Stereo Hybrid Event-Frame (SHEF) Cameras for 3D Perception

Ziwei Wang1, Liyuan Pan∗,2,3, Yonhon Ng∗,1, Zheyu Zhuang1 and Robert Mahony1

Abstract—Stereo camera systems play an important role in robotics applications to perceive the 3D world. However, conventional cameras have drawbacks such as low dynamic range, motion blur and latency due to the underlying frame-based mechanism. Event cameras address these limitations as they report the brightness changes of each pixel independently with a fine temporal resolution, but they are unable to acquire absolute intensity information directly. Although integrated hybrid event-frame sensors (e.g., DAVIS) are available, the quality of data is compromised by coupling at the pixel level in the circuit fabrication of such cameras. This paper proposes a stereo hybrid event-frame (SHEF) camera system that offers a sensor modality with separate high-quality pure event and pure frame cameras, overcoming the limitations of each separate sensor and allowing for stereo depth estimation. We provide a SHEF dataset targeted at evaluating disparity estimation algorithms and introduce a stereo disparity estimation algorithm that uses edge information extracted from the event stream correlated with the edge detected in the frame data. Our disparity estimation outperforms the state-of-the-art stereo matching algorithm on the SHEF dataset.

Code, Datasets and Video:
https://github.com/ziweiWWANG/SHEF.git

I. INTRODUCTION

Event cameras based on bio-inspired retina vision sensor, such as Dynamic Vision Sensor (DVS) [1] asynchronously measure per-pixel log intensity changes with a fine time resolution (<10μs), low latency (<0.5ms), high dynamic range (>120dB vs 60dB of conventional cameras), and low power consumption (<0.1W) [1], [2], [3]. These properties make event cameras of significant interest in robotics applications where the robustness of the sensing system is crucial for challenging scenarios with fast motion and high contrast scenes. However, event cameras do not capture the absolute intensity and static image information. There is considerable potential for robotic systems to exploit the complementary characteristics of both event and frame-based cameras, motivating the consideration of hybrid and stereo hybrid event-frame sensor systems. The dynamic and active pixel vision sensor (DAVIS) [2] combines the DVS and an active pixel sensor (APS) circuit. By sharing the same photodiode, DAVIS provides registered event data and frames at the pixel level that allows easy access to dual event-frame data. However, the coupling between APS and DVS components generates biased event noise triggered by the shutter of each frame [2]. This hybrid configuration problem is significant, and limits the quality of both event and frame data. Currently, the highest quality of the hybrid event-frame camera is DAVIS346 with a limited resolution of 346 × 260 in both event data and frames. In contrast, pure event cameras have achieved 1280 × 960 resolution in Samsung Gen4 and are developing more quickly than the hybrid event-frame sensors. The stereo hybrid event-frame sensor modality offers significant performance advantages as well as allowing simple integration into existing sensor suites, such as the multi-camera systems commonly found on modern mobile phones and robotic systems. In addition, adding event sensors to existing robotic systems can compensate for failure of an existing conventional frame camera in low-light or high-velocity scenarios. However, in order to exploit event data with existing systems, it is necessary to estimate disparity between the separate event sensor and the existing frame sensor. Such a disparity, and consequently depth, estimate will also provide additional sensing capability for future robotic systems.

To the authors’ best knowledge, this paper is the first work that tackles the problem of fusing event and frame data (separate sensing modalities) to compute stereo disparity. In this paper, we introduce the first dataset targeted specifically at stereo hybrid event-frame data for near field scenes and a baseline algorithm for computing disparity. Our stereo hybrid...
event-frame (SHEF) camera system consists of a separate event camera and a frame-based camera alongside each other. We mount the system rigidly on a UR5 robot manipulator to get the controlled motion and accurate camera poses. Then we build the ground truth 3D environment reconstruction using a sequence of RGB images with known camera positions from the UR5. Our baseline (edge-based) disparity estimation algorithm computes cross-correlation between event data and frames with a coarse-to-fine framework. As an indication of the potential of our system, we also train a Disparity Completion Net (DCNet) to interpolate and optimize the sparse disparity map to generate a dense map, taking advantage of the absolute intensity information from the frame data. Our dataset and code will be made available online for future studies and comparisons.

In summary, the contributions of the paper are:

- Formulate the multi-modality stereo hybrid event-frame depth estimation problem and introduce a baseline disparity estimation algorithm.
- Introduce the first open-source high-quality stereo hybrid event-frame dataset targeted specifically at depth estimate for near field scenes, including event data, high frame rate RGB image sequences, accurate camera poses, point clouds of 3D environment reconstructions and ground truth depth maps.

II. RELATED WORK

Since the introduction of integrated hybrid event-frame cameras, in particular the DAVIS cameras, many datasets that incorporate both event and frame data are developed. These datasets target a wide range of applications, such as optical flow [4], tracking [5], [6], [7], video reconstruction [8], [9], [10], [11], [12] and deblurring [13], [14]. However, the hybrid DAVIS cameras generate biased event noise triggered by the shutter of each frame when operating in hybrid mode. Some researchers provide datasets with separate event-frame sensors. Wang et al. [15] propose a hybrid system with a DAVIS 240b event camera (resolution of 180 × 190, operating in pure event mode) and a FLIR RGB frame-based camera. Events and frames are registered through a beam-splitter. The system overcomes the inherent noise of DAVIS and acquires high-resolution frames from a separate commercial frame sensor. Despite having two separate sensors, the system is effectively a monocular set-up and it only provides relatively low-resolution event data from the DAVIS 240b event camera. Along with other on-board sensors, stereo DA VIS system proposed by Zhu et al. [16] provide stereo event data, stereo frames, camera poses, reference depth images from embedded GPS, VI sensor and Lidar. Because the system provides depth ground truth, many recent stereo event-based 3D perception works are evaluated on the released dataset of this system, e.g., [17], [18], [19]. However, this dataset is inherently affected by bias and shutter noise due to the use of DAVIS’s hybrid event-frame mode [2]. Furthermore, in the widely-used indoor flying sequences for the 3D perception tasks [16], the event data is sparse and heavily biased (the ratio of positive to negative events is around 2.5-5) [16]. Recent datasets that use separate event and frame cameras [20], [21], [22], [12] have been shown to lead to higher quality depth estimation and image reconstruction.

A number of authors have considered stereo event-event and monocular “multi-view” event depth estimation as well as using event data to augment frame based stereo or multi-view depth estimation. Event-driven stereo matching methods were proposed in [23] and [24]. The disparity uniqueness and depth continuity in stereo event depth estimation has also been considered by [25] and [26]. Osswald et al. [27] propose a spiking stereo neural network to compute disparity that can be implemented in parallel. Zhou et al. [17] compute a semi-dense depth map from stereo events at multiple viewpoints. Monocular multi-view event-based depth estimation algorithms are usually used to solve SLAM and visual odometry tasks [28], [29], [30], computing relative camera poses and 3D reconstructions over time. Mitrokhin et al. [31] present a dataset with accurate motion masks and ground truth depth that can be used in monocular event-based depth estimation. To exploit the advantage of different sensing modalities, Hadviger et al. [19] compute dense disparity from frames, and predict future disparity with visual odometry, then update disparity using event-based optical flow.

III. STEREO HYBRID EVENT-FRAME DATASET

In this section, we introduce the proposed SHEF system and provide details of the dataset. It covers the sensor parameters, sampling rate, hardware set up, three dataset scenes, camera motion control, system calibration, 3D reconstruction model and depth ground truth.

A. Methodology

Sensors: We build a stereo hybrid event-frame camera system with a Prophesee event camera (VGA, 640 × 480 pixels) and a FLIR RGB camera (Chameleon3USB3, 2048 × 1536 pixels, 54fps) mounted side-by-side on a camera rig. This rig is then
Camera motions: The camera motion is controlled by servoing UR5’s end-effector to follow the pre-defined smooth trajectories to 0.1mm precision [32]. The camera trajectories in the dataset consist of, a circle, and three ‘8’ shaped Lissajous trajectories. The three ‘8’ shaped Lissajous trajectories include a planer trajectory parallel to the cameras’ image plane, one with moderate variation along the cameras’ optical axis, and a complex trajectory with significant variation along the cameras’ optical axis. In addition, a square trajectory of four pure translations is included. Since using a robotic arm increases the reproducibility of the experiment, we are able to perform each trajectory six times at three different speeds, for two different lighting conditions and each of the three scenes described above. The orientation of the stereo camera rig remains unchanged in all trajectories. Images used to reconstruct the ground truth 3D scene are captured from multiple views to reduce the effect of occlusion.

B. Calibration

Stereo camera calibration: We calibrate the intrinsic and extrinsic parameters of SHEF using the MATLAB stereo camera calibrator [33]. To continuously trigger events for calibration, we move a blinking checkerboard in front of the system and reconstruct intensity images using the event-based high pass filter [9]. We synchronize the two cameras by sending a trigger signal from the frame camera to the event camera.

Hand-eye calibration: The SHEF rig is mounted on the end-effector of the UR5 manipulator. A hand-eye calibration procedure is performed to compute the rigid-body transformation from the frame camera to the end-effector so that the accurate camera poses with respect to the robot base can be directly computed from the manipulator’s forward kinematics model. During the calibration procedure, the robot end-effector, i.e the SHEF rig, is moved to observe a static flat calibration pattern from various viewpoints.

Denote the 6 DoF poses of the end-effector frame \( \{H\} \), robot base frame \( \{W\} \), and the calibration pattern \( \{T\} \) as elements in the Special Euclidean Group \( \text{SE}(3) \). Then denote the pose of an arbitrary frame \( \{B\} \) with respect to an arbitrary
reference frame \{A\} as $^A X_B$. The pose of the calibration pattern with respect to the robot base can be expressed as $^W X_T = ^W X_H ^H X_C ^C X_T$. The same pose can be expressed from different viewpoints as

$$^W X_H ^H X_C ^C X_T = ^W X_H ^H X_C ^C X_T,$$  

(1)

where $i, j \in \{1, 2, 3, \ldots\}$ represent the index of different viewpoints. By rearranging Eq.(1), we have

$$A_{ij}^H X_C = ^H X_B B_{ij},$$  

(2)

where $A_{ij} = ^W X_H ^W X_H$, $B_{ij} = C_{ij} X_P^C X_P^{-1}$. Both $A_{ij}$ and $B_{ij}$ are known; the former is computed from the forward kinematics model of the UR5 manipulator, and the latter is estimated from image frames. The transformation matrix from the camera to the robot base is estimated from Eq.(2) using the hand-eye calibration method proposed in [34].

**Ground truth:** The dataset requires the ground truth depth map associated with each frame. For each dataset environment, we use the FLIR camera to take around 100 RGB images from different views and compute camera poses from the manipulator. Then we build a dense point cloud and a 3D reconstruction model for each setup (see Fig. 2) using a commercial software 3DFLOW Zephyr [37]. With the known camera pose and intrinsic parameter for each frame in our dataset, we compute the camera ray for each pixel in the frame and find the corresponding 3D point along the ray. Then, we transform the 3D point cloud into a 2D ground truth depth map for each frame.

### IV. STEREO DISPARITY ESTIMATION

Different from the standard stereo frame-based disparity estimation problem, our stereo hybrid event-frame system provides an event stream at one view and an intensity frame at the other view. As the sensing modality of the event data is different from its corresponding intensity frame data, correspondence-based algorithms that rely on matching cost functions cannot be applied directly. In this section, we provide a stereo disparity estimation baseline algorithm that correlates the edge information extracted from event data and frames. A coarse-to-fine framework is also applied to compute denser and smoother disparity.

**Binary edge images:** To estimate the disparity of event and frame image at a certain frame timestamp $t$, we first use high pass filter [9] to reconstruct intensity images from events at time $t$ with a relatively high cut-off frequency (120 rad/s), resulting in a short filter memory and only preserving high frequency edge information in a deterministic and asynchronous manner. On the reconstructed image, we perform non-maximal suppression to generate a binary edge map $E_b$. The binary edge maps are generated at the same timestamp as the paired frame camera. For the frame data, we compute an edge map using classical $3 \times 3$ Sobel filters and then apply non-maximal suppression to generate the corresponding binary edge map $E_f$ (Fig. 3).

**Cross-correlation:** We slide a window along the epipolar line in the right-hand binary edge image and calculate the cross-correlation between the reference window in the left-hand binary edge image. By minimizing the matching cost, we get a sparse disparity map $D_s \in \mathbb{R}^{H \times W}$. Here, $H$ and $W$ are the image height and width.

We estimate the disparity of a pixel $x = (x, y)$ by solving,

$$D_s(x) = \min_{\triangle x} \lambda \{e(x + \triangle x)\}.$$  

(3)

The matching cost $e(\cdot)$ is given by,

$$e(x + \triangle x) = \sum_{i, j \in \mathcal{W}} L_b(x + i, y + j) \odot E_b(x + i + \triangle x, y + j),$$  

(4)

where $D_s$ denotes the sparse disparity map, $(i, j)$ are the pixels in the sliding window $\mathcal{W}$, and $\lambda = -1$ is the weight parameter. Here, $\triangle x$ is the estimated disparity.

**Coarse-to-fine framework:** To make our sparse disparity map denser and smoother, we implement a traditional coarse-to-fine framework using an image pyramid with 5 levels and a scale factor of 0.9. Disparity estimation then operates coarse-to-fine over the defined pyramids by

$$D_{k}^p = I_1 D_{k+1}^p + I_2 \min_{\triangle x} \lambda \{e(x + \triangle x)\}_{k},$$  

(5)

where $k \in \{1, \ldots, 5\}$ is the pyramid level, $\uparrow$ denotes upsampling by the scale factor, and $I_1$ and $I_2$ are weight matrices. We represent the improve disparity map as $D_p$. Note that, $I_1 \in \mathbb{R}^{H \times W}$, where $I_1 + I_2 = 1$ and all elements in the two matrices are in the set $\{0, 0.5, 1\}$. Finally, we remove disparity outliers that do not have sufficient surrounding support pixels, where the local estimated disparity should be similar.

### V. EXPERIMENT

**A. Experimental Setup and Main Results.**

We evaluate our disparity estimation algorithm on three scenes: *simple boxes*, complex boxes and *picnic*. For each scene, we evaluate on 50 event-frame image pairs under the same motion speed (fast), lighting condition (bright) and trajectory (circle). The algorithm is run offline in batch mode. To evaluate the quality of the disparity map, we first compute the ground truth disparity $D = \frac{bf}{d}$, where $b, f, d$ denotes the camera baseline, focal length and depth respectively.

**Comparison:** Since there is no similar existing dataset and algorithm, to prove the effectiveness of our disparity estimation algorithm, we compare our method with the state-of-the-art stereo frame-based matching Adaptive Aggregation Network (AANet) [35] using our dataset. We first use the state-of-the-art pure-event reconstruction method ECNN [36] to generate an event image to each corresponding intensity image, and then estimate dense disparity from the stereo image pairs using the AANet. We represent the dense disparity map computed by AANet as $D_{AA}$.

The quality of sparse $D_s$ using cross-correlation and the improved $D_p$ with an image pyramid are separately evaluated against the ground-truth; only the pixels with estimated
disparity are considered. We evaluate the disparity of $D_{AA}$ on the edge pixels of $D_p$ in Table II and compare the dense disparity $D_{AA}$ with our completed dense disparity $D_c$ (discussed in §V-B) in Table III. In the quantitative evaluation of all methods, we ignore pixels on the right and left image boundaries because they do not have corresponding matching pixels in the other stereo image pair.

**Evaluation metrics:** We evaluate the disparity estimation quality using three metrics: RMSE (Root-Mean-Square Error), bad-p and the inlier ratios [38]. ‘bad-p’ is the standard metric of KITTI [39] which computes the percentage of bad pixels averaged over pixels with the estimated disparity of all testing images. For our experiment, we consider a pixel to be correctly estimated if the disparity end-point error is $<5$ pixels or $<5\%$. The $\delta$ inlier ratios measure the percentage of the disparity error within the preset maximal mean relative error of $\delta_i = 1.25^i$, where $i \in \{1, 2, 3\}$.

**Main results:** In Figure 4, we evaluate our sparse disparity map $D_s$, the improved disparity map $D_p$ (using a coarse-to-fine framework) and $D_{AA}$ at edge pixels of $D_p$. Because the binary edges from the event camera are very noisy and they are highly dependent on camera motion (see Fig. 3), some of the edges cannot be captured by $D_s$. Using the coarse-to-fine framework, the disparity of each pixel in $D_p$ contains a local average corresponding to its neighborhood pixels that average out the noise and takes advantage of the spatial information. Figure 4 shows that our coarse-to-fine framework leads to a smoother and more accurate disparity map. Especially in the simple boxes dataset (the first row of Fig. 4), the image pyramid helps $D_p$ to obtain more accurate disparity estimation (yellow to green).

Table II shows that our coarse-to-fine result $D_p$ achieves more than 51\% relative improvement in RMSE and 55\% relative improvement in bad-p respectively versus $D_s$. It also outperforms the learning-based stereo method AANet [35] (evaluated on edges) with a significant margin in all three metrics. Since event cameras only report intensity changes at textured areas, ECNN [36] relies on regularisation and spatial priors to reconstruct a smooth and dense image. This will distort the edge information in the reconstructed image and degrade the performance of AANet. For example, in Fig. 4 highlighted by the red bounding box, $D_{AA}$ failed to estimate disparity correctly but our $D_p$ achieves the more precise disparity estimation.

To provide a visualisation of the results, we plot the 3D point cloud from the computed disparity ($D_p$) in Fig. 5, with point colour taken from the frame image. The resulting 3D scene is a reasonable geometric representation of a part of the scene in Fig. 2(b) and it would be sufficient to undertake robotic tasks like grasping.

**B. Disparity completion.**

Compared to a stereo camera system with two event sensors, our SHEF camera system also provides RGB image intensities. To demonstrate the potential of our system and for better

![Fig. 4. Examples of the comparison of our results in three setups. (a) Intensity image. (b) Ground truth disparity map at intensity image edges. (c) Dense disparity map $D_{AA}$ computed by the state-of-the-art stereo-frame-based disparity method AANet, using the input of intensity frames and pure-event reconstruction by the state-of-the-art algorithm ECNN [36]. (d) Our sparse disparity map $D_s$. (e) Our sparse disparity map $D_p$ with the coarse-to-fine framework. Disparity maps are color coded, from blue (far) to yellow (close), in the range of 0-80 pixels. Area of significant differences is highlighted with red rectangle.](image)
where $D_p$ is the estimated disparity map and $D_{c}$ is the ground truth disparity map. We use ground-truth disparity map $D$ for training, and adopt the widely used smooth $\ell_1$ loss $\mathcal{S}(\cdot)$ [42] to penalizes the differences between $D$ and $D_{c}$. Our loss function $\mathcal{L}$ is defined as

$$\mathcal{L} = \frac{1}{N} \sum_{x,y} \mathcal{S}(D(x,y) - D_{c}(x,y))$$  \hspace{1cm} (6)

where $N$ is the number of pixels.

The training data is based on the Middlebury dataset [43], which provides stereo image pairs, camera parameters, and the associated ground-truth disparity map. We take the left-view as the reference, and use the right-view image to generate the event data. With a given rotation and translation matrix, we reproject the right-view image to a new view (horizontal shifting only), and then generate the event stream. Our training data only includes 450 event-frame pairs. For a fair comparison, both AANet and DCNet is trained and fine-tuned on the same dataset.

**Implementation details:** Our network is implemented in Pytorch and is trained from scratch using the Adam optimizer [44] with a learning rate of $10^{-4}$ and a batch size of 10. Our model is trained on a single NVIDIA Titan XP GPU.

**Result:** We demonstrate the dense disparity evaluation of our result $D_c$ by DCNet and $D_{AA}$ by AANet in Fig. 6. From both dense disparity result, $D_{AA}$ and our $D_{c}$, the outlines of the objects are distinguishable, but AANet leads to large areas with wrong disparity estimation. For example, in the simple boxes scene (first line in Fig. 6), $D_{AA}$ has a smooth dense disparity of the left box, but wrongly estimate the disparity of the entire object. This explains the higher errors in $D_{AA}$ for all three metrics.

The Table III shows that the large majority of the pixels of $D_{AA}$ and our $D_{c}$ has the maximal mean relative error of $\delta_2$, indicating only a small amount of outliers in the disparity map are introduced from both the methods. However, $D_{AA}$ leads to only around 64% pixels with the maximal mean relative error of $\delta_1$ and more than 56% pixels have disparity error larger than 5 pixels or 5%, which may be caused by the fact that $D_{AA}$ has more pixels with imprecise disparity around the texture-less region. Comparing to $D_{AA}$, our $D_{c}$ is more reliable and it has around 16% smaller RMSE.

**VI. CONCLUSIONS**

In this paper, we discuss the advantages of the stereo hybrid event-frame (SHEF) camera configuration and introduce a high-quality dataset targeted at stereo hybrid event-frame disparity estimation in near field scenes. We provide a baseline disparity estimation method that computes cross-correlation between binary edge image pairs of event data and frames with a coarse-to-fine framework. Our method outperforms the state-of-the-art stereo matching method AANet on the proposed dataset.

**VII. ACKNOWLEDGMENTS**

The authors would like to thank Pieter van Goor for helping ground truth depth collection, and thank Prophesee for providing the event camera that was used in the work.
REFERENCES

[1] P. Lichtsteiner, C. Posch, and T. Delbrück, “A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor,” IEEE J. Solid-State Circuits, vol. 43, no. 2, pp. 566–576, 2008.

[2] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbrück, “A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor,” IEEE Journal of Solid-State Circuits, vol. 49, no. 10, pp. 2333–2341, 2014.

[3] Y. Hu, S.-C. Liu, and T. Delbrück, “V2e: From video frames to realistic dvs events,” in IEEE Conf. Comput. Vis. Pattern Recogn. Workshops (CVPRW), 2021, pp. 1312–1312.

[4] L. Pan, M. Liu, and R. Hartley, “Single image optical flow estimation with an event camera,” in IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR). IEEE, 2020, pp. 1669–1678.

[5] D. Tedaldi, G. Gallego, E. Mueggeler, and D. Scaramuzza, “Feature detection and tracking with the dynamic and active-pixel vision sensor (davis),” in 2016 Second International Conference on Event-based Control, Communication, and Signal Processing (EBCSCSP). IEEE, 2016, pp. 1–7.

[6] H. Liu, D. P. Moey, G. Das, D. Neil, S.-C. Liu, and T. Delbrück, “Combined frame-and event-based detection and tracking,” in 2016 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2016, pp. 2511–2514.

[7] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, “EKLT: Asynchronous photometric feature tracking using events and frames,” Int. J. Comput. Vis., 2019.

[8] C. Brandli, L. Muller, and T. Delbrück, “Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor,” pp. 686–689, 2014.

[9] C. Scheerlinck, N. Barnes, and R. Mahony, “Continuous-time intensity estimation using event cameras,” in Asian Conference on Computer Vision. Springer, 2018, pp. 308–324.

[10] Z. Wang, Y. Ng, P. van Goor, and R. Mahony, “Event camera calibration of per-pixel biased contrast threshold,” Australasian Conference on Robotics and Automation, ACRA, 2019.

[11] L. Pan, R. I. Hartley, C. Scheerlinck, M. Liu, X. Yu, and Y. Dai, “High frame rate video reconstruction based on an event camera,” IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[12] Z. Wang, Y. Ng, C. Scheerlinck, and R. Mahony, “An asynchronous kalman filter for hybrid event cameras,” in Int. Conf. Comput. Vis. (ICCV), 2021.

[13] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, “Learning event-based motion deburring,” in IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR), 2020, pp. 3320–3329.

[14] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, “Bringing a blurry frame alive at high frame-rate with an event camera,” in IEEE Conf. Comput. Vis. Pattern Recogn. (CVPR), 2019, pp. 6820–6829.

[15] Z. W. Wang, P. Duan, O. Cossaïrt, A. Katsaggelos, T. Huang, and B. Shi, “Joint filtering of intensity images and neuromorphic events camera dataset for 3d perception,” IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[16] A. Mitrokhin, C. Ye, C. Fermüller, Y. Aloimonos, and T. Delbruck, “Ev-imo: Motion segmentation dataset and learning pipeline for event cameras,” in IEEE/RSJ Int. Conf. Intel. Robot. Syst. (IROS). IEEE, 2019, pp. 6105–6112.

[17] “Universal robots - ur5 technical specifications,” 2016. [Online]. Available: https://www.universal-robots.com/media/50588/ur5_en.pdf

[18] MathWorks, “Matlab stereo camera calibrator app.” [Online]. Available: https://au.mathworks.com/help/vision/us/ stereo-camera-calibrator-app.html

[19] R.-h. Liang and J.-f. Mao, “Hand-eye calibration with a new linear decomposition algorithm,” Journal of Zhejiang University-SCIENCE A, vol. 9, no. 10, pp. 1363–1368, 2008.

[20] H. Xu and J. Zhang, “Anet: Adaptive aggregation network for efficient stereo matching,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1959–1968.

[21] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, “Reducing the Sim-to-Real gap for event cameras,” in Eur. Conf. Comput. Vis. (ECCV), 2020.

[22] 3DFLOW, “3dflow zephyr - the complete photogrammetry solution.” [Online]. Available: https://www.3dflow.net/

[23] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbrück, “A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor,” IEEE Transactions on Neural Systems and Learning Systems, vol. 11, p. 535, 2017.

[24] M. Rebecq, G. Gallego, E. Mueggeler, and D. Scaramuzza, “EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time,” Int. J. Comput. Vis., vol. 126, no. 12, pp. 1394–1414, Dec. 2018.

[25] G. Gallego, M. Firouzi and J. Conradt, “Asynchronous event-based cooperative stereo matching using neumorphic silicon retinas,” Neural Processing Letters, vol. 43, no. 2, pp. 311–326, 2016.

[26] Z. Xie, S. Chen, and G. Orchard, “Event-based stereo depth estimation using belief propagation,” Frontiers in neuroscience, vol. 11, p. 535, 2017.

[27] H. Rebecq, G. Gallego, E. Mueggeler, and D. Scaramuzza, “EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real-time,” IEEE Robot. Autom. Lett., vol. 2, no. 2, pp. 593–600, 2017.

[28] MathWorks, “Matlab stereo camera calibrator app.” [Online]. Available: https://au.mathworks.com/help/vision/us/ stereo-camera-calibrator-app.html

[29] M. Osswald, S.-H. Ieng, R. Benosman, and G. Indiveri, “A spiking neural network model of 3D perception for event-based neumorphic stereo vision systems,” Scientific Reports, vol. 7, no. 1, Jan. 2017.

[30] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbrück, “Event-driven stereo visual tracking algorithm to solve object occlusion,” IEEE transactions on neural networks and learning systems, vol. 29, no. 9, pp. 4223–4237, 2017.

[31] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbrück, “A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor,” IEEE Journal of Solid-State Circuits, vol. 49, no. 10, pp. 2333–2341, 2014.

[32] T. Takatani, and Y. Fu, “Learning to reconstruct high speed and high dynamic range videos from events,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 2024–2033.