Optics Letters

Event-based depth estimation with dense occlusion

KANGRUI ZHOU,[1] TAIHANG LEI,[1,2] BANGLEI GUAN,[1,2,*] AND QIFENG YU[1,2]

[1]College of Aerospace Science and Engineering, National University of Defense Technology, Changsha, Hunan 410073, China
[2]Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, Changsha, Hunan 410073, China
*guanbanglei12@nudt.edu.cn

Received 22 February 2024; revised 4 May 2024; accepted 13 May 2024; posted 13 May 2024; published 5 June 2024

Occlusions pose a significant challenge to depth estimation in various fields, including automatic driving, remote sensing observation, and video surveillance. In this Letter, we propose a novel, to the best of our knowledge, depth estimation method for dense occlusion to estimate the depth behind occlusions. We design a comprehensive procedure using an event camera that consists of two steps: rough estimation and precise estimation. In the rough estimation, we reconstruct two segments of the event stream to remove occlusions and subsequently employ a binocular intersection measurement to estimate the rough depth. In the precise estimation, we propose a criterion that the maximum total length of edges of reconstructed images corresponds to the actual depth and search for the precise depth around the rough depth. The experimental results demonstrate that our method is implemented with relative errors of depth estimation below 1.05‰.

https://doi.org/10.1364/OL.521988

Introduction. Depth estimation is one of the most important areas of study in three-dimensional (3D) metrology [1]. In recent years, depth estimation based on event cameras has experienced a surge in popularity [2]. Event cameras possess outstanding properties when compared to standard cameras. They provide a high temporal resolution (in the order of μs), low latency, and very high dynamic range (140 dB vs. 60 dB) [3]. Hence, event cameras have significant potential for depth estimation to address challenging scenarios that often present difficulties for standard cameras [4,5], such as scenes with dense occlusion [6], high-speed motion [7,8], and high dynamic range [9–11].

Occlusions present a significant challenge in depth estimation [12]. Occlusions result in a reduction of light information from the occluded target and disrupt the continuity of the information of targets. Researchers usually use a moving camera or a camera array to capture the information of the scene from multiple views. Among them, the integral imaging-based standard camera demonstrates excellence in removing occlusion [13,14] performing the 3D reconstruction [15] and improving the target detection [16,17]. The integral imaging works based on the intensity, so it may not be suitable for direct application to the event stream that just records the intensity changes. The brightness reconstruction also inevitably leads to the loss of certain brightness details and all color information that may influence the result based on integral imaging. Besides, dense occlusions prohibit the efficient imaging of scenes, because the collected light information is very limited and moreover severely disturbed [6]. Event-based synthetic aperture imaging (SAI) approaches have been developed for dense occlusion scenes. Zhang *et al.* [6] proposed the hybrid SNN–CNN network to make full use of the spatiotemporal information of the event and ensure the overall performance of the occluded target reconstruction. It relies on prior information on target depth which limits its application. To realize the reconstruction in the unknown depth scenes, Yu *et al.* [18] proposed Refocus-Net that can adaptively align signal events and largely facilitate the event-based SAI for real-world scenarios. The aforementioned studies have made significant advances. However, few researchers pay attention to the event-based depth estimation with dense occlusion which is useful in various fields.

To address these problems, we first combine SAI algorithms with edge detection to reconstruct the occluded target and estimate the target depth efficiently. The overview of our method is shown in Fig. 1. The contributions of this Letter are as follows:

1) We design a complete procedure that consists of rough estimation and precise estimation to estimate the target depth behind the occlusion. To the best of our knowledge, this is the first time that we have proposed a novel event-based depth estimation method for scenes with dense occlusion.
2) We propose rough estimation based on a binocular intersection measurement. Two segments of the event stream are reconstructed to remove occlusions. They are considered as the images taken by the left and right virtual cameras, respectively. A rough depth is estimated by intersection.
3) We propose precise estimation according to edge detection. Based on the principle of SAI, we get the sharpest reconstructed image at the actual depth theoretically. Since event cameras accurately capture edge information, we adopt the total length of edges as a metric to evaluate reconstruction quality. A precise depth is obtained when the metric peaks.

Synthetic aperture imaging. SAI enlarges the effective imaging system aperture by the movement of cameras [19]. As the event camera keeps moving, there is a brightness difference between the target and occlusions, thus triggering events. We define $C_{t,ref}$ as the reference camera coordinate system and x_i as the pixel coordinate of the target in another camera coordinate system C_t. We assume that the event camera remains fixed on the camera plane, and axes of all camera coordinate systems are parallel. The coordinates x_i of an event caused by the target at C_t

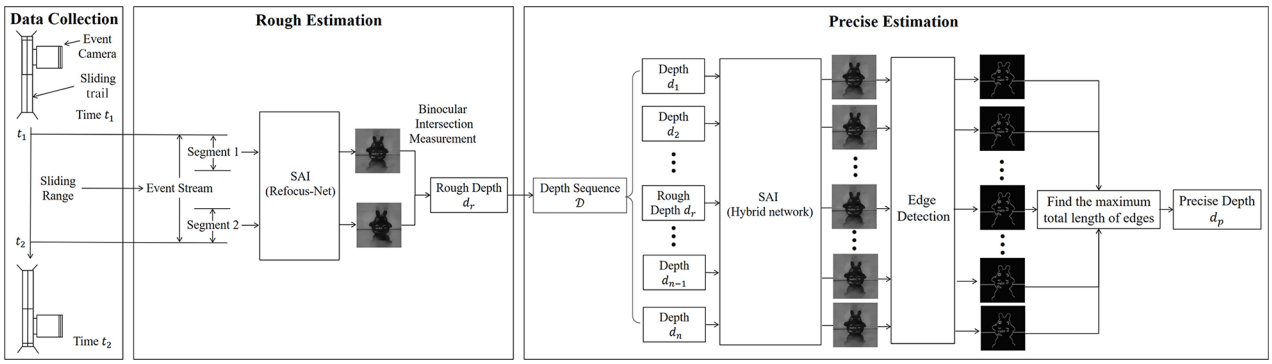

Fig. 1. Overview of our method. It is a two-step process: rough estimation and precise estimation.

are aligned on $C_{t,ref}$ by the mapping relationship as follows [20]:

$$x_{ref,i} = KRK^{-1}x_i + \frac{KT}{d}, \tag{1}$$

where $x_{ref,i}$ represents the pixel coordinate of the event on $C_{t,ref}$, target depth d is the distance between the target and camera plane, K is the intrinsic matrix of the camera, and R and T are the rotation and translation matrix from C_t to $C_{t,ref}$. The intensity of the scene can be reconstructed based on the number of events.

Rough estimation. We select two segments of the event stream and take their starting points as the reference position. Then, we initially reconstruct the occluded target in two segments separately by Refocus-Net [18] and obtain two reconstructed images, which are considered as the images taken by the left and right virtual cameras, respectively.

This Letter uses the SURF algorithm [21] to extract and match the feature points in the reconstructed images. The sliding trail is used as the x axis to establish the coordinate system, and the motion between the left and right virtual cameras can be regarded as the rigid body motion. We calculate the depth of each pair of feature points in two reconstructed images. The average of these depths of feature points is rough depth d_r:

$$d_r = \frac{1}{N}\sum_{k=1}^{n} f(\widetilde{x}_{lk}, \widetilde{x}_{rk}, R, t, K), \tag{2}$$

where N represents the total number of feature point pairs. $\widetilde{x}_{lk}$ and $\widetilde{x}_{rk}$ represent the coordinates of the k-th feature point pair. f maps feature points, the intrinsic and extrinsic to depth.

In the rough estimation, we only reconstruct a slightly blurred image due to the lack of depth information. Then, we conduct an error analysis of the feature points' position mainly caused by the blur and consider the error into the generation of the depth sequence in precise estimation, where we use edge information as a cue to obtain a precise result within the depth sequence.

Precise estimation. According to Eq. (1), the event stream E can be refocused to $C_{t,ref}$ denoted as E_c by mapping θ, which is monotonic about d. Then, events offset in the same direction when there is an error between the target depth d and the actual depth, and this offset becomes more pronounced as the error increases.

During the refocusing, all signal events are successfully aligned. We use a hybrid network [18] to further separate noise events, meanwhile preserving the refocused events from the targets based on the leakage mechanism of LIF neuron, where E_c is mapped to E_{SAI} and the mapping is denoted as H. We can describe the relationship between the intensity image and the number of events by accumulating the event polarities of each pixel of E_{SAI} [22] over a time interval t and produce an intensity image I_{SAI} . $\mathbf{x}$ presents all pixels of E. The total length of the edges L can be expressed as

$$L = G\left(I_0 \cdot exp\left(\int_0^t H(\theta(E(\mathbf{x}, s), P(d)))ds\right)\right), \tag{3}$$

where I_0 is the initial value of the intensity reconstructed image. G is the mapping from the intensity image to the total length of the edges L by the Canny algorithm. $P(d)$ is the mapping matrix when E is focused on d. According to the analysis above, when there is an error in d compared to the actual depth, the events offset to the same direction leading to the smoothness of the intensity I_{SAI} especially around the edges. Then, the gradient of the intensity decreased and the total length of the edges decrease. Therefore, a criterion is proposed that the maximum total length of edges corresponds to the precise depth. The depth sequence $\mathcal{D}$ takes the rough depth d_r as the center and considers the estimated depth error as the length. We calculate L by each depth in $\mathcal{D}$ and the maximum corresponds to the precise depth d_p:

$$d_p = \arg\max_{d_j \in \mathcal{D}} G(\phi(E, P(d_j))), \tag{4}$$

where d_j represents the j-th element in the depth sequence $\mathcal{D}$. ϕ presents the mapping from the event stream and depth to the intensity image. Then, we obtain a more precise result within the depth sequence.

Experimental setup. In this Letter, bamboo fences are made to simulate dense occlusion. The experimental scene is shown in Fig. 2. The resolution of the event camera is 1280×720 pixels.

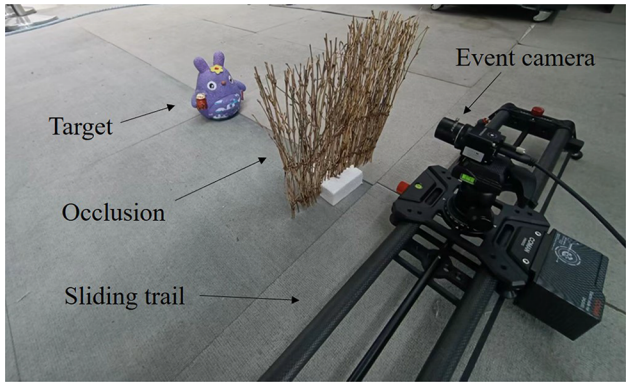

Fig. 2. Experimental scene. Event camera moves on a programmable sliding trail at a constant velocity.

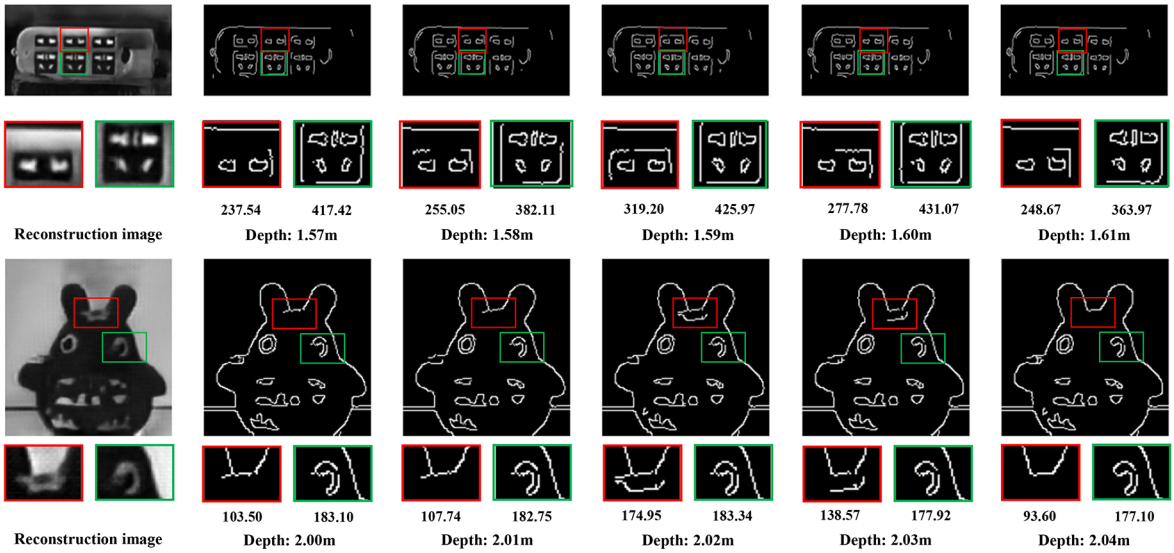

Fig. 3. Quantitative comparisons under different target depths: the global edge detection (row 1), the local edge detection (row 2), and the total length of local edges (row 3).

The time interval is set to 0.7 s, and the velocity is measured at 0.046 m/s. The intrinsic parameters of the event camera are obtained based on camera calibration [23].

Criterion analysis. In the precise estimation, we propose a criterion that the maximum total length of edges corresponds to the actual depth. To verify the effectiveness of this criterion, we conduct experiments to reconstruct occluded targets and measure the total length of edges near their precise depth, as shown in Fig. 3. As indicated by the red rectangle in the Doll scene, the edges are most distinct at the depth of 2.02 m and become less clear as the target depth deviates from the precise depth.

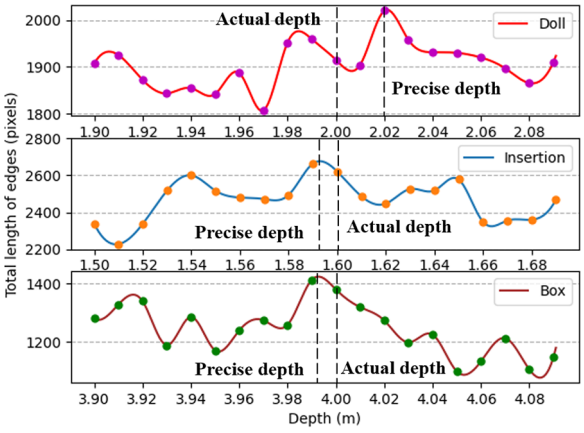

Fig. 4. Total length of edges corresponds to the target depth within the vicinity of the rough depth.

In particular, the total length of edges is sensitive to the target depth. In the red rectangular of the Insertion scene, the total length changes over 15 pixels as the target depth transitions from 1.57 to 1.61 m and the changes in the edge's profile are also obvious. The sensitivity of edges to depth is beneficial to get a precise result.

Depth estimation accuracy. We first estimate the rough depth and then take it as a cue to estimate the precise depth within the vicinity of the rough depth. The details of the precise estimation are shown in Fig. 4. We calculate the total length of edges at an interval of 0.01 m and interpolate them to a finer depth solution within the depth sequence. According to our criterion, the maximum length corresponds to the precise depth.

As shown in Table 1, the precise depth exhibits a lower degree of error when compared to the rough depth. Besides, the flat targets (Insection, UAV photo, and Box) show less error compared to the stereo targets (Bottle and Doll). One possible explanation for this is that we consider the center of the targets as the actual depth. The edges of targets may not be sharpest at the center for the stereo targets, as it may represent a depth within the range of the target. Our method was implemented with relative errors of 1.05% for the stereo target at 2 m and 0.50%, 0.50%, and 0.20% for the flat target at 1.6 m, 2 m, and 4 m, respectively.

Extension for occluded target reconstruction. Since the total length of edges of the reconstructed images is used as the metric, we reconstruct the sharpest images when the metric peaks. Then, we can estimate the precise depth and reconstruct the occluded targets simultaneously. As shown in Fig. 5, we

Table 1. Quantitative Comparisons among Rough Estimation, Precise Estimation, and Real Depth

Scenes	Rough Depth(/m)	Precise Depth(/m)	Actual Depth(/m)	Relative Error
Insection	1.528	1.592	1.600	0.500%
Bottle	2.073	1.979	2.000	1.050%
Doll	2.082	2.021	2.000	1.050%
UVA photo	2.061	2.010	2.000	0.500%
Box	4.097	3.992	4.000	0.200%

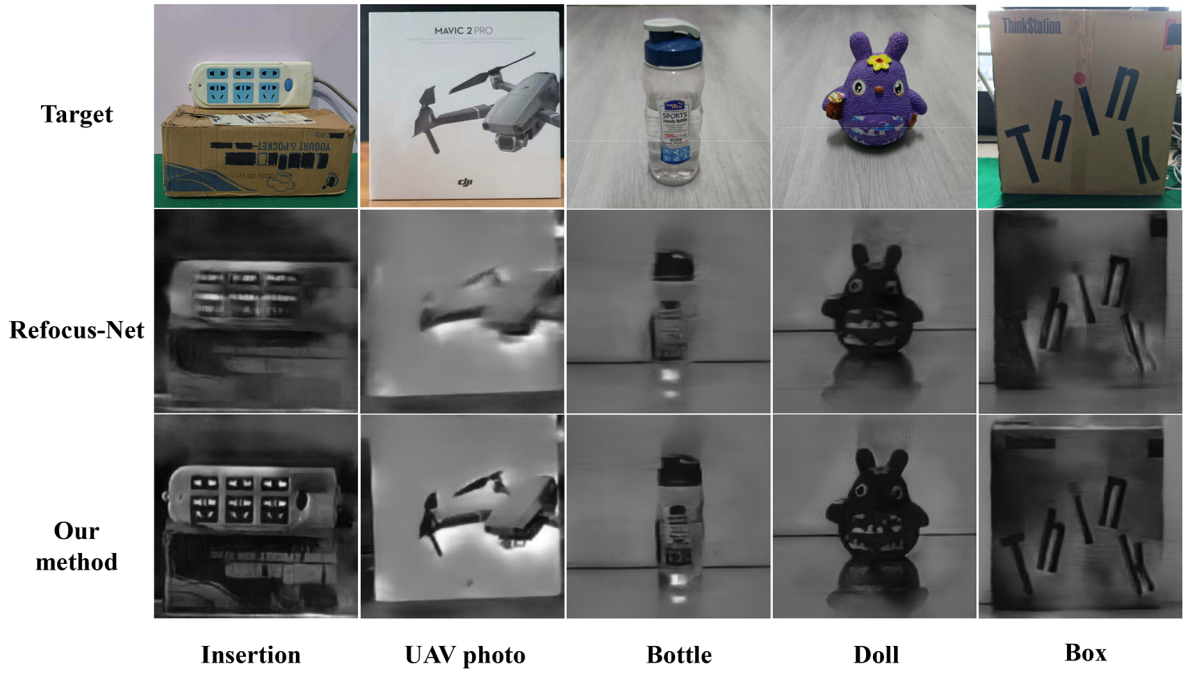

Fig. 5. Qualitative comparisons among target (row 1), reconstructed images by Refocus-Net [18] (row 2), and our method (row 3).

conduct a comparison of the reconstructed images obtained using our method and Refocus-Net. The edges and features of the targets using our method appear sharper than those obtained with Refocus-Net. This work also expands the potential application scenarios of SAI in scenes with unknown target depth.

Conclusion. In this Letter, we propose an event-based depth estimation method for dense occlusion. To achieve accurate depth estimation in scenarios with dense occlusion, we design a complete procedure including rough estimation and precise estimation. The rough estimation reconstructs two segments of the event stream and utilizes a binocular intersection measurement to estimate the rough depth. The precise estimation searches for the precise depth within the vicinity of the rough depth based on the criterion that the maximum total length of edges corresponds to the precise depth. Experimental results on real-world datasets verified the effectiveness of our method.

Funding. National Natural Science Foundation of China (12372189); Science Fund for Distinguished Young Scholars of Hunan Province (2023JJ20045).

Acknowledgment. We would like to acknowledge support from the Key Laboratory of Image Measurement and Vision Navigation.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this Letter are not publicly available at this time but may be obtained from the authors upon reasonable request.

REFERENCES

1. K. Beomjun, H. Daerak, M. Woonchan, *et al.*, Curr. Opt. Photon. **5**, 514 (2021).
2. A. Z. Zhu, D. Thakur, T. Özaslan, *et al.*, IEEE Robot. Automat. Lett. **3**, 2032 (2018).
3. G. Gallego, T. Delbrück, G. Orchard, *et al.*, IEEE Trans. Pattern Anal. Mach. Intell. **44**, 154 (2020).
4. J. Hidalgo-Carrió, D. Gehrig, and D. Scaramuzza, in *International Conference on 3D Vision* , 2020 , pp. 534–542.
5. X. Shao, M. M. Eisa, Z. Chen, *et al.*, Opt. Express **24**, 30124 (2016).
6. X. Zhang, W. Liao, L. Yu, *et al.*, in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14235–14244.
7. T. Huang, Y. Zheng, Z. Yu, *et al.*, Engineering **25**, 110 (2023).
8. Z. Yu, T. Bu, Y. Zhang, *et al.*, *IEEE Trans. Neural Net. Learn. Syst.*, 1 (2024).
9. B. Pan, Q. Kemao, L. Huang, *et al.*, Opt. Lett. **34**, 416 (2009).
10. S. Dong, J. Li, J. Ma, *et al.*, Measurement **226**, 114088 (2024).
11. T. Zhang, Y. Ye, S. Zhang, *et al.*, Opt. Lasers Eng. **154**, 107032 (2022).
12. X. Zhang, Y. Zhang, T. Yang, *et al.*, Pattern Recognition **62**, 175 (2017).
13. X. Xiao, M. Daneshpanah, and B. Javidi, J. Disp. Technol. **8**, 483 (2012).
14. J. Martínez Sotoca, P. Latorre-Carmona, F. Pla, *et al.*, IEEE Access **7**, 1052 (2019).
15. X. Shen, A. Markman, and B. Javidi, Appl. Opt. **56**, D151 (2017).
16. K. Usmani, T. O'Connor, P. Wani, *et al.*, Opt. Express **31**, 479 (2023).
17. V. J. Traver, P. Latorre-Carmona, E. Salvador-Balaguer, *et al.*, IEEE Signal Process. Lett. **24**, 171 (2017).
18. L. Yu, X. Zhang, W. Liao, *et al.*, IEEE Trans. Pattern Anal. Mach. Intell. **45**, 8660 (2022).
19. J.-S. Jang and B. Javidi, Opt. Lett. **27**, 1144 (2002).
20. X. Zhang, W. Liao, L. Yu, *et al.*, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*, 2021, pp. 14230–14239.
21. H. Bay, T. Tuytelaars, and L. Van Gool, in *European Conference on Computer Vision* , 2006 , pp. 404–417.
22. C. Scheerlinck, N. Barnes, and R. Mahony, in *Asian Conference on Computer Vision (ACCV)* (2018), pp. 308–324.
23. Z. Zhang, in *International Conference on Computer Vision*, 1999 , pp. 666–673.