The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4-2024
ISPRS TC IV Mid-term Symposium “Spatial Information to Empower the Metaverse”, 22–25 October 2024, Fremantle, Perth, Australia

How Many Events are Needed for One Reconstructed Image Using an Event Camera?

Tingting Lei[1, 2], Xueli Guo[1], You Li[1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China – (leitingting, guoxueli, liyou)@whu.edu.cn
[2] TUM School of Engineering and Design, Technical University of Munich, Munich, Germany

Keywords: Event Camera, Event-based Vision, Image Reconstruction.

Abstract

Event cameras offer significant advantages over traditional cameras, including high temporal resolution, high dynamic range, and low power consumption. However, reconstructing images from the asynchronous events generated by these cameras presents unique challenges. This paper investigates the optimal number of events needed for high-quality image reconstruction using event cameras. We evaluate two primary reconstruction strategies—fixed time window and fixed number of events—across various dynamic and static scenes. Our study includes scenarios with different lighting conditions and camera movements. Using the state-of-the-art E2VID algorithm, we perform both qualitative and quantitative analyses of the reconstructed images, comparing them with reference frames from a traditional RGB camera. Our results demonstrate the trade-offs between temporal resolution and image quality for each reconstruction strategy, providing insights into the optimal settings for different applications. This research offers practical guidelines for selecting appropriate reconstruction parameters to achieve the better image quality from event cameras.

1. Introduction

The event camera is an optical sensor that generates images based on changes in brightness rather than uniform exposure at a fixed frame rate. Each pixel asynchronously generates a series of events. Compared to traditional cameras, event cameras offer advantages such as high temporal resolution (approximately 1 microsecond), high dynamic range (over 140 dB), no motion blur, lower data redundancy, and lower power consumption (Gehrig et al., 2020). This makes event cameras a new option for robotic applications, serving as auxiliary sensors in scenarios where traditional cameras face challenges, and they hold significant potential in many application domains such as feature detection, tracking, and visual simultaneous localization and mapping (SLAM).

A widely used approach to processing event data involves converting the data into lower frequency frames, such as reconstructed images or videos. This serves as an interface between event cameras and traditional frame-based computer vision (Gallego et al., 2017). While this method may result in the loss of some advantages, such as ultra-high frequency data, it also offers clear benefits. Image reconstruction can provide humans with an intuitive understanding of the rich information encoded by events, enabling people to visually interpret events and gain an intuitive understanding of information embedded in the event data. Reconstructed images can also serve as a useful representation for traditional frame-based computer vision. Additionally, existing visual localization algorithms can be directly applied to event data when events are reconstructed for intensity frames.

Although event cameras generate asynchronous events, an individual event only marks the brightness change of a certain pixel, thus one event alone does not provide sufficient information for estimation. Therefore, past events or additional information are required. When reconstructing images from event data, each intensity image is reconstructed from a certain number of event data. However, this approach presents a problem: how many events are needed for one reconstructed image using an event camera? The number of events selected for image reconstruction is crucial for event cameras. If too few events are used, the image reconstruction may fail due to insufficient information; if too many events are used, ghosting may appear in the reconstructed image due to overlapping events with different timestamps. For example, the background objects vary in different environments, thus the event data volume changes in various scenarios. Besides, the complexity of camera motions can lead to the change in event data volume, resulting in image blurring, large feature extraction errors, and dispersion of localization results. Therefore, it is important to strike a balance and select the appropriate number of events for accurate image reconstruction.

This paper aims to investigate the effect of event data volume on image reconstruction, with respect to various environments as well as various camera motions. Firstly, event data in various scenarios is collected, including different environments and camera movements. These scenarios included indoor and outdoor environments, event camera movements at different speeds, and straight and turning movements of the camera. Then, different image reconstruction strategies are used to process the event data, including reconstruction with a fixed number of events and with a fixed time window. Finally, the reconstructed images under different reconstruction strategies are compared, and the relationship between image reconstruction and different reconstruction strategies in different scenarios is analysed.

The contributions of this article are as follows: (1) Collect event camera data in various scenarios, and intuitively compare the performance of event camera image reconstruction methods in various scenarios. (2) Use different image reconstruction strategies for event camera data in different scenarios, and comprehensively compare the differences between different reconstruction strategies. (3) Establish the connection between the scene and the image reconstruction strategy, and provide some guiding suggestions for the setting of image reconstruction parameters. Conclusions from this paper can pave the way for optimized event camera reconstruction strategies across varied scenarios.

2. Related Work

In recent years, reconstruction models based on deep neural networks have shown remarkable performance in processing event camera data. These models aim to reconstruct high-quality intensity images or videos from the sparse and asynchronous events generated by event cameras. The development of these models addresses the need for bridging the gap between the unique data structure of event cameras and traditional frame-based computer vision methods.

Munda et al. (Munda et al., 2018) presented a method for intensity reconstruction by framing it as an energy minimization problem. Their approach allows for image reconstruction at arbitrary frame rates, but their experiments were conducted in static environments, lacking significant camera motion changes, which limits their applicability in dynamic real-world scenarios. Barua et al. (Barua et al., 2016) used K-SVD to map small patches of events to an image gradient and applied Poisson integration to reconstruct intensity images. This method performs well in static scenes but struggles with dynamic scenes, as it does not account for rapid changes in the environment or the camera's motion.

These two works reconstruct independent intensity images from small windows of events, while Rebecq et al. (Rebecq et al., 2019a, Rebecq et al., 2019b) introduced E2VID, a recurrent neural network designed to handle the high-speed and high dynamic range nature of event streams. E2VID reconstructs high-quality videos from long event streams, capturing temporal dependencies in the data, which is crucial for reconstructing accurate intensity frames from events. Similarly, Scheerlinck et al. (Scheerlinck et al., 2020) developed FireNet, a fully convolutional network that performs fast video reconstruction from events. FireNet is optimized for efficiency, requiring fewer parameters and less memory compared to E2VID, making it suitable for real-time applications.

Later, Wang et al. (Wang et al., 2020) proposed an unsupervised pipeline that first reconstructs low-resolution images from event streams and then enhances the image quality through super-resolution techniques. Their method effectively upsamples the enhanced images, providing high-quality reconstructions without relying on ground-truth data for training. Besides, Ercan et al. (Ercan et al., 2024) introduced HyperE2VID, which utilized hypernetworks to combine current events with previously reconstructed images. This approach improves the reconstruction quality by incorporating temporal information from past frames, thus maintaining consistency in the reconstructed video. For high dynamic range (HDR) video reconstruction, Zou et al. (Zou et al., 2021) proposed a convolutional recurrent neural network, while Yang et al. (Yang et al., 2023) introduced a multimodal learning framework that combines low dynamic range videos and event data.

However, these methods often rely on ground-truth data from conventional cameras, which may not accurately capture HDR scenarios. Paredes-Vallés et al. (Paredes-Vallés and De Croon, 2021) took a different approach by using self-supervised learning to reconstruct intensity images from events. Their method estimates optical flow simultaneously with intensity reconstruction, eliminating the need for ground-truth data and making the system more adaptable to various scenarios.

Despite these advances, challenges remain in optimizing reconstruction algorithms for diverse environments and dynamic camera motions. This paper builds upon existing research by evaluating different reconstruction strategies across a variety of scenarios, aiming to identify optimal parameters for high-quality image reconstruction from event data.

3. Image Reconstruction Methodology

The process of reconstructing images from event data is critical for utilizing the high temporal resolution and dynamic range offered by event cameras. Unlike traditional cameras, which capture frames at fixed intervals, event cameras generate a continuous stream of events representing pixel-level brightness changes. This asynchronous nature of event data necessitates specialized strategies and algorithms for effective image reconstruction.

Image reconstruction from event data can be broadly categorized into strategies based on the integrated process of events over fixed durations or fixed quantities. These strategies aim to balance the trade-off between temporal resolution and the quality of the reconstructed images. Fixed-duration strategies ensure consistent frame rates but can lead to variable image quality depending on scene dynamics. Conversely, fixed-number strategies maintain a consistent number of events per frame but result in variable frame rates.

This section looks into the fundamental aspects of event representation, outlines the prominent reconstruction strategies, and introduces a prominent reconstruction algorithm, E2VID, which is used to reconstruct images from events in this paper. The subsequent sections provide a detailed examination of these methodologies, highlighting their implementation and effectiveness in various scenarios.

3.1 Event Representation

An event camera detects changes in brightness at the pixel level and triggers asynchronous events when the brightness surpasses a predefined threshold. Given an event sequence $\mathbf{E} = \{\mathbf{e}_i\}$ with a total time duration of T and comprising a total of N events, where W and H represent the width and height of the event camera, respectively, each event can be defined as:

$$\mathbf{e}_i = (x_i, y_i, t_i, p_i) \tag{1}$$

where (x_i, y_i) represents the pixel position, t_i represents the timestamp of the event, and p_i is a polarity flag that signifies whether the brightness increases or decreases. Here, $i \in [0, N-1], x_i \in \{0, 1, \dots, W-1\}, y_i \in \{0, 1, \dots, H-1\}, t_i \in [0, T], p_i \in \{-1, 1\}$.

3.2 Reconstruction Strategy

The current event-to-image reconstruction strategy is mainly based on event-batching algorithms, which integrate a fixed time interval or a fixed number of events to reconstruct one intensity image. The basic concept is to divide the continuous stream of events into consecutive windows. All the events within each window will be used to reconstruct an intensity frame. Two primary strategies are employed: fixed-duration and fixed-number reconstruction.

3.2.1 Fix-duration Reconstruction Strategy In this strategy, events are accumulated over a fixed time interval to reconstruct an intensity image. Formally, the events within each time window are:

$$I_{t_k} \leftarrow E_{t_k} = \{e_i | kT \leq t_i \leq (k+1)T\}, k \in \{0, 1, \dots\} \tag{2}$$

where T denotes the fixed duration. This method ensures a consistent frame rate, but the number of events in each reconstructed image can vary significantly depending on the scene's dynamics, as shown in Figure. 1.

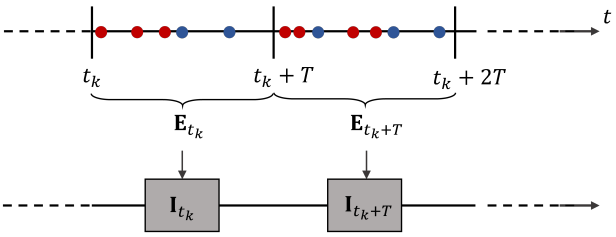

Figure 1. Fixed-duration reconstruction strategy.

3.2.2 Fix-number Reconstruction Strategy In this approach, a fixed number of events are used to reconstruct each intensity image. The events are accumulated as:

$$I_j \leftarrow E_j = \{e_i | kN \leq i \leq (k+1)N-1\}, k \in \{0, 1, \dots\} \quad (3)$$

where N represents the fixed number of events. This method maintains a consistent number of events per frame, but the frame rate may vary depending on the event generation rate, as shown in Figure. 2.

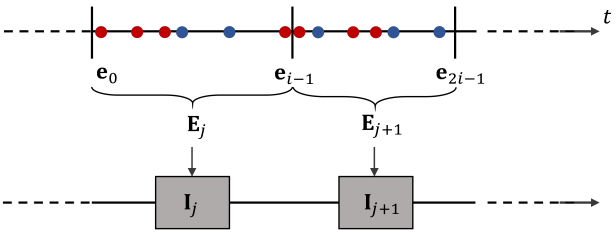

Figure 2. Fixed-number reconstruction strategy.

3.3 Reconstruction Algorithm

Among the various reconstruction algorithms, E2VID has laid the groundwork and shown superior performance in generating high-quality images from event data. Developed by Rebecq et al. (Rebecq et al., 2019a), E2VID uses a recurrent neural network (RNN) to process streams of event data and produce intensity frames with high temporal resolution.

E2VID processes the event data through several stages: Firstly, events are grouped into batches, ensuring a manageable and consistent data input size. Then the network employs convolutional layers to extract spatiotemporal features from the accumulated events. A Long Short-Term Memory (LSTM) module processes the extracted features, maintaining temporal context and effectively handling the asynchronous nature of the events. This recurrent layer ensures that the temporal dependencies between events are preserved. Next residual connections are integrated into the network to facilitate the flow of information and gradients, aiding in the learning process and improving the network's ability to reconstruct fine details. After recurrent processing, the network uses upsampling techniques to increase the spatial resolution of the feature maps. This step ensures that the final reconstructed image has the desired resolution and quality. The final stage involves additional convolutional layers that transform the upsampled feature maps into an intensity image. These layers refine the features and produce high-quality frames that accurately capture the underlying scene dynamics. Figure. 3 displays the architecture of the E2VID reconstruction algorithm.

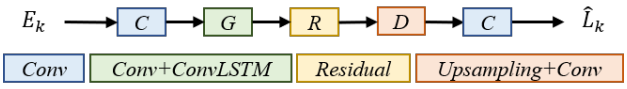

Figure 3. Architecture of the E2VID reconstruction algorithm.

This paper uses E2VID as an example to process events, aiming to provide a comprehensive analysis of image reconstruction using an event camera across diverse scenarios.

4. Evaluation

In this section, we compare reconstructed images using different reconstruction strategies across various scenarios, and present both quantitative and qualitative results. Our evaluation relies on authentic event data captured by a DVXplorer sensor with a resolution of 640×480. To establish a benchmark, an RGB camera, specifically the Intel RealSense, is used to provide ground-truth frames at a rate of 30 frames per second (fps). Both the event camera and the RGB camera are mounted on a wheeled robot, as illustrated in Figure. 4.

Figure 4. Setup of event camera and RGB camera mounted on a wheeled robot.

These scenarios include diverse lighting conditions, ranging from bright indoor to dim indoor and outdoor environments. Additionally, the scenarios include in camera motion speed, spanning from slow to swift movements. Moreover, the scenarios feature challenges such as high dynamic range situations, where the camera faces direct sunlight, as well as environments with many moving objects in the background.

For image reconstruction strategies, we employ two reconstruction principles to process the event data: reconstruction with a fixed time window and reconstruction with a fixed number of events. In terms of reconstruction with a fixed number of events, we use large, medium, and small event volumes to reconstruct one intensity image; in terms of reconstruction with a fixed time window, we use long, medium, and short time windows to reconstruct one intensity image. These reconstruction strategies will be applied to all the scenarios mentioned earlier, and the reconstruction results will be compared and analyzed.

For each reconstructed image, we query the corresponding ground-truth frame with the closest timestamp to the reconstructed image, and then compare the similarity of the two frames according to several quality metrics. Prior to the comparison, we employ local histogram equalization to both the ground-truth and

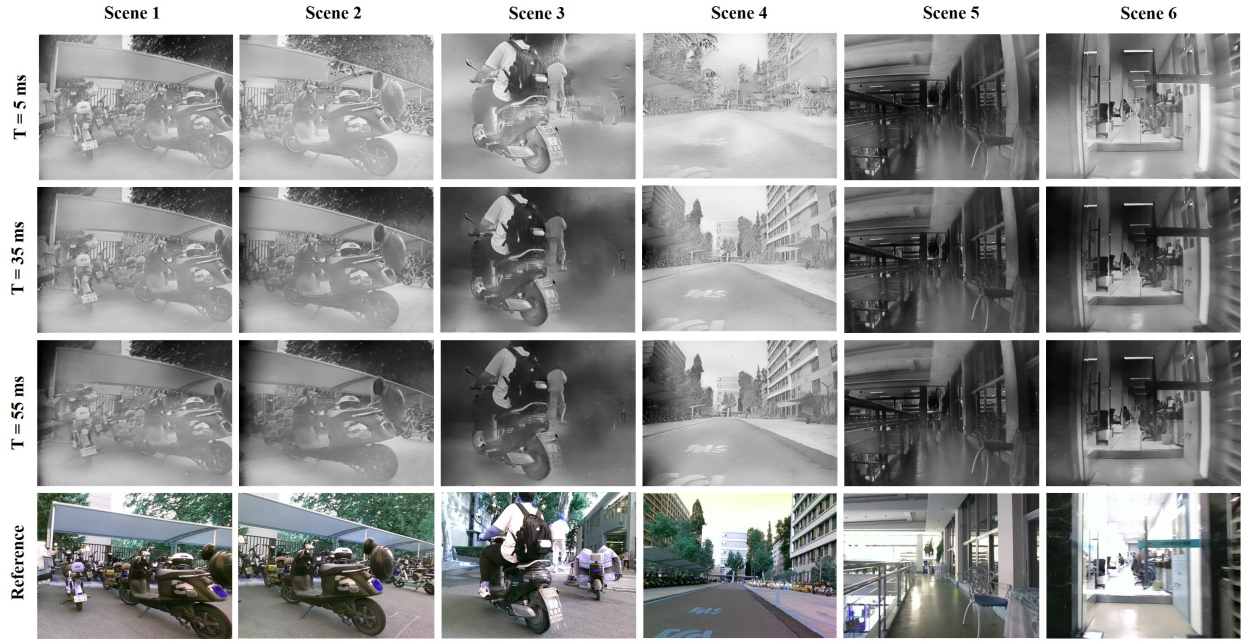

Figure 5. Reconstruction results using fixed-duration strategy across different scenes.

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6
T = 5 ms	0.58	0.63	0.58	0.59	0.66	0.50
T = 30 ms	0.55	0.58	0.67	0.59	0.61	0.41
T = 55 ms	0.56	0.66	0.67	0.59	0.61	0.41

Table 1. pHash similarity results for fixed-duration reconstruction.Low value represents higher similarity.

reconstructed frames, ensuring that both grayscale images are standardized to the same intensity range and can be compared.

To compare the reconstructed images using different reconstruction strategies, we use a quantitative metric that is widely used in this field as well as the reference images obtained from an RGB camera. The primary evaluation of the reconstruction strategies relies on the similarity between the reconstructed and reference frames. The metric is pHash similarity (Fridrich and Goljan, 2000), a perceptual hashing method that quantifies the similarity between images. Lower pHash values indicate higher similarity.

4.1 Reconstruction with a Fixed Time Window

The fixed-duration reconstruction strategies are evaluated at 5 ms, 30 ms, and 55 ms intervals, respectively. The six scenes include making a turn, moving objects, slow motion, indoor dim lighting with a hand-held camera, and indoor bright lighting with a hand-held camera.

Figure. 5 displays the qualitative results, and Table. 1 displays pHash results. Low pHash value represents higher similarity.

In the first two scenes, the camera is making a turn, introducing significant motion blur and dynamic changes in the field of view. For the fixed-duration strategy, shorter duration of time windows achieves higher similarity, effectively capturing rapid changes without significant overlap of events. The longest duration (55 ms) sees a decrease in similarity, likely due to the compounded motion blur.

The third scene involves significant motion within the frame, such as people walking or vehicles moving. For short time window, the short accumulation period captures the movement crisply, but may lack sufficient context for complex scenes. For medium time window, the balance between context and detail is optimal, capturing movements effectively without excessive blurring. For long time window, the increased time window introduces more blurring, especially noticeable in fast-moving objects.

The fourth scene features slow and steady movements. All time windows (5 ms, 30 ms, 55 ms) perform relatively well due to the slow motion, but the 5 ms window might result in too sparse data, creating false ghost. The 30 ms window seems to offer the best balance, providing enough events to create a smooth and detailed reconstruction. The 55 ms window performs similarly and might slightly blur very slow movements.

The last two scenes involve indoor lighting and hand-held cameras. Low light conditions often pose challenges for traditional cameras due to noise and poor contrast.For short time windows, the short exposure time limits the amount of data collected, leading to potentially noisy reconstructions. For longer time windows, there is a significant improvement as more events are captured, enhancing the image quality. However, a problem arises when the reference images themselves face certain challenges, such as high dynamic range in this case, reducing their reference value.

4.2 Reconstruction with a Fixed Number of Events

The fixed-number reconstruction strategies are evaluated using 50000, 100000, and 200000 event volumes, respectively.

Figure. 6 displays the qualitative results, and Table. 2 displays pHash results. Low pHash value represents higher similarity.

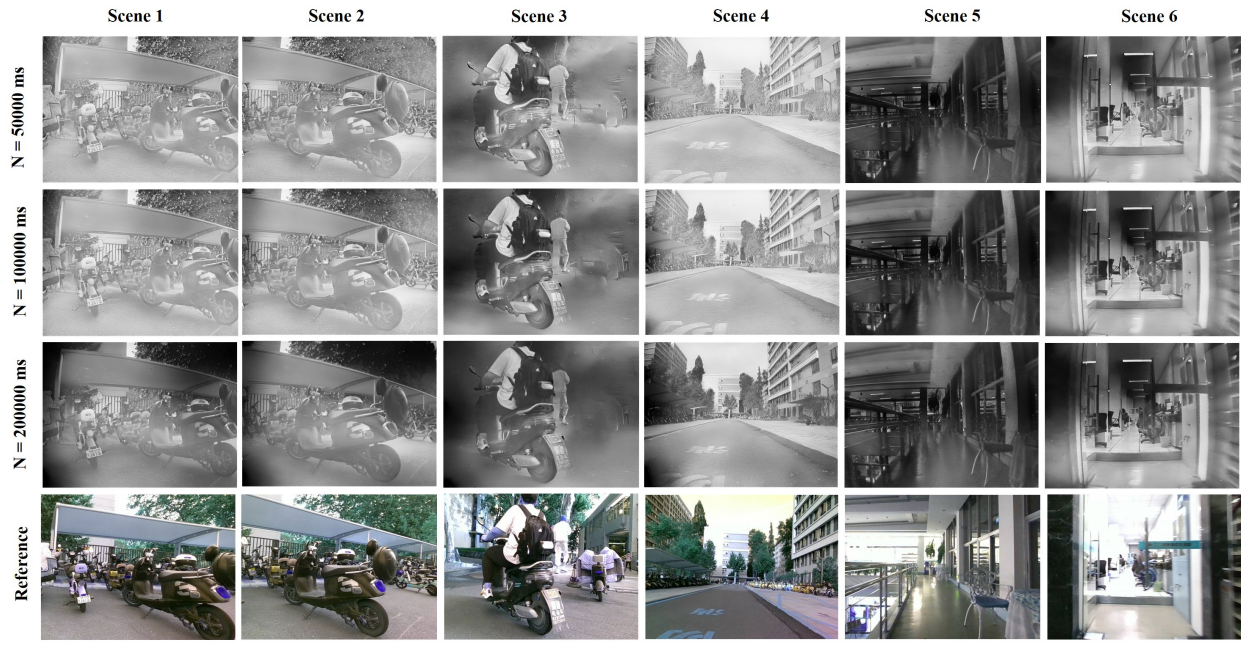

Figure 6. Reconstruction results using fixed-number strategy across different scenes.

	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6
$N = 50000$	0.61	0.61	0.67	0.64	0.60	0.53
$N = 100000$	0.64	0.58	0.66	0.63	0.61	0.48
$N = 200000$	0.55	0.64	0.67	0.61	0.60	0.47

Table 2. pHash similarity results for fixed-number reconstruction. Low value represents higher similarity.

For Scene 1, the best reconstruction quality (lowest pHash value) is observed with 200,000 events, indicating that a higher number of events leads to more accurate reconstruction during the turn. While in Scene 2, using 100,000 events results in the lowest pHash value, suggesting it provides the best reconstruction quality for this particular turning scenario. The two scenes both involve cameras making a turn. The difference between Scene 1 and Scene 2 highlights the variability in the amount of data required for similar types of motion.

Scene 3 shows minimal variation in pHash values across different event numbers, indicating that the reconstruction quality is relatively stable regardless of the number of events. This could imply that moving objects might be sufficiently captured with a moderate amount of event data.

For slow motion in Scene 4, the quality of reconstruction improves slightly as the number of events increases, with the best results at 200,000 events. Slow motion might benefit from more events to better capture finer details and gradual changes.

Scene 5 demonstrates that the pHash values are quite close across all event numbers, suggesting that dim lighting conditions might not significantly benefit from more events. The similarity in pHash values implies that the reconstruction is relatively unaffected by the number of events in low-light conditions. While in Scene 6, the pHash values show improvement with an increasing number of events, with the lowest value at 200,000 events. This indicates that bright lighting conditions benefit from more events to achieve higher reconstruction quality, likely due to the greater contrast and details that can be captured.

Overall, the analysis shows that the optimal number of events for image reconstruction varies with the scene and motion dynamics. Fast motions, such as turning, generally benefit from more events, while scenes with moderate changes, like moving objects or slow motion, show stable reconstruction quality across different event numbers. Lighting conditions also play a significant role, with bright conditions benefiting more from higher event counts compared to dim lighting scenarios. These insights can guide the selection of appropriate event volumes for different scenarios to achieve the best image reconstruction quality.

4.3 Comparison and Analysis

It is clear that the optimal reconstruction strategy is highly dependent on the specific scene and the nature of the motion. Fixed time window strategies tend to perform better in scenarios with consistent motion or lighting conditions, providing a reliable temporal resolution. However, fixed number of events strategies offer better adaptability to varying scene complexities and motion speeds, ensuring that sufficient data is always captured to reconstruct high-quality images.

In scenes with rapid motion, shorter time windows or smaller event volumes help in maintaining sharpness and reducing blurring. In contrast, scenes with slower motion or more complex lighting conditions benefit from longer time windows or higher event volumes to capture more contextual information, enhancing the detail and quality of the reconstructed images.

Overall, the choice of reconstruction strategy should consider the specific characteristics of the scene and the intended application. Fixed-duration strategies are more consistent in timing but can vary in event density, while fixed-number strategies provide consistent event density but variable timing. Under-

standing the trade-offs between these approaches is crucial for optimizing image reconstruction from event cameras.

5. Conclusion

This paper investigates the optimal number of events needed for high-quality image reconstruction using event cameras. We evaluated two primary reconstruction strategies—fixed-duration and fixed-number—across various dynamic and static scenes using the E2VID algorithm. The results show how different reconstruction strategies affect the quality of the reconstructed images, providing insights into selecting appropriate parameters for specific scenarios. Our findings demonstrate the trade-offs between temporal resolution and image quality for each strategy, providing insights into the optimal settings for different applications.

The conclusions from this study offer practical guidelines for selecting appropriate reconstruction parameters, enabling the effective use of event cameras in various application domains. Future work could explore adaptive reconstruction strategies that dynamically adjust parameters based on scene dynamics and camera motion, further enhancing the performance of event-based vision systems.

References

Barua, S., Miyatani, Y., Veeraraghavan, A., 2016. Direct face detection and video reconstruction from event cameras. *2016 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 1–9.

Ercan, B., Eker, O., Saglam, C., Erdem, A., Erdem, E., 2024. Hypere2vid: Improving event-based video reconstruction via hypernetworks. *IEEE Transactions on Image Processing*.

Fridrich, J., Goljan, M., 2000. Robust hash functions for digital watermarking. *Proceedings International Conference on Information Technology: Coding and Computing (Cat. No. PR00540)*, IEEE, 178–183.

Gallego, G., Lund, J. E., Mueggler, E., Rebecq, H., Delbruck, T., Scaramuzza, D., 2017. Event-based, 6-DOF camera tracking from photometric depth maps. *IEEE transactions on pattern analysis and machine intelligence*, 40(10), 2402–2412.

Gehrig, D., Rebecq, H., Gallego, G., Scaramuzza, D., 2020. EKLT: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3), 601–618.

Munda, G., Reinbacher, C., Pock, T., 2018. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12), 1381–1393.

Paredes-Vallés, F., De Croon, G. C., 2021. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3446–3455.

Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D., 2019a. Events-to-video: Bringing modern computer vision to event cameras. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3857–3866.

Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D., 2019b. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6), 1964–1980.

Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D., 2020. Fast image reconstruction with an event camera. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 156–163.

Wang, L., Kim, T.-K., Yoon, K.-J., 2020. Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8315–8325.

Yang, Y., Han, J., Liang, J., Sato, I., Shi, B., 2023. Learning event guided high dynamic range video reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13924–13934.

Zou, Y., Zheng, Y., Takatani, T., Fu, Y., 2021. Learning to reconstruct high speed and high dynamic range videos from events. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024–2033.