# Active Event-based Stereo Vision
# — Supplementary Material —

Jianing Li Yunjian Zhang Haiqian Han Xiangyang Ji *
Tsinghua University
lijianing@pku.edu.cn, sdtczyj@gmail.com, hanhq23@mails.tsinghua.edu.cn, xyji@tsinghua.edu.cn

## A. Overview

In this supplementary material, we give detailed descriptions of our active event-based stereo vision as follows:

1) Sec. B provides *event camera working principles*, particularly when combined with structured light.
2) Sec. C presents the details about our *active event-based stereo camera prototype* and *the newly built datasets*.
3) Sec. D provides more details of the proposed method, including an in-depth description of the network architecture (i.e., ActiveEventNet).
4) Sec. E revisits some widely used *evaluation metrics* in the stereo matching task.
5) Sec. F reports *additional experimental results* to verify the effectiveness of our ActiveEvenNet, including more visualization results and quantitative metrics.

## B. Event Camera Working Principles

Event cameras [4], such DVS [5, 10, 16], implementing an abstract of the photoreceptor-bipolar-ganglion cell information flow, model the structure of the biological retina. In contrast to conventional frames, each pixel in DVS independently responds to light changes in illumination intensity $I(\boldsymbol{u}, t)$ with a stream of events. More specifically, an event $e_n$ is a four-attribute tuple $\langle x_n, y_n, t_n, p_n \rangle$ using the address event representation (AER), triggered for the pixel $\boldsymbol{u}$=$\langle x_n, y_n \rangle$ at the timestamp $t_n$ when the log-intensity changes over the pre-defined threshold $\theta_{th}$. This process can be depicted as:

$$\ln I(\boldsymbol{u}_n, t_n) - \ln I(\boldsymbol{u}_n, t_n - \Delta t_n) = p_n \theta_{th}, \tag{S1}$$

where $\Delta t_n$ is the temporal sampling interval at a pixel, the polarity $p_n \in \{1, -1\}$ denotes whether the brightness is increasing or decreasing.

The event train $T_s = \{t_n \in \Gamma : n = 1, ..., N\}$ is a sequence of ordered event firing timestamps for one pixel, which can be mathematically described as:

$$T_s(t) = \{p_n \delta(t - t_n)\}_{n=1}^{N_e}, \tag{S2}$$

where $N_e$ is the number of events in single pixel during the time interval, and $\delta(\cdot)$ refers to the Dirac delta function, with $\delta(t) = 0, \forall t \neq 0$ and $\int \delta(t)\, dt = 1$.

Intuitively, asynchronous events appear as sparse and discrete points in the spatiotemporal domain [9, 13, 14, 22], which can be described as follows:

$$S(x, y, t) = \{p_n \delta(x - x_n, y - y_n, t - t_n)\}_{n=1}^{N_e}. \tag{S3}$$

In general, passive event-based vision usually detects changes in light intensity across the 400 nm-780 nm spectrum. Besides, the chip of DAVIS346 camera [24] is sensitive to 300 nm-1000 nm, the generation of the event streams is mainly affected by natural light, laser light, and noise.

In this study, we introduce an innovative approach involving an 850 nm infrared 2D structured light pattern combined with an event camera. This solution enables the generation of dynamic events within static scenes by adjusting laser intensity or frequency. Consequently, event-based camera systems, by introducing structured light, may overcome some challenges in texture-less regions or extremely low light scenes. In fact, bio-inspired event cameras are increasingly being utilized in combination with infrared structured light for high-speed depth sensing. As illustrated in Table S1, structured light sources are commonly categorized into three types (i.e., point, line, and 2D pattern).

## C. Camera Prototype and Datasets

**Realistic Synthetic Dataset.** To meet a huge demand for labor-saving, high-quality synthetic, and labeled disparity maps for benchmarking learning-based stereo matching algorithms, we establish a highly realistic synthetic dataset, namely RealSense-Event-Sim, which uses the V2E simulator [7] to convert infrared stereo videos into event streams. To be precise, an Intel RealSense D435 camera, offering RGB images and infrared images in stereo pairs, is utilized to record 119 video sequences. The raw recordings cover diverse indoor and outdoor scenes, varying lighting conditions, and different camera movement speeds. As shown in Fig. S1, take the left camera for example, we first interpolate

*Corresponding author: Xiangyang Ji.

| Method | Venue | Year | Type | Density | Camera | Resolution | Projector | Framework | Depth label | Code |
|---|---|---|---|---|---|---|---|---|---|---|
| Manasi *et al.* [19] | 3DV | 2021 | Monocular | Sparse | Gen3 | 640×480 | Laser point, 60 Hz | Model-based | Sparse | ✓ |
| Muglikar *et al.* [20] | 3DV | 2021 | Monocular | Dense | Gen3 | 640×480 | Laser point, 60 Hz | Learning-based | Dense | ✗ |
| Brandli *et al.* [2] | FNR | 2014 | Monocular | Sparse | DVS128 | 128×128 | Laser line, 500 Hz | Model-based | No | ✗ |
| Wieland *et al.* [18] | CVPRW | 2023 | Monocular | Sparse | Gen3 | 640×480 | Laser line, 60 Hz | Model-based | No | ✓ |
| Takatani *et al.* [23] | CVPR | 2021 | Monocular | Sparse | DAVIS346 | 346×260 | Laser beam | Model-based | No | ✗ |
| Leroux *et al.* [12] | arXiv | 2018 | Monocular | Sparse | ATIS | 304×240 | Laser 2D pattern | Model-based | No | ✗ |
| Ashish *et al.* [17] | SPL | 2020 | Monocular | Sparse | DAVIS346 | 346×260 | Laser 2D pattern | Model-based | No | ✗ |
| Huang *et al.* [8] | OE | 2021 | Monocular | Sparse | Celex-V | 1280×800 | Laser 2D pattern, 9500 Hz | Model-based | No | ✗ |
| Fu *et al.* [3] | OE | 2023 | Monocular | Dense | EKV4 | 1280×720 | Laser 2D pattern, 60 Hz | Model-based | No | ✗ |
| Bajestani *et al.* [1] | WACV | 2023 | Monocular | Sparse | Gen3 | 640×480 | Laser 2D pattern, 4225 Hz | Model-based | No | ✗ |
| Li *et al.* [15] | IEEE SJ | 2024 | Monocular | Sparse | DAVIS346 | 346×260 | Laser 2D pattern, 60 Hz | Model-based | No | ✗ |
| Wang *et al.* [26] | MTF | 2022 | Monocular | Dense | DVXplorer | 640×480 | Laser 2D pattern | Learning-based | Dense | ✗ |
| **Ours** | - | **2024** | **Stereo** | **Dense** | **DAVIS346** | **346×260** | **Laser 2D pattern** | **Learning-based** | Dense | ✓ |

Table S1. A literature review of event-based vision system using active infrared light.

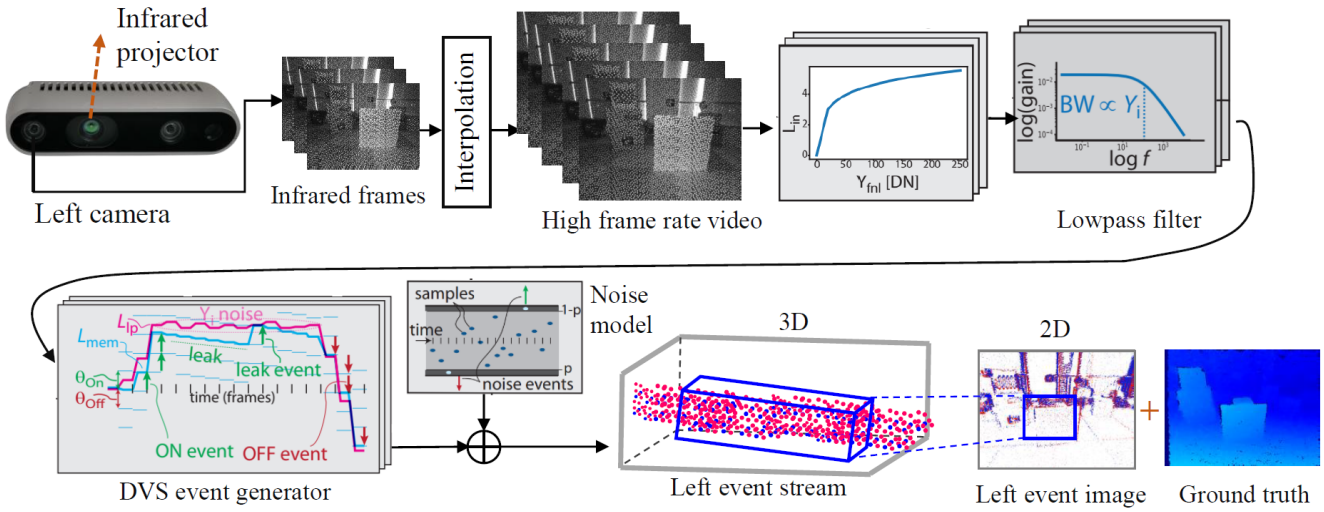


Figure S1. The pipeline of generating active event-based stereo matching simulated dataset. Initially, the simulator interpolates the low-frame-rate infrared video to create a high-frame-rate video. Then, we use a linear-to-logarithmic mapping operation, implement a low-pass filter, and apply a differential comparison principle to generate events. Lastly, we incorporate the noise model into raw event streams.

the low-frame-rate infrared video to create a high-frame-rate video. Subsequently, we utilize a linear-to-logarithmic mapping operation, implement a low-pass filtering procedure, and apply a differential comparison principle to generate dynamic events. Lastly, we incorporate the noise model from the event camera into raw event streams. In particular, since the raw infrared videos are associated with active stereo vision, the synthetic event data comprises 2D infrared patterns with structured light. Compared to passive stereo vision, active event-based stereo vision can overcome some challenges in texture-less regions or dark conditions. Statistically, our newly built dataset (i.e., RealSense-Event-Sim) consists of stereo pairs of event streams and 23.8k synchronized ground truth labels recorded at 20 Hz. This highly synthetic dataset is divided into 16k samples for training, 3.8k for validation, and 4k for testing.

**Spatiotemporal Calibration for Camera Prototype.** We build a prototype stereo camera system by integrating two DAVIS346 cameras (i.e., resolution 346×260), an infrared 2D pattern projector, and an Intel RealSense D455 camera. Spatiotemporal calibration is a critical step for hybrid multi-camera systems. For ***temporal calibration***, we synchronize the two stereo event cameras and the RealSense D455 camera in the temporal domain by publishing the timestamp of each topic in the robot operating system (ROS). For ***spatial calibration***, one objective is to establish a horizontal baseline correction for stereo matching between the two event cameras. Another goal is to ensure that the RealSense camera shares the same view as the left event camera. The main reason is that we use the flagship RealSense D455 stereo camera at 15 FPS to capture depth ground truth in normal scenes and also as a fair comparison in high-speed motion scenarios. More specifically, a standard checkerboard is first placed in front 1 m away from our active event-based stereo camera system to make a full view [27]. Then, we utilize a professional binocular stereo matching correction toolbox to perform baseline correction on RGB images from two DAVIS346 cameras. Meanwhile, the homogra-

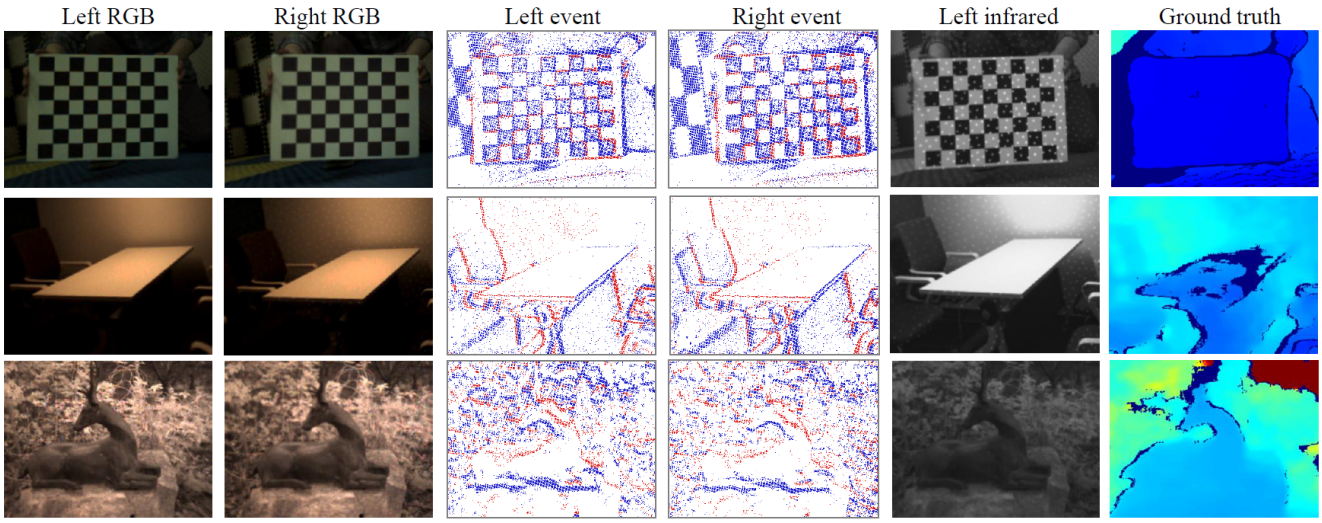


Figure S2. Representative spatiotemporal synchronized examples of our Active-Event-Stereo dataset. We use binocular DAVIS346 cameras and a RealSense D455 camera to record stereo event streams and depth values in both indoor and outdoor scenarios.

phy adopts an affine transformation that connects two sets of coordinates as follows:

$$[x_i^1, y_i^1, 1]^T = \boldsymbol{R} \cdot [x_i^2, y_i^2, 1]^T, \tag{S4}$$

where a 3×3 matrix $\boldsymbol{R}$ is computed using the checkerboard keypoints between two cameras. $[x_i^1, y_i^1, 1]^T$ and $[x_i^2, y_i^2, 1]^T$ are the coordinates in two RGB images from the left DAVIS346 camera and the RealSense D455.

**Real-world Dataset.** We use the built stereo camera prototype to record 85 sequences including event stream pairs, RGB frames, infrared frames, and depth values. After spatiotemporal calibration, all labels are provided at a frequency of 15 Hz by the RealSense D455. As a result, the newly built dataset (i.e., Active-Event-Stereo) offers event streams in stereo pairs and 21.5k synchronized true labels. Afterward, we split them into 14.6k for training, 3.6k for validation, and 3.3k for testing. As illustrated in Fig. S2, we show some representative spatiotemporal synchronized examples using our stereo camera prototype. In particular, these scenarios take velocity distribution, illumination change, scene diversity, and varying distances into account. All in all, such a novel event-based stereo system with structured light and professional design enables our Active-Event-Stereo to be a competitive dataset.

## D. Architecture Network Details

This work aims at designing a lightweight yet effective active event-based stereo matching network, termed ***ActiveEventNet***, which generates high-speed dense disparity maps via integrating binocular event cameras and infrared structured light. More specifically, the input event stream is first divided into temporal bins, with each bin transformed into a 2D image-like representation, referred to as event tensors. To improve the efficiency of event-based stereo matching models, we incorporate lightweight MobileNet blocks alongside standard convolutions, which serve as critical components for tasks like feature extraction and encoder-decoder processing. Then, event stereo embeddings are passed through a feature extraction module and a channel reduction module, producing compact yet highly informative features. Furthermore, we present a novel 3D cost volume design that dynamically exchanges channels and concatenates interaction features from stereo inputs. This cost volume represents the matching costs of corresponding pixels between two event streams captured from slightly different viewpoints. Finally, the 3D cost volume is processed through an encoder-decoder module equipped with a stack of lightweight convolutional layers. This is followed by dense disparity map generation using regression modules.

## E. Evaluation Metrics

This section gives the more details of six metrics from the stereo matching task [11, 21], which includes EPE, RMSE, D1-all, and bad pixel ratios (i.e., >1px, >2px, and >3px).

**End-Point-Error (EPE).** The EPE is calculated as the average $L_1$ distance between each estimated disparity $\boldsymbol{d}_e$ and its corresponding ground truth disparity $\boldsymbol{d}_{gt}$. It can be formulated as follows:

$$\text{EPE} = \frac{1}{N_p} \sum_{i=1}^{N_p} |\boldsymbol{d}_e^{(i)} - \boldsymbol{d}_{gt}^{(i)}|, \tag{S5}$$

where $N_p$ is the total number of pixels in the disparity map.

**Root Mean Square Error (RMSE).** The RMSE measures the average magnitude of the errors between each estimated

| Scenario | Sequence | EPE ↓ | RMSE ↓ | D1-all ↓ | >1px ↓ | >2px ↓ | >3px ↓ | Runtime (ms) |
|---|---|---|---|---|---|---|---|---|
| Indoor with normal light | 002_indoor_boxes | 1.064 | 1.834 | 0.063 | 0.329 | 0.131 | 0.063 | 23.5 |
| | 003_indoor_desk | 1.320 | 2.125 | 0.097 | 0.408 | 0.175 | 0.097 | 22.3 |
| | 010_indoor_desk | 2.408 | 4.671 | 0.161 | 0.538 | 0.288 | 0.181 | 19.8 |
| | 042_indoor_car | 0.761 | 1.164 | 0.020 | 0.538 | 0.070 | 0.020 | 20.4 |
| | 050_indoor_office_cabinet | 1.108 | 1.780 | 0.063 | 0.371 | 0.129 | 0.062 | 20.7 |
| | 059_indoor_conference_room | 0.777 | 1.184 | 0.025 | 0.248 | 0.066 | 0.025 | 29.2 |
| | 064_indoor_office_checkerboard | 0.983 | 1.948 | 0.042 | 0.257 | 0.085 | 0.042 | 23.8 |
| | 073_indoor_floor | 2.531 | 4.518 | 0.186 | 0.562 | 0.303 | 0.186 | 20.2 |
| | 078_indoor_office_room | 1.091 | 2.195 | 0.055 | 0.272 | 0.099 | 0.055 | 19.5 |
| | 106_indoor_checkerboard | 1.075 | 2.855 | 0.043 | 0.224 | 0.075 | 0.043 | 20.1 |
| | 107_indoor_checkerboard | 0.863 | 2.007 | 0.027 | 0.226 | 0.061 | 0.027 | 28.4 |
| Indoor with low light | 015_indoor_desk_night | 1.351 | 2.120 | 0.059 | 0.338 | 0.120 | 0.059 | 24.5 |
| | 057_indoor_office_night | 1.403 | 2.481 | 0.105 | 0.390 | 0.179 | 0.105 | 28.5 |
| | 108_indoor_checkerboard_night | 0.661 | 1.007 | 0.015 | 0.188 | 0.038 | 0.002 | 29.8 |
| Outdoor with normal light | 090_outdoor_floor | 0.779 | 1.235 | 0.027 | 0.233 | 0.067 | 0.027 | 20.5 |
| | 097_outdoor_bridge | 0.606 | 1.035 | 0.021 | 0.146 | 0.044 | 0.021 | 28.2 |
| | 100_outdoor_people | 0.714 | 1.308 | 0.022 | 0.162 | 0.040 | 0.022 | 24.9 |
| Outdoor with low light | 118_outdoor_wall_night | 1.577 | 2.980 | 0.123 | 0.375 | 0.193 | 0.123 | 29.8 |
| | 119_outdoor_wall_night | 1.858 | 3.301 | 0.149 | 0.420 | 0.226 | 0.149 | 26.3 |
| **All** | **Average** | **1.223** | **2.320** | **0.070** | **0.314** | **0.127** | **0.070** | **21.9** |

Table S2. Performance evaluation of our highly synthetic RealSense-Event-Sim dataset in various scenarios.

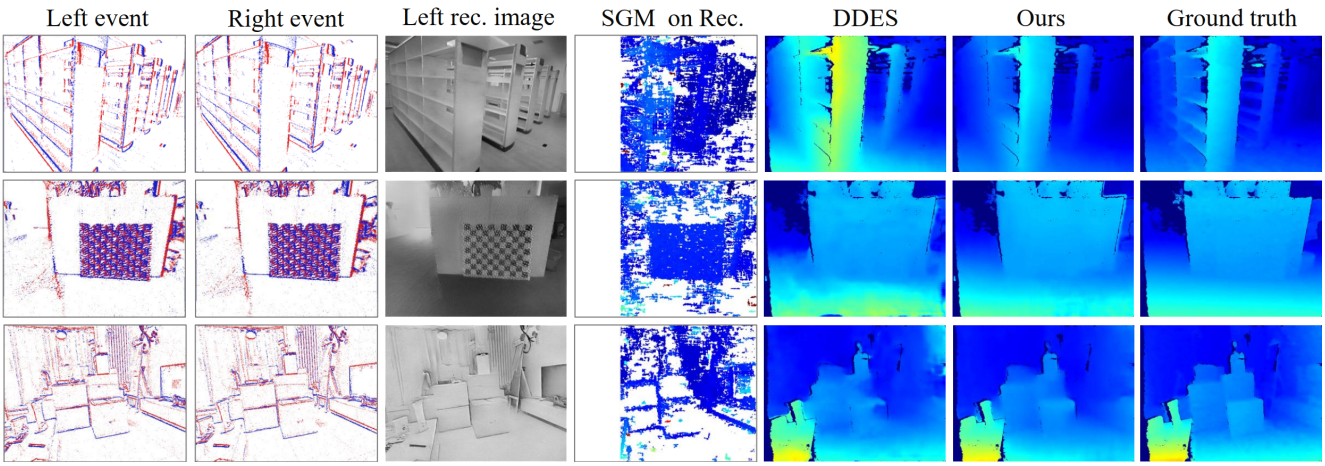


Figure S3. Representative examples of different stereo matching results on our synthetic RealSense-Event-Sim dataset. To enhance visualization and comparison, we implement a mask to the void areas of the ground truth to the predicted dense maps.

disparity $\boldsymbol{d}_e$ and its corresponding ground truth disparity $\boldsymbol{d}_{gt}$. It can be described as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\boldsymbol{d}_e^{(i)} - \boldsymbol{d}_{gt}^{(i)})^2}. \tag{S6}$$

**D1-all.** The D1-all usually refers to the percentage of pixels where either the absolute disparity error is greater than 3 pixels meanwhile the relative disparity error exceeds 0.05. Thus, this metric can be depicted as follows:

$$\text{D1-all} = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbb{I}[|\boldsymbol{d}_e^{(i)} - \boldsymbol{d}_{gt}^{(i)}| > 3) \ \wedge \ \frac{\boldsymbol{d}_e^{(i)} - \boldsymbol{d}_{gt}^{(i)}}{\boldsymbol{d}_{gt}^{(i)}} > 0.05]),$$

where $\mathbb{I}$ represents the indicator function that takes on the value 1 if a specified condition is true, and 0 otherwise.

**Bad Pixel Ratios (BPR, >1px, >2px, and >3px).** The three metrics refer to the bad pixel ratio that measures the percentage of pixels with disparity errors exceeding a threshold $\theta_\delta$ (e.g., 1 pixel, 2 pixels, and 3 pixels) as follows:

$$\text{BPR} = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbb{I}[|\boldsymbol{d}_e^{(i)} - \boldsymbol{d}_{gt}^{(i)}| > \theta_\delta)]. \tag{S7}$$

# F. Additional Experiments

## F.1. Effective Test on Synthetic Dataset

**Evaluation on Various Sequences.** To give a detailed assessment of our synthetic RealSense-Event-Sim dataset, we report the quantization results of each sequence in Table S2.

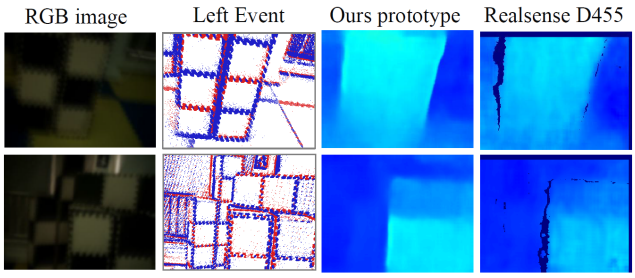


Figure S4. Comparison with our stereo camera prototype and the RealSense D455 in high-speed motion blur scenarios.

Notably, our ActiveEventNet achieves a satisfactory performance, yielding an average end-point-error (EPE) of 1.223 and the D1-all metric of 0.070, while delivering real-time inference at a speed of nearly 50 frames per second (FPS) for a resolution of 640×480.

**Comparison with SOTA Methods.** To further verify the effectiveness of our ActiveEventNet, we report representative visualization results of stereo matching algorithms on our synthetic RealSense-Event-Sim dataset in Fig. S3. Note that, the classical SGM [6] only obtains sparse disparity maps from stereo reconstructed images. Our ActiveEventNet produces a higher quality dense disparity map than the DDES method [25]. For example, our method discerns that the sharpness of the edge contour closely resembles the ground truth, whereas the edge contour produced by the DDES method lacks such sharpness.

### F.2. Scalability Test on Real-world Dataset

To verify the effectiveness of our solution for high-speed depth sensing, we compare our stereo camera prototype with the RealSense D455 in high-speed motion blur scenes. As illustrated in Fig. S4, RGB images exhibit motion blur in high-speed scenes, making it challenging to discern the object's outline. Conversely, the event camera excels in capturing the edge contour of the object. In particular, when comparing disparity maps generated by our method with that of the RealSense D455, our solution produces a higher-quality disparity map. In other words, our solution using event cameras outperforms conventional frames for high-speed depth sensing.

## References


[1] Seyed Ehsan Marjani Bajestani and Giovanni Beltrame. Event-based rgb sensing with structured light. In *WACV*, pages 5458–5467, 2023. 2

[2] Christian Brandli, Thomas A Mantel, Marco Hutter, Markus A Höpflinger, Raphael Berner, Roland Siegwart, and Tobi Delbruck. Adaptive pulsed laser line extraction for terrain reconstruction using a dynamic vision sensor. *Front. Neurosci.*, 7:275, 2014. 2

[3] Jiacheng Fu, Yueyi Zhang, Yue Li, Jiacheng Li, and Zhiwei Xiong. Fast 3d reconstruction via event-based structured light with spatio-temporal coding. *Opt. Eng.*, 31(26):44588–44602, 2023. 2

[4] 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 TPAMI*, 44(1):154–180, 2020. 1

[5] Menghan Guo, Shoushun Chen, Zhe Gao, Wenlei Yang, Peter Bartkovjak, Qing Qin, Xiaoqin Hu, Dahai Zhou, Qiping Huang, Masayuki Uchiyama, et al. A three-wafer-stacked hybrid 15-mpixel cis +1-mpixel evs with 4.6-gevent/s readout, in-pixel tdc, and on-chip isp and esp function. *IEEE JSSC*, 2023. 1

[6] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2007. 5

[7] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *CVPRW*, pages 1312–1321, 2021. 1

[8] Xueyan Huang, Yueyi Zhang, and Zhiwei Xiong. High-speed structured light based 3d scanning using an event camera. *Opt. Eng.*, 29(22):35864–35876, 2021. 2

[9] Zhaodong Kang, Jianing Li, Lin Zhu, and Yonghong Tian. Retinomorphic sensing: A novel paradigm for future multimedia computing. In *ACM MM*, pages 144–152, 2021. 1

[10] Kazutoshi Kodama, Yusuke Sato, Yuhi Yorikado, Raphael Berner, Kyoji Mizoguchi, Takahiro Miyazaki, Masahiro Tsukamoto, Yoshihisa Matoba, Hirotaka Shinozaki, Atsumi Niwa, et al. 1.22 $\mu$m 35.6 mpixel rgb hybrid event-based vision sensor with 4.88 $\mu$m-pitch event pixels and up to 10k event frame rate by adaptive control on event sparsity. In *ISSCC*, pages 92–94, 2023. 1

[11] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE TPAMI*, 44 (4):1738–1764, 2020. 3

[12] T Leroux, S-H Ieng, and Ryad Benosman. Event-based structured light for depth reconstruction using frequency tagged light patterns. *arXiv*, 2018. 2

[13] Dianze Li, Jianing Li, and Yonghong Tian. Sodformer: Streaming object detection with transformer using events and frames. *IEEE TPAMI*, 45(11):14020–14037, 2023. 1

[14] Jianing Li, Xiao Wang, Lin Zhu, Jia Li, Tiejun Huang, and Yonghong Tian. Retinomorphic object detection in asynchronous visual streams. In *AAAI*, pages 1332–1340, 2022. 1

[15] Yuhui Li, Heng Jiang, Chen Xu, and Lilin Liu. Event-driven fringe projection structured light 3d reconstruction based on time-frequency analysis. *IEEE Sensor J.*, 24(4):5097–5106, 2024. 2

[16] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15$\mu$s latency asynchronous temporal contrast vision sensor. *IEEE JSSC*, 43(2):566–576, 2008. 1

[17] Ashish Rao Mangalore, Chandra Sekhar Seelamantula, and Chetan Singh Thakur. Neuromorphic fringe projection profilometry. *IEEE SPL*, 27:1510–1514, 2020. 2

[18] Wieland Morgenstern, Niklas Gard, Simon Baumann, Anna Hilsmann, and Peter Eisert. X-maps: Direct depth lookup for event-based structured light systems. In *CVPRW*, pages 4006–4014, 2023. 2

[19] Manasi Muglikar, Guillermo Gallego, and Davide Scaramuzza. Esl: Event-based structured light. In *3DV*, pages 1165–1174, 2021. 2

[20] Manasi Muglikar, Diederik Paul Moeys, and Davide Scaramuzza. Event guided depth sensing. In *3DV*, pages 385–393, 2021. 2

[21] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: A survey. *IEEE TPAMI*, 44(9):5314–5334, 2021. 3

[22] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proc. IEEE.*, 102(10):1470–1484, 2014. 1

[23] Tsuyoshi Takatani, Yuzuha Ito, Ayaka Ebisu, Yinqiang Zheng, and Takahito Aoto. Event-based bispectral photometry using temporally modulated illumination. In *CVPR*, pages 15638–15647, 2021. 2

[24] 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 TCS-II*, 65(5):677–681, 2018. 1

[25] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *ICCV*, pages 1527–1537, 2019. 5

[26] Huijiao Wang, Tangbo Liu, Chu He, Cheng Li, Jianzhuang Liu, and Lei Yu. Enhancing event-based structured light imaging with a single frame. In *IEEE MFI*, pages 1–7, 2022. 2

[27] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000. 2