Active Event-based Stereo Vision

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

Abstract

Conventional frame-based imaging for active stereo systems has encountered major challenges in fast-motion scenarios. However, how to design a novel paradigm for high-speed depth sensing still remains an open issue. In this paper, we propose a novel problem setting, namely active event-based stereo vision, which provides the first insight of integrating binocular event cameras and an infrared projector for high-speed depth sensing. Technically, we first build a stereo camera prototype system and present a real-world dataset with over 21.5k spatiotemporal synchronized labels at 15 Hz, while also creating a realistic synthetic dataset with stereo event streams and 23.8k synchronized labels at 20 Hz. Then, we propose ActiveEventNet, a lightweight yet effective active event-based stereo matching neural network that learns to generate high-quality dense disparity maps from stereo event streams with low latency. Experiments demonstrate that our ActiveEventNet outperforms state-of-the-art methods meanwhile significantly reducing computational complexity. Our solution offers superior depth sensing compared to conventional stereo cameras in high-speed scenes, while also achieving the inference speed of up to 150 FPS with our prototype. We believe that this novel paradigm will provide new insights into future depth sensing systems. Our project can be available at https://github.com/jianing-li/active_event_based_stereo.

1. Introduction

Stereo vision [24, 36, 57], one of the longstanding and fundamental topics, supports a wide range of computer vision and robotic tasks. Passive stereo vision typically struggles in texture-less regions and low-light environments [33]. In contrast, active stereo vision [4, 19] addresses these challenges by projecting an infrared pattern, enabling more accurate depth maps compared to passive stereo systems. Nevertheless, conventional active stereo cameras are generally constrained by their depth frame rates (e.g., 30 FPS for Kinect V2 and 90 FPS for RealSense D435), limiting their effectiveness in high-speed scenarios [18, 46]. For instance, a racing drone could suffer a severe collision in a short period between two adjacent depth frames. This may raise a key question: *How can we develop a novel active stereo sensing paradigm for high-speed depth perception to overcome the limitations of conventional stereo cameras?*

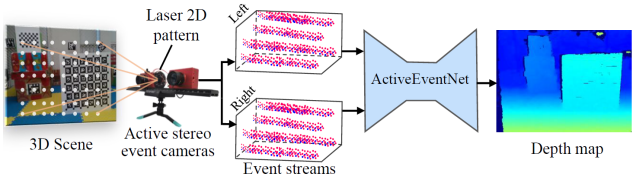


Figure 1. Our active event-based stereo camera system integrates binocular event cameras and an infrared 2D pattern laser for high-speed depth sensing. Our lightweight ActiveEventNet effectively converts stereo event streams into high-quality depth maps.

Event cameras [21, 38, 42], also known as silicon retinas, operate differently from conventional frame-based cameras. Instead of capturing frames at a fixed rate, they detect changes in intensity at each pixel, generating asynchronous events with microsecond-level temporal resolution [7, 39]. This unique capability makes them ideal for various high-speed vision tasks [26, 40, 44, 55, 61, 69, 81, 83, 86] that require low-latency processing of visual information. Consequently, there is growing research interest in leveraging event cameras for high-speed depth sensing (i.e., monocular and binocular) in agile robots [17]. In particular, event-based stereo vision [53, 67, 85, 87] offers the advantage of providing more accurate and reliable depth information compared to monocular depth estimation [11, 27, 45, 89].

One problem is that current event-based stereo depth systems [1, 2, 14, 43, 68, 82] are passive stereo, resulting in *inaccurate depth maps in texture-less regions and dark scenes*. In other words, passive event-based stereo methods rely on feature matching, which may be challenging or even impossible in areas with little texture or low contrast. While some structured light systems with a single event camera [3, 20, 32, 41, 51, 52, 70, 71, 75] have attempted to achieve high-speed depth sensing, these active monocu-

*Corresponding author: Xiangyang Ji.

lar depth methods may not always match the accuracy of stereo depth systems at long distances or in ambient light conditions. In fact, by combining active infrared structured light and passive visible light, stereo camera systems could be empowered to generate robust and accurate depth maps in diverse real-world settings, regardless of lighting conditions or surface textures. Yet, there is still no active stereo vision system that combines binocular event cameras and structured light. Meanwhile, most existing passive event-based stereo vision datasets [5, 8, 23, 30, 88] typically provide only sparse LiDAR depth maps, and there is a lack of event streams with structured light and dense depth labels.

Another problem is that most existing event-based stereo matching algorithms [10, 13, 14, 43, 50, 53, 59, 67, 68] prioritize maximizing accuracy via more complex deep learning-based models, leading to *slower inference on resource-constrained devices*. This is contrary to the initial objectives of designing high-speed and energy-efficient event cameras. For example, constructing an efficient 3D or 4D cost volume is one critical step in end-to-end pipelines. This volume represents the similarity between pixels in a stereo pair, affecting the accuracy and computational speed of stereo matching models. Some event-based stereo matching algorithms [12, 15, 53, 68, 87] aim to build higher dimensional cost volumes to further enhance accuracy, there have been few endeavors to design a lightweight end-to-end model that specifically addresses computational cost and improves inference speed without sacrificing accuracy.

To address the aforementioned problems, we propose a novel paradigm for high-speed depth sensing, namely active event-based stereo vision, which first integrates binocular event cameras and an infrared pattern projector for stereo matching with low-latency (see Fig. 1). In fact, the goal of this work is not to optimize passive event-based stereo matching algorithms for higher accuracy. In contrast, our goal is to overcome the following challenges: (i) *Lack of prototype system and dataset* - How do we establish an active stereo camera setup and build high-quality event-based datasets including both simulated and real-world scenes? (ii) *Lightweight yet effective model* - How do we design a lightweight stereo matching model that effectively reduces the computational expense without sacrificing accuracy?

To this end, we first build an active event-based stereo camera prototype and present a real-world dataset with over 21.5k spatiotemporal synchronized true labels at 15 Hz, while also creating a highly realistic synthetic dataset with 23.8k synchronized labels at 20 Hz. Then, we design a lightweight neural network (i.e., ActiveEventNet) for active event-based stereo matching, which mainly involves two strategies: incorporating lightweight blocks into event-based stereo matching frameworks and designing a novel cost volume for similarity measurement in stereo pairs. The results show that our ActiveEventNet achieves better performance than state-of-the-art methods meanwhile significantly reducing computational complexity. Our prototype, integrating binocular event cameras and an infrared projector, offers superior depth sensing compared to conventional stereo cameras in high-speed scenarios, while also achieving the inference speed of up to 150 FPS. Our solution also highlights that active stereo surpasses passive stereo in low-texture regions and low-light scenarios. We believe that our prototype will provide novel insight into developing the next-generation neuromorphic stereo cameras.

The main contributions of this work are summarized as:

- We present a novel problem setting of *active event-based stereo vision* that combines infrared structured light and binocular event cameras for high-speed depth sensing.
- We propose a *lightweight yet effective neural network* for event-based stereo matching, namely ActiveEventNet, which significantly reduces the computational complexity meanwhile maintaining comparable accuracy.
- We establish a *real-world dataset* using our active event-based stereo camera prototype, along with a *highly realistic synthetic dataset* that contains temporally continuous labels. We believe these two standardized datasets open up opportunities for research in this novel problem.

2. Related Work

Event-based Stereo Vision. Event-based stereo matching methods can be broadly divided into two categories. Early model-based works [85, 87] usually find reliable local correspondences and global optimization algorithms to calculate the disparity. Although these model-based methods achieve sparse or semi-dense depth maps in real-time, they are hard to obtain global dense depth maps. Nowadays, deep learning-based methods [1, 14, 43, 53, 67, 68, 74, 82] have exhibited superior performance in predicting dense depth maps from stereo event streams. Moreover, some cross-modal learning-based methods [12, 13, 50] attempt to generate dense depth maps from stereo hybrid event-frame cameras. Nevertheless, these learning-based methods substantially improve accuracy meanwhile increasing computational complexity. In other words, they require a considerable amount of GPU memory, making them impractical for agile robots or mobile devices. Thus, this work aims to design a lightweight yet effective neural network for event-based stereo matching with low latency.

Event-based Vision with Structured Light. Bio-inspired event cameras are increasingly being utilized in combination with infrared structured light for high-speed depth sensing. In general, structured light sources are commonly categorized into three types (i.e., point, line, and 2D pattern). For instance, Bramdli *et al.* [6] first integrates a laser line-projector and an event camera for 3D reconstruction. Martel *et al.* [48] combine a laser light source with an event-based stereo setup. Muglikar *et al.* [51] design a structured light

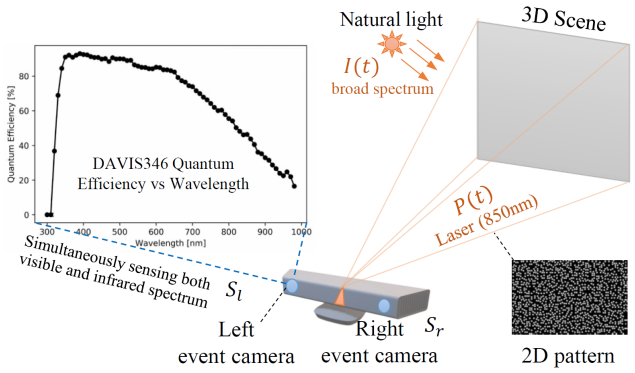


Figure 2. Sensing mechanism of active event-based stereo vision. Projecting a 2D pattern from an infrared projector enhances scene texture, enabling improved stereo matching performance.

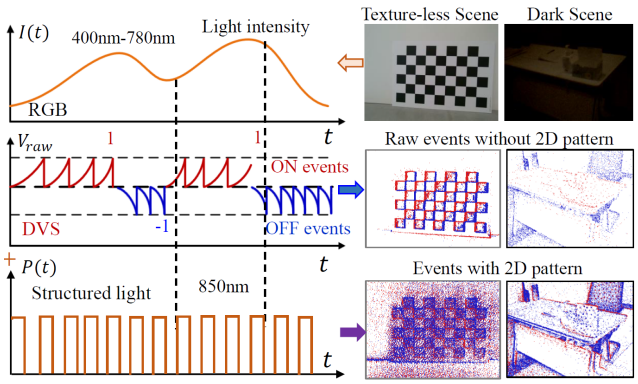


Figure 3. Comparison with event streams using or without structured light. Our solution enables the event camera to generate dynamic events even in static scenarios or low-light conditions.

system with a laser point-projector and an event camera for depth estimation. Huang *et al.* [32] present a structured light system using an event camera and a laser pattern-projector for high-speed 3D scanning. Most systems [6, 32] using model-based methods achieve high-speed depth sensing but yield sparse depth maps. More recently, deep learning models [52, 71] have explored event-based structured light systems for dense depth estimation. While these monocular active event-based vision systems are generally simpler, stereo systems often provide more accurate depth maps due to triangulation. Hence, we design an active event-based stereo structured light system, which uses a lightweight learning-based model to generate dense depth maps.

3. Problem Formulation

Event cameras [21, 80], such as DVS [25, 35, 42, 54], respond to light changes with continuous asynchronous event streams. Each event e_n is depicted by a tuple $\langle x, y, t, p \rangle$, including spatial coordinates $\langle x, y \rangle$, timestamp t, and polarity p. This sensing mechanism enables event-based stereo vision to achieve reliable disparity maps in high-speed dynamic scenes [56, 58, 72, 76, 79, 90]. Yet, passive event-based stereo vision meets challenges in scenarios with texture-less and low light. This work aims to overcome this gap by integrating binocular event cameras and an infrared pattern projector. We formulate this novel problem setting called ***active event-based stereo vision*** as follows.

The goal of this work is to calculate the disparity map $\boldsymbol{d}_e$ from the stereo stream pair $\boldsymbol{S}_l$ and $\boldsymbol{S}_r$. Since the chip of event camera [65] is sensitive to a spectrum of 300-1000 nm (see Fig. 2), the generation of the event stream is mainly affected by natural light, laser light (850 nm), and noise, which can be mathematically formulated as:

$$\begin{aligned} \boldsymbol{S}_l &= \mathcal{G}\left(\boldsymbol{W}_t^l \cdot [I(t) + P(t)] + \mathcal{N}\right), \\ \boldsymbol{S}_r &= \mathcal{G}\left(\boldsymbol{W}_t^r \cdot [I(t) + P(t)] + \mathcal{N}\right), \end{aligned} \tag{1}$$

where $\mathcal{G}$ denotes the event generation process of event cameras. $I(t)$ and $P(t)$ are the scene light intensity and the infrared laser intensity at time t. $\boldsymbol{W}_t^l$ and $\boldsymbol{W}_t^r$ refers to the left and right warping operation via projecting a real 3D scene to each 2D camera plane. $\mathcal{N}$ is the event camera noise. *Note that, the proposed system enables the event camera to generate dynamic events by adjusting the laser's frequency or intensity, even in static scenarios with constant light intensity $I(t)$ or in extremely low-light conditions.*

In general, event-based stereo matching estimates disparities between corresponding points in two streams as:

$$\boldsymbol{d}_e = \mathcal{M}_d(\boldsymbol{S}_l, \boldsymbol{S}_r, \theta), \tag{2}$$

where $\mathcal{M}_d$ is the proposed active event-based stereo matching model, and θ denotes the optimized parameters of $\mathcal{M}_d$.

Then, we can solve the following minimization problem:

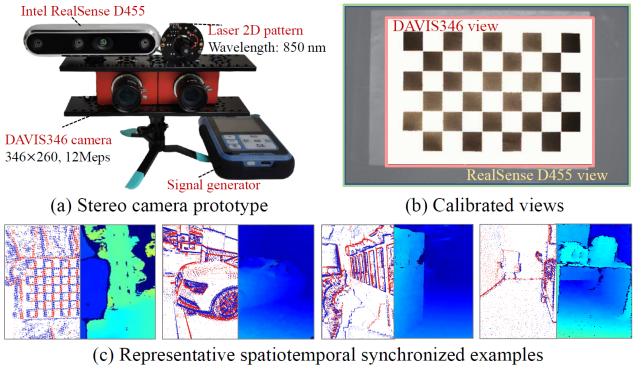
$$\hat{\theta} = \arg\min_{\theta} L_{\mathcal{M}}\left(\boldsymbol{d}_e, \boldsymbol{d}_{gt}\right) + \lambda \Phi\left(\theta\right), \tag{3}$$

where $L_{\mathcal{M}}\left(\boldsymbol{d}_e, \boldsymbol{d}_{gt}\right)$ is the loss function between the predicted disparity $\boldsymbol{d}_e$ and the ground truth $\boldsymbol{d}_{gt}$, and $\Phi\left(\theta\right)$ is the regularization term, and λ is the trade-off parameter.

4. Active Event Camera Stereo Dataset

This section first describes how we built our Active-Event-Stereo dataset with our camera prototype and then provides statistics for a better understanding of this new dataset.

Stereo Camera Prototype. To verify the effectiveness of our solution, we build a prototype stereo camera system by integrating binocular DAVIS346 cameras (i.e., resolution 346×260), an infrared pattern projector, and an Intel RealSense D455 camera (i.e., resolution 640×480). Unlike conventional passive vision, our prototype can detect dynamic events even in static scenarios or dark environments by adjusting the laser frequency or intensity of the signal generator (see Fig. 4 (a)). This capability is enabled by an



(a) Stereo camera prototype (b) Calibrated views

(c) Representative spatiotemporal synchronized examples

Figure 4. An event-based stereo camera prototype and the newly built dataset. (a) The experimental setup combines binocular event cameras and an infrared pattern projector. (b) Spatiotemporal calibration using a standard checkerboard. (c) Examples of the real-world dataset, the left in each image is from the left DAVIS346, and the right is the depth map from the RealSense D455.

infrared projector operating at a wavelength of 850 nm. Besides, 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.

Spatiotemporal Calibration. In general, 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 by publishing each topic's timestamp in the robot operating system (ROS). *For spatial calibration*, our objectives are to establish a horizontal baseline correction for stereo matching between binocular event cameras and to align the RealSense camera's view with that of the left event camera. To achieve this, we place a standard checkerboard 1 meter in front of our prototype to ensure full visibility (see Fig. 4 (b)). We adopt a professional binocular stereo correction toolbox to correct the baseline alignment on RGB images from the two DAVIS346 cameras. Simultaneously, checkerboard keypoints are extracted from the RGB images of both the left DAVIS346 and the RealSense D455 cameras, and an affine transformation aligns the two coordinate sets [84].

Data Recordings and Statistics. Our Active-Event-Stereo dataset contains indoor and outdoor challenging scenarios (see Fig. 4 (c)) by considering velocity distribution, illumination change, scene diversity, varying distances, etc. 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 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. We compare our Active-Event-Stereo with the relevant camera prototype and representative datasets in Table 1. Notably, this is the first work to build an active event-based stereo vision dataset.

Method	Type	Camera	Resolution	Projector	Labels
Manasi [51]	Monocular	Gen3	640×480	Point	Sparse
Muglikar [52]	Monocular	Gen3	640×480	Point	Dense
Brandli [6]	Monocular	DVS128	128×128	Line	No
Wieland [49]	Monocular	Gen3	640×480	Line	No
Takatani [64]	Monocular	DAVIS346	346×260	Line	No
Leroux [37]	Monocular	ATIS	304×240	Pattern	No
Ashish [47]	Monocular	DAVIS346	346×260	Pattern	No
Huang [32]	Monocular	CeleX-V	1280×800	Pattern	No
Fu [20]	Monocular	EVK4	1280×720	Pattern	No
Bajestani [3]	Monocular	Gen3	640×480	Pattern	No
Li [41]	Monocular	DAVIS346	346×260	Pattern	No
Wang [71]	Monocular	DVXplorer	640×480	Pattern	Dense
Ours	**Stereo**	DAVIS346	346×260	Pattern	**Dense**

Table 1. Comparison with event-based vision systems using active infrared light and depth labels. The laser shape emitted by the infrared projector can be classified into point, line, and 2D patterns.

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 with *multiple characteristics*: (i) *High temporal resolution from event streams*; (ii) *Dynamic event generation with structured light even in static scenes or dark environments*; (iii) *Temporally long-term stereo event streams with depth labels at 15 Hz*; (iv) *Real-world recordings with abundant diversities in moving speed, light change, scene category, and distance variation*.

5. Methodology

5.1. Architecture Overview

This work aims at designing a novel lightweight yet effective active event-based stereo matching neural network, termed ***ActiveEventNet***, which generates high-speed dense disparity maps via integrating binocular event cameras and infrared structured light. As illustrated in Fig. 5, our framework mainly consists of four modules: *event representation*, *feature extraction*, ***dynamic interaction for cost volume*** and encoder-decoder. More precisely, the continuous event stream is first divided into event temporal bins, and each bin can be converted into a 2D image-like representation (i.e., event tensors [22]). Then, we introduce the lightweight MobileNet blocks [29, 62] to the corresponding 3D convolutions and show their necessity for event-based stereo matching models. The event embeddings in stereo pairs are fed into the feature extraction backbone and the channel reduction module to obtain compact yet powerful features. Moreover, we design a novel 3D cost volume via dynamically exchanging the channels and concatenating interaction stereo features, which refers to the costs of matching corresponding pixels between two event streams from slightly different viewpoints. Finally, the 3D cost volume is taken into an encoder-decoder module with a stack of

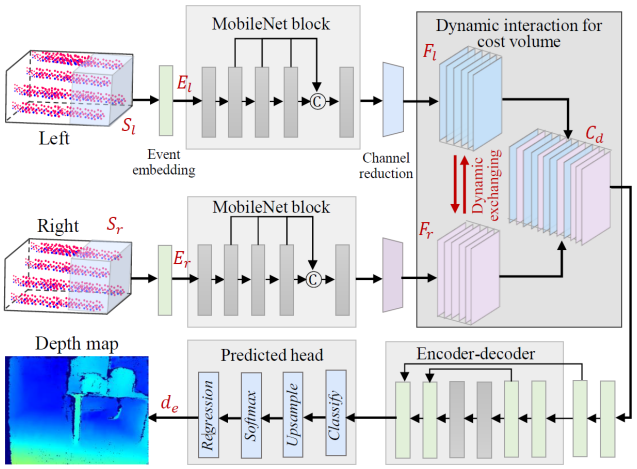


Figure 5. Overview of *active event-based stereo matching neural network (ActiveEventNet)*. Each event stream is first split into event temporal bins and encoded into event tensors [22]. Then, we incorporate MobileNet blocks for feature extraction and channel reduction. Meanwhile, we design a novel cost volume using dynamic interaction for similarity measurement in stereo pairs. Finally, an encoder-decoder component via a stack of lighter convolutional layers is utilized to predict dense disparity maps (along with a mask for enhanced visualization).

lighter convolutional layers to predict dense disparity maps.

5.2. Raising MobileNet for Event-based Stereo

To achieve a trade-off between fidelity and inference speed, we first incorporate the lightweight yet effective MobileNet blocks [29, 62, 63] instead of standard convolution operations for active event-based stereo matching. As a pioneering work, MobileNet v1 utilizes depthwise convolutions followed by pointwise convolutions to achieve standard convolution while reducing computational complexity. In general, MoileNet v1 achieves significant computation compared to standard convolutions by effectively utilizing the depth separable convolutions. Furthermore, MobileNet v2 introduces linear bottlenecks and inverted residuals to improve accuracy while keeping comparable memory-efficient inference. We formulate the output $\boldsymbol{F}_{res}$ of an inverted residual block as follows:

$$\begin{aligned}
\boldsymbol{F}_1 &= \sum_{c}^{tC} \boldsymbol{k}_p(x,y,c) \cdot \boldsymbol{F}(x,y,c), \\
\boldsymbol{F}_2 &= \sum_{x,y} \boldsymbol{k}_d(x,y,tc) \cdot \boldsymbol{F}_1(x-1,y-1,tc), \\
\boldsymbol{F}_3 &= \sum_{c} \boldsymbol{k}_p(x,y,c) \cdot \boldsymbol{F}_2(x,y,c), \\
\boldsymbol{F}_{\text{res}} &= \boldsymbol{F} + \boldsymbol{F}_3,
\end{aligned} \tag{4}$$

where $\boldsymbol{k}_d(x,y,c)$ denotes the depthwise convolution kernel at position (x,y,c) of the input feature map $\boldsymbol{F}$, and $\boldsymbol{k}_p$ is the 1×1 pointwise convolution kernel. $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, and $\boldsymbol{F}_3$ are intermediate feature maps in the inverted residual block. The channel dimension c is expanded with an expansion factor t in the pointwise and depthwise convolution operations.

Overall, our ActiveEventNet mainly replaces standard 2D convolutions of lightweight MobileNet v2 blocks in the feature extraction module and the encoder-decoder module.

5.3. Dynamic Interaction for Cost Volume

Cost volume [36, 57, 66] is a crucial component in the stereo matching pipeline, which is the matching cost between pixels at different disparities. A 3D cost volume $\boldsymbol{C_d}$ can be constructed via computing the dissimilarity between the left feature map $\boldsymbol{F}_l$ and the right feature map $\boldsymbol{F}_r$ as:

$$\boldsymbol{C}_d(x,y,d) = \mathcal{M}_c(\boldsymbol{F}_l(x,y,c), \boldsymbol{F}_r(x-d,y,c)), \tag{5}$$

where d is the disparity between pixels in a stereo pair, and $\mathcal{M}_c$ is a similarity measurement function. For learning-based models, $\mathcal{M}_d$ usually adopts concatenation or correlation operations [50, 53, 73] between two feature maps.

This work aims at designing a lightweight yet powerful 3D cost volume via dynamically exchanging the stereo channels and concatenating interacted features, which achieves satisfactory performance while maintaining comparable computational complexity to typical aggregation operations. As a result, we can mathematically describe the construction process of 3D cost volume as follows:

$$\begin{aligned}
\hat{\boldsymbol{F}}_l &= \sum_{c} \boldsymbol{F}_l(x,y,d-c) + \boldsymbol{W}_r \cdot \boldsymbol{F}_r(x,y,d-c), \\
\hat{\boldsymbol{F}}_r &= \sum_{c} \boldsymbol{F}_r(x,y,d-c) + \boldsymbol{W}_l \cdot \boldsymbol{F}_l(x,y,d-c), \\
\boldsymbol{C}_d &= \left[\hat{\boldsymbol{F}}_l(x,y,d), \hat{\boldsymbol{F}}_r(x,y,d)\right],
\end{aligned} \tag{6}$$

where $\hat{\boldsymbol{F}}_l$ and $\hat{\boldsymbol{F}}_r$ are the left and the right interacted feature maps after the dynamic interaction operation. $\boldsymbol{W}_l$ and $\boldsymbol{W}_r$ are weight matrices that determine how feature maps from the left and right views interact dynamically.

In this study, the scaling factor of the batch normalization (BN) layer reflects the importance of the feature map in the c-th channel. The feature map is dynamically interacted by the corresponding channel from the other view once the scaling factor is smaller than a presetting threshold θ_γ. For example, the left camera's feature map is replaced by the corresponding feature map $F'_{r,m,c}$ of the right camera as:

$$F'_{r,m,c} = \gamma_{r,m,c} \frac{F_{r,m,c} - \mu_{r,m,c}}{\sqrt{\sigma^2_{r,m,c} + \varepsilon}} + \beta_{r,m,c}, \ \gamma_{r,m,c} < \theta_\gamma, \tag{7}$$

where F_{rmc} is the c-th channel before the m-th BN layer in the right branch. ε is a small constant, and $\mu_{r,m,c}$ and $\sigma_{r,m,c}$ denote the mean and the standard deviation. $\gamma_{r,m,c}$ and $\beta_{r,m,c}$ are the trainable scaling factor and the offset.

Scenario	Sequence	RGB frames						Events					
		EPE↓	RMSE↓	D1-all↓	>1px↓	>2px↓	>3px↓	EPE↓	RMSE↓	D1-all↓	>1px↓	>2px↓	>3px↓
Normal light	05_indoor_boxes	1.889	7.988	0.066	0.359	0.114	0.066	1.941	7.972	0.069	0.397	0.129	0.069
	10_indoor_boxes	2.636	8.918	0.196	0.589	0.338	0.196	2.594	8.922	0.171	0.613	0.321	0.171
	26_indoor_checkerboard	2.048	7.9114	0.086	0.495	0.198	0.086	1.732	7.855	0.063	0.308	0.102	0.063
	70_outdoor_car	1.025	3.515	0.015	0.319	0.123	0.015	1.150	3.555	0.037	0.399	0.132	0.037
	80_outdoor_deer	2.421	8.217	0.16	0.589	0.318	0.16	2.316	8.122	0.107	0.624	0.285	0.107
Low light	15_indoor_office_desk_dark	3.312	12.018	0.158	0.0575	0.276	0.158	3.281	12.067	0.161	0.523	0.288	0.161
	17_indoor_office_desk_dark	3.453	11.805	0.212	0.631	0.364	0.212	2.894	11.756	0.115	0.414	0.196	0.114
	22_indoor_printer_dark	2.514	7.758	0.187	0.645	0.353	0.187	1.597	7.517	0.062	0.256	0.103	0.061
	27_indoor_checkerboard_dark	2.198	7.825	0.225	0.477	0.295	0.225	1.736	7.759	0.050	0.409	0.110	0.050
	32_indoor_conference_desk_dark	2.945	7.118	0.337	0.735	0.514	0.337	1.706	6.650	0.074	0.460	0.164	0.074
	58_indoor_washroom_dark	2.381	6.018	0.268	0.626	0.365	0.269	1.031	5.338	0.025	0.188	0.052	0.025
	64_outdoor_car_dark	1.088	3.673	0.019	0.398	0.087	0.019	0.839	3.591	0.011	0.217	0.044	0.011
All	**Average**	2.238	8.095	0.159	0.545	0.281	0.159	**1.995**	**7.821**	**0.083**	**0.399**	**0.163**	**0.082**

Table 2. Performance evaluation of our real-world Active-Event-Stereo dataset. Note that, our solution, combining binocular event cameras with an infrared pattern projector, outperforms conventional frame-based stereo vision, particularly in challenging low-light conditions.

Method	Event representation	Backbone	EPE ↓	RMSE ↓	D1-all ↓	>1px ↓	>2px ↓	>3px ↓	# Params. (M)	Runtime (ms)
SGM [28]	Reconstructed images	No learning	3.625	16.567	0.586	0.751	0.684	0.585	-	32.1
PSMNet [9]	Event images	2D CNN	2.894	11.756	0.204	0.603	0.352	0.204	5.22	15.6
DeepPruner-Fast [16]	Event images	2D CNN	2.514	8.758	0.113	0.545	0.253	0.112	7.39	39.4
AANet [77]	Event images	2D CNN	2.317	8.123	0.106	0.525	0.287	0.106	3.68	16.7
Unimatch [78]	Event images	Transformer	**1.902**	**7.759**	**0.068**	**0.361**	**0.116**	**0.066**	4.70	87.6
DDES [67]	Event embeddings	2D CNN	2.643	8.920	0.122	0.591	0.341	0.122	2.33	19.5
Our ActiveEventNet	Event images	MobileNet	1.995	7.821	0.083	0.399	0.163	0.082	**2.23**	**6.5**
Our ActiveEventNet*	Voxel grids	MobileNet	1.987	7.813	0.079	0.386	0.158	0.078	2.23	6.8
Our ActiveEventNet$^{\diamondsuit}$	Reconstructed images	MobileNet	1.972	7.780	0.076	0.382	0.156	0.076	2.23	7.0

Table 3. Comparison with state-of-the-art methods on our real-world Active-Event-Stereo dataset.

6. Experiment

6.1. Experimental Settings

Realistic Synthetic Dataset. To obtain a large amount of labor-saving yet high-quality synthetic data, we build an active event-based stereo matching simulated dataset, namely RealSense-Event-Sim. An Intel RealSense D435 sensor is first utilized to record 119 infrared video sequences that consider velocity distribution, light condition, scene diversity, etc. Then, we use the V2E simulator [31] to convert infrared videos into dynamic events. As a result, the newly built dataset offers event streams in stereo pairs and 23.8k synchronized true labels. Finally, we split them into 16k for training, 3.8k for validation, and 4k for testing.

Implementation Details. We select event images [22] as the event representation to achieve an accuracy-speed trade-off. We set the maximum disparity to 192 for stereo matching in all cases. We set the threshold θ_γ to 10^{-2} in dynamic interaction for cost volume. All networks are trained for 50 epochs using the Adam optimizer [34] on an NVIDIA 3090 GPU with a learning rate of 10^{-3}. For training losses, we utilize an L_1 loss to measure the absolute difference between the predicted maps and the labels.

Evaluation Metrics. Mean average end-point-error (EPE), root mean square error (RMSE), the percentage of pixels with disparity error than 3 pixels and $0.05\boldsymbol{d}_{gt}$ (D1-all), the percentage of pixels with disparity errors greater than 1 pixel, 2 pixels, and 3 pixels (i.e., >1px, >2px, and >3px) are used to evaluate the accuracy in the stereo matching task. The model parameters (#Params) and the running time (ms) are adopted to evaluate the computation speed.

6.2. Effective Test

Evaluation on RGB Frames and Events. To compare our solution with conventional frame-based stereo, we report quantization results for each test set sequence using both RGB frames and events with structured light on our Active-Event-Stereo dataset (see Table 2). We can find that our solution with structured light is able to acquire high-quality dense disparity maps in both indoor and outdoor scenarios with light changes. Our solution achieves better performance than passive stereo vision with RGB frames in most indoor and outdoor scenes, especially in low-light scenarios. More precisely, our solution significantly reduces errors across five metrics compared to passive stereo vision, with EPE, RMSE, and D1-all decreased by 0.243, 0.274, and 0.076, respectively. To our surprise, our method using sparse events outperforms dense RGB images in most sequences, even under normal light scenes, with a few recordings showing comparable results. This may be the incorporation of structured light in stereo vision, which enhances scene textures and further boosts performance.

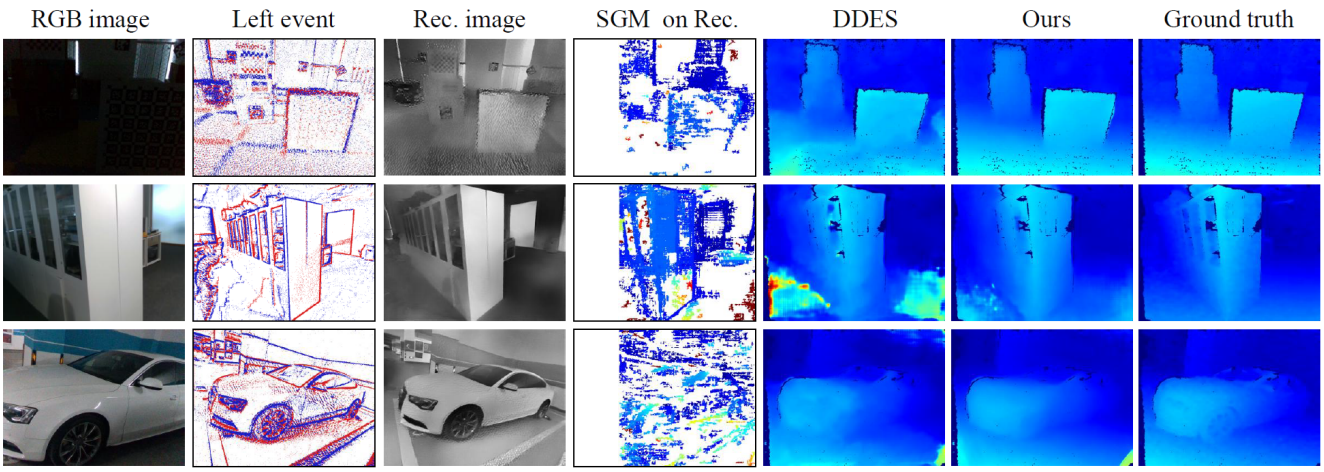


Figure 6. Representative examples of different stereo matching results on our real-world Active-Event-Stereo dataset. To enhance visualization and comparison, we implement a mask to the void areas of the ground truth to the predicted dense maps.

Method	EPE ↓	RMSE ↓	D1-all ↓	Runtime (ms)
SGM [28]	4.298	8.326	0.681	43.2
PSMNet [9]	2.786	5.583	0.211	20.6
DeepPruner-Fast [16]	1.463	2.513	0.087	76.0
AANet [77]	1.397	2.463	0.077	28.5
Unimatch [78]	**1.108**	**1.78**	**0.063**	172.6
DDES [67]	1.696	3.241	0.125	36.4
Our ActiveEventNet	1.223	2.320	0.070	**21.9**

Table 4. Comparison with state-of-the-art methods on our highly synthetic RealSense-Event-Sim dataset.

Method	Baseline	(a)	(b)	**Ours**
MobileNet blocks		✓		✓
Dynamic interaction cost volume			✓	✓
EPE ↓	2.124	2.21	**1.968**	1.993
RMSE ↓	7.935	7.952	**7.750**	7.821
D1-all ↓	0.092	0.095	**0.075**	0.083
Runtime (ms)	33.8	**6.2**	35.7	6.5

Table 5. The contribution of each component to our ActiveEventNet on our real-world Active-Event-Stereo dataset.

Comparison with State-of-the-Art Methods. To make a comparison with stereo matching methods as fair as possible, we first convert event streams to videos using the E2VID [60] and then use reconstructed gray images as the input of the classical SGM [28]. In addition, we compare our ActiveEventNet with four frame-based stereo matching networks (i.e., PSMNet [9], DeepPruner-Fast [16], AANet [77], and Unimatch [78]) and a popular event-based stereo matching framework (i.e., DDES [67]). For *real-world dataset evaluation*, we compare our ActiveEventNet with other methods in our Active-Event-Stereo dataset in Table 3. Note that, our ActiveEventNet achieves superior performance compared to five state-of-the-art methods while maintaining smaller parameters and faster inference times. Compared to the top-performing Transformer-based Unimatch [78], our approach delivers comparable performance with a 12× improvement in inference speed. Furthermore, we compare three typical event representations (i.e., event images [22], voxel grids [89], and reconstructed images [60]) to verify the generality of our method for various event representations. It indicates that this improvement in event representation comes with an associated increase in computational speed. Furthermore, we present some representative examples of visualization comparison results on our Active-Event-Stereo dataset in Fig. 6. Apparently, the conventional model-based SGM only generates sparse disparity maps, but our ActiveEventNet excels in obtaining high-quality disparity maps, even under varying light conditions in both indoor and outdoor scenes. For *simulated dataset evaluation*, we report quantization results in Table 4, showing that the conclusions from the simulation dataset are consistent with those from the real dataset.

6.3. Ablation Test

Contribution of Each Component. To explore the impact of each component on the final performance, our baseline uses standard convolutions and adopts a typical concatenation operation in the cost volume module. As shown in Table 5, three methods, namely (a), (b), and our ActiveEventNet, consistently indicate that leveraging MobileNet blocks boosts inference speed, while the introduction of the dynamic interaction cost volume enhances accuracy. Besides, our approach, using dynamic interaction for cost volume, obtains a 0.156 reduction in EPE while keeping a comparable computation speed. To our surprise, our method, employing MobileNet blocks to replace standard convolutions, achieves a nearly 6× increase in inference speed.

Influence of MobileNet Blocks. To analyze MobileNet blocks in our ActiveEventNet, we deploy MobileNet v2 blocks instead of standard convolutions in the feature ex-

Feature extraction	Encoder-decoder	EPE↓	D1-all↓	Runtime (ms)
Standard Conv.	Standard Conv.	**1.968**	**0.075**	35.7
Standard Conv.	MobileNet v2	1.972	0.076	22.2
MobileNet v2	Standard Conv.	1.983	0.078	13.4
MobileNet v2	MobileNet v2	1.993	0.083	**6.5**

Table 6. The influence of MobileNet blocks in Our ActiveEventNet on our real-world Active-Event-Stereo dataset.

Method	EPE ↓	RMSE ↓	D1-all ↓	Runtime (ms)
Concatenation	2.527	8.823	0.113	**5.8**
Correlation [53]	2.316	8.124	0.107	6.2
Dynamic Interaction	**1.993**	**7.821**	**0.083**	6.5

Table 7. Comparison of our ActiveEventNet with various aggregation strategies of cost volume on our Active-Event-Stereo dataset.

Setup	EPE ↓	RMSE ↓	D1-all ↓	Runtime (ms)
Without structured light	2.614	8.875	0.197	6.3
Structure light	**1.993**	**7.821**	**0.083**	6.5

Table 8. The impact of structured light on event-based stereo vision. Some static and slow-motion sequences are evaluated with and without structured light under the same scene.

traction and the encoder-decoder modules. As shown in Table 6, Our approach, introducing MobileNet v2 blocks, consistently achieves faster computational speed. Comparing MobileNet v2 and standard convolution, the absolute decrease in EPE is only 0.025, while the inference time is reduced by nearly 6×. It indicates that our approach using MobileNet v2 instead of standard convolution achieves an accuracy-speed trade-off for practical applications.

Influence of Cost Volume Construction. To evaluate the effectiveness of dynamic interaction for cost volume, we compare it with some typical aggregation operations in Table 7. Note that, our approach obtains the best performance against two aggregation operations (i.e., concatenation and correlation [53]). This is due to our dynamic interaction strategy, which effectively exchanges stereo channels and aggregates interacted features to enhance performance while keeping comparable computational speed.

6.4. Scalability Test

Analyzing the Role of Structured Light. To evaluate the impact of structured light on event-based stereo vision, we selected static and slow-motion sequences for quantitative assessment. As illustrated in Table 8, the stereo matching performance significantly improves by incorporating structured light. Besides, we present a representative instance in an extremely slow motion scenario (see Fig. 7). While the passive stereo generates almost no events, our camera prototype excels in producing dynamic events through the use of structured light. In other words, the solution, integrating the structured light for binocular event cameras, can overcome the limitation of most existing passive event-based stereo in static scenarios or dark environments.

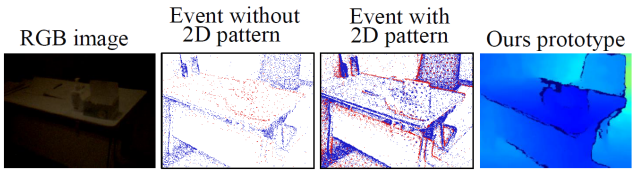


Figure 7. Representative instance of our camera prototype in extreme slow motion scenes. Unlike passive vision, our solution with structured light produces dynamic events even in static scenes.

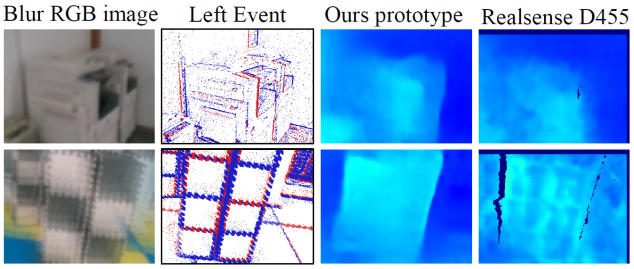


Figure 8. Comparison with a conventional active stereo camera in high-speed motion scenarios. Note that, our camera prototype outperforms Realsense D455 for high-speed depth sensing.

Test Camera Prototype in High-Speed Scenes. To verify our solution for high-speed depth sensing, we compare our camera prototype with a conventional RealSense D455. As shown in Fig. 8, our prototype empowers the acquisition of high-quality disparity maps in high-speed scenes, while RealSense D455 at 90 FPS is notably ineffective. In fact, conventional images from the RealSense D455 may suffer from motion blur, which impairs depth perception in high-speed scenes. In contrast, our solution leverages the event camera with high temporal resolution and structured light to improve scene texture, resulting in superior depth accuracy. Furthermore, our ActiveEventNet efficiently processes stereo event stream pairs, achieving an inference speed of up to 150 FPS on an NVIDIA 3090 GPU.

7. Conclusion

This paper presents a novel paradigm for high-speed depth sensing, called *active event-based stereo vision*, which first integrates binocular event cameras and active infrared structured light. Towards this end, we establish a real-world dataset using our active event-based stereo camera prototype, along with a highly realistic synthetic dataset. Then, we design a lightweight yet effective event-based stereo matching model, which significantly reduces the computational cost meanwhile keeping the comparable accuracy. We believe that our standardized datasets will open up an opportunity for the research of this challenging problem.

Acknowledgments. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102801, and the National Natural Science Foundation of China under Grant 61827804.

References

[1] Soikat Hasan Ahmed, Hae Woong Jang, SM Nadim Uddin, and Yong Ju Jung. Deep event stereo leveraged by event-to-image translation. In *AAAI*, pages 882–890, 2021. 1, 2

[2] Alexander Andreopoulos, Hirak J Kashyap, Tapan K Nayak, Arnon Amir, and Myron D Flickner. A low power, high throughput, fully event-based stereo system. In *CVPR*, pages 7532–7542, 2018. 1

[3] Seyed Ehsan Marjani Bajestani and Giovanni Beltrame. Event-based rgb sensing with structured light. In *WACV*, pages 5458–5467, 2023. 1, 4

[4] Luca Bartolomei, Matteo Poggi, Fabio Tosi, Andrea Conti, and Stefano Mattoccia. Active stereo without pattern projector. In *CVPR*, pages 18470–18482, 2023. 1

[5] Luca Bartolomei, Matteo Poggi, Andrea Conti, and Stefano Mattoccia. Lidar-event stereo fusion with hallucinations. In *ECCV*, pages 125–145, 2025. 2

[6] 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, 4

[7] Bharatesh Chakravarthi, Aayush Atul Verma, Kostas Daniilidis, Cornelia Fermuller, and Yezhou Yang. Recent event camera innovations: A survey. *arXiv*, 2024. 1

[8] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *CVPRW*, pages 4016–4023, 2023. 2

[9] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 6, 7

[10] Wu Chen, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Event-based stereo depth estimation by temporal-spatial context learning. *IEEE SPL*, 2024. 2

[11] Stefano Chiavazza, Svea Marie Meyer, and Yulia Sandamirskaya. Low-latency monocular depth estimation using event timing on neuromorphic hardware. In *CVPR*, pages 4071–4080, 2023. 1

[12] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *AAAI*, pages 454–462, 2022. 2

[13] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *ECCV*, pages 470–486, 2022. 2

[14] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *CVPR*, pages 17797–17807, 2023. 1, 2

[15] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *ECCV*, pages 294–314, 2025. 2

[16] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *CVPR*, pages 4384–4393, 2019. 6, 7

[17] Davide Falanga, Suseong Kim, and Davide Scaramuzza. How fast is too fast? the role of perception latency in high-speed sense and avoid. *IEEE RAL*, 4(2):1884–1891, 2019. 1

[18] Davide Falanga, Kevin Kleber, and Davide Scaramuzza. Dynamic obstacle avoidance for quadrotors with event cameras. *Sci. Robot.*, 5(40):eaaz9712, 2020. 1

[19] Sean Ryan Fanello, Julien Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir Tankovich, Philip Davidson, and Shahram Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *CVPR*, pages 6535–6544, 2017. 1

[20] 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. 1, 4

[21] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, et al. Event-based vision: A survey. *IEEE TPAMI*, 44(1):154–180, 2020. 1, 3

[22] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *ICCV*, pages 5633–5643, 2019. 4, 5, 6, 7

[23] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE RAL*, 6(3):4947–4954, 2021. 2

[24] Suman Ghosh and Guillermo Gallego. Event-based stereo depth estimation: A survey. *arXiv*, 2024. 1

[25] 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. 3

[26] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *CVPR*, pages 17804–17813, 2022. 1

[27] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *3DV*, pages 534–542, 2020. 1

[28] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE TPAMI*, 30(2):328–341, 2007. 6, 7

[29] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, 2017. 4, 5

[30] Rui Hu, Jürgen Kogler, Margrit Gelautz, Min Lin, and Yuanqing Xia. A dynamic calibration framework for the event-frame stereo camera system. *IEEE RAL*, 2024. 2

[31] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *CVPRW*, pages 1312–1321, 2021. 6

[32] 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. 1, 3, 4

[33] Hae-Gon Jeon, Joon-Young Lee, Sunghoon Im, Hyowon Ha, and In So Kweon. Stereo matching with color and monochrome cameras in low-light conditions. In *CVPR*, pages 4086–4094, 2016. 1

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

[35] Kazutoshi Kodama, Yusuke Sato, Yuhi Yorikado, Raphael Berner, Kyoji Mizoguchi, Takahiro Miyazaki, Masahiro Tsukamoto, Yoshihisa Matoba, Hirotaka Shinozaki, Atsumi Niwa, et al. 1.22 μm 35.6 mpixel rgb hybrid event-based vision sensor with 4.88 μm-pitch event pixels and up to 10k event frame rate by adaptive control on event sparsity. In *ISSCC*, pages 92–94, 2023. 3

[36] 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. 1, 5

[37] T Leroux, S-H Ieng, and Ryad Benosman. Event-based structured light for depth reconstruction using frequency tagged light patterns. *arXiv*, 2018. 4

[38] Jianing Li and Yonghong Tian. Recent advances in neuromorphic vision sensors: A survey. *Chinese J. Comput.*, 44 (6):1258–1286, 2021. 1

[39] Jianing Li, Yihua Fu, Siwei Dong, Zhaofei Yu, Tiejun Huang, and Yonghong Tian. Asynchronous spatiotemporal spike metric for event cameras. *IEEE TNNLS*, 34(4):1742–1753, 2021. 1

[40] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE TIP*, 31:2975–2987, 2022. 1

[41] 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. 1, 4

[42] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. *IEEE JSSC*, 43(2):566–576, 2008. 1, 3

[43] Peigen Liu, Guang Chen, Zhijun Li, Huajin Tang, and Alois Knoll. Learning local event-based descriptor for patch-based stereo matching. In *ICRA*, pages 412–418, 2022. 1, 2

[44] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *arXiv*, 2024. 1

[45] Xu Liu, Jianing Li, Jinqiao Shi, Xiaopeng Fan, Yonghong Tian, and Debin Zhao. Event-based monocular depth estimation with recurrent transformers. *IEEE TCSVT*, 2024. 1

[46] Antonio Loquercio, Elia Kaufmann, René Ranftl, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Learning high-speed flight in the wild. *Sci. Robot.*, 6(59):eabg5810, 2021. 1

[47] Ashish Rao Mangalore, Chandra Sekhar Seelamantula, and Chetan Singh Thakur. Neuromorphic fringe projection profilometry. *IEEE SPL*, 27:1510–1514, 2020. 4

[48] Julien NP Martel, Jonathan Müller, Jörg Conradt, and Yulia Sandamirskaya. An active approach to solving the stereo matching problem using event-based sensors. In *ISCAS*, pages 1–5, 2018. 2

[49] 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. 4

[50] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *CVPR*, pages 4258–4267, 2021. 2, 5

[51] Manasi Muglikar, Guillermo Gallego, and Davide Scaramuzza. Esl: Event-based structured light. In *3DV*, pages 1165–1174, 2021. 1, 2, 4

[52] Manasi Muglikar, Diederik Paul Moeys, and Davide Scaramuzza. Event guided depth sensing. In *3DV*, pages 385–393, 2021. 1, 3, 4

[53] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *CVPR*, pages 6114–6123, 2022. 1, 2, 5, 8

[54] Atsumi Niwa, Futa Mochizuki, Raphael Berner, Takuya Maruyarma, Toshio Terano, Kenichi Takamiya, Yasutaka Kimura, Kyoji Mizoguchi, Takahiro Miyazaki, Shun Kaizu, et al. A 2.97 μm-pitch event-based vision sensor with shared pixel front-end circuitry and low-noise intensity readout mode. In *ISSCC*, pages 4–6, 2023. 3

[55] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *CVPR*, pages 6820–6829, 2019. 1

[56] Xin Peng, Ling Gao, Yifu Wang, and Laurent Kneip. Globally-optimal contrast maximisation for event cameras. *IEEE TPAMI*, 44(7):3479–3495, 2021. 3

[57] 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. 1, 5

[58] 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. 3

[59] Ulysse Rançon, Javier Cuadrado-Anibarro, Benoit R Cottereau, and Timothée Masquelier. Stereospike: Depth learning with a spiking neural network. *IEEE Access*, 2022. 2

[60] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, pages 3857–3866, 2019. 7

[61] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 43(6):1964–1980, 2019. 1

[62] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 4, 5

[63] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *WACV*, pages 2417–2426, 2022. 5

[64] 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. 4

[65] 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. 3

[66] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *NeurIPS*, 31, 2018. 5

[67] 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. 1, 2, 6, 7

[68] SM Nadim Uddin, Soikat Hasan Ahmed, and Yong Ju Jung. Unsupervised deep event stereo for depth estimation. *IEEE TCSVT*, 32(11):7489–7504, 2022. 1, 2

[69] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE TIP*, 31:7237–7251, 2022. 1

[70] Guijin Wang, Chenchen Feng, Xiaowei Hu, and Huazhong Yang. Temporal matrices mapping-based calibration method for event-driven structured light systems. *IEEE Sensor J.*, 21 (2):1799–1808, 2020. 1

[71] 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. 1, 3, 4

[72] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Joint framework for single image reconstruction and super-resolution with an event camera. *IEEE TPAMI*, 44(11):7657–7673, 2021. 3

[73] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. In *NeurIPS*, pages 4835–4845, 2020. 5

[74] Yixuan Wang, Jianing Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Learning stereo depth estimation with bio-inspired spike cameras. In *ICME*, pages 1–6, 2022. 2

[75] Ziyi Wu, Mathias Gehrig, Qing Lyu, Xudong Liu, and Igor Gilitschenski. Leod: Label-efficient object detection for event cameras. In *CVPR*, pages 16933–16943, 2024. 1

[76] Tianyi Xiong, Jiayi Wu, Botao He, Cornelia Fermuller, Yiannis Aloimonos, Heng Huang, and Christopher A Metzler. Event3dgs: Event-based 3d gaussian splatting for fast egomotion. *arXiv*, 2024. 3

[77] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020. 6, 7

[78] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 6, 7

[79] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *CVPR*, pages 13924–13934, 2023. 3

[80] Zheyu Yang, Taoyi Wang, Yihan Lin, Yuguo Chen, Hui Zeng, Jing Pei, Jiazheng Wang, Xue Liu, Yichun Zhou, Jianqiang Zhang, et al. A vision chip with complementary pathways for open-world sensing. *Nature*, 629(8014):1027–1033, 2024. 3

[81] Bohan Yu, Jieji Ren, Jin Han, Feishi Wang, Jinxiu Liang, and Boxin Shi. Eventps: Real-time photometric stereo using an event camera. In *CVPR*, pages 9602–9611, 2024. 1

[82] Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, and Luziwei Leng. Discrete time convolution for fast event-based stereo. In *CVPR*, pages 8676–8686, 2022. 1, 2

[83] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Generalizing event-based motion deblurring in real-world scenarios. In *ICCV*, pages 10734–10744, 2023. 1

[84] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000. 4

[85] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *ECCV*, pages 235–251, 2018. 1, 2

[86] Yi Zhou, Guillermo Gallego, and Shaojie Shen. Event-based stereo visual odometry. *IEEE TRO*, 37(5):1433–1450, 2021. 1

[87] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Realtime time synchronized event-based stereo. In *ECCV*, pages 433–447, 2018. 1, 2

[88] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE RAL*, 3(3):2032–2039, 2018. 2

[89] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, pages 989–997, 2019. 1, 7

[90] Yi-Fan Zuo, Jiaqi Yang, Jiaben Chen, Xia Wang, Yifu Wang, and Laurent Kneip. Devo: Depth-event camera visual odometry in challenging conditions. In *ICRA*, pages 2179–2185, 2022. 3