, and Tongliang Liu. E2HQV: High-quality video generation from event camera via theory-inspired model-aided deep learning. 2024. 5, 6, 7

[26] 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*, 43(6):1964–1980, 2019. 1, 2, 4, 5, 6, 7, 8

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 4, 6

[28] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Proc. of Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. 4

[29] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. pages 156–163, 2020. 2, 5, 6, 7

[30] Richard Shaw, Sibi Catley-Chandar, Ales Leonardis, and Eduardo Pérez-Pellitero. HDR reconstruction from bracketed exposures and events. 2022. 2

[31] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 5, 6

[32] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. *Proc. of European Conference on Computer Vision*, 2020. 2

[33] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 1

[34] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based video frame interpolation. In *Proc. of Computer Vision and Pattern Recognition*, 2021. 2

[35] Lin Wang, 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 *Proc. of Computer Vision and Pattern Recognition*, pages 10081–10090, 2019. 2

[36] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. VisEvent: Reliable object tracking via collaboration of frame and event flows. *ArXiv*, abs/2108.05015, 2021. 1

[37] Zihao Winston Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proc. of Computer Vision and Pattern Recognition*, 2020. 1

[38] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. of International Conference on Computer Vision*, 2021. 5, 6, 7

[39] Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. EvUnroll: Neuromorphic events based rolling shutter image correction. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 1

[40] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proc. of Computer Vision and Pattern Recognition*, pages 989–997, 2019. 5

[41] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proc. of Computer Vision and Pattern Recognition*, 2022. 5, 6, 7

[42] 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) Workshops*, pages 0–0, 2018. 6