IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

Real-time Hetero-Stereo Matching for Event and Frame Camera with Aligned Events Using Maximum Shift Distance

Haram Kim[1], Sangil Lee[1,2], Junha Kim[1] and H. Jin Kim[1]

***Abstract*—Event cameras can show better performance than frame cameras in challenging scenarios such as fast-moving environments or high-dynamic-range scenes. However, it is still difficult for event cameras to replace frame cameras in non-challenging normal scenarios. In order to leverage the advantages of both cameras, we conduct a study for the heterogeneous stereo camera system which employs both an event and a frame camera. The proposed system estimates the semi-dense disparity in real-time by matching heterogeneous data of an event and a frame camera in stereo. We propose an accurate, intuitive and efficient way to align events with 6-DOF camera motion, by suggesting the maximum shift distance method. The aligned event image shows high similarity to the edge image of the frame camera. The proposed method can estimate poses of an event camera and depth of events in a few frames, which can speed up the initialization of the event camera system. We verified our algorithm in the DSEC dataset. The proposed hetero-stereo matching outperformed other methods. For real-time operation, we implemented our code using parallel computation with CUDA and release our code open source: https://github.com/Haram-kim/Hetero_Stereo_Matching**

***Index Terms*—Computer Vision for Automation**

I. Introduction

Event cameras can stably measure visual information in high-dynamic-range and high speed environments that are challenging for conventional cameras. However, because of the frameless and the asynchronous characteristics of event data, conventional vision systems could not be directly employed. For several years, various applications for event camera have been studied [1]–[5]. In particular, active research on employing both event and frame cameras aims to demonstrate improved performance than the event-only methods and to extend the event research into practice.

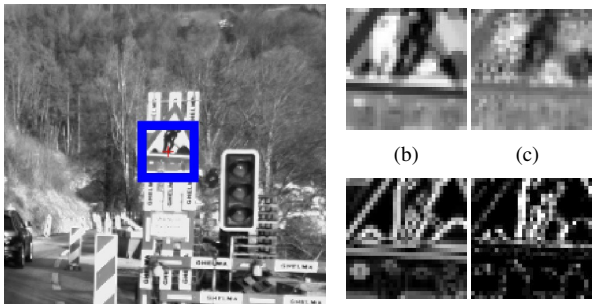


Fig. 1: Illustrations of the stereo matching results. (a) frame image. Red cross represents the edge pixel $\mathbf{f}$ and blue square represents patch $\mathbf{P}(\mathbf{f}, {}^{F}\mathbf{I}_n)$ of the n-th frame image ${}^{F}\mathbf{I}_n$. (b) and (d) are the frame patch of the temporal gradient image ${}^{F}\mathbf{I}_n^{\Delta\tau}$ and the edge image ${}^{F}\mathbf{I}_n^{\Delta\mathbf{x}}$ represented as $\mathbf{P}(\mathbf{f}, {}^{F}\mathbf{I}_n^{\Delta\tau})$, $\mathbf{P}(\mathbf{f}, {}^{F}\mathbf{I}_n^{\Delta\mathbf{x}})$, respectively. (c) and (e) are the event patch of the raw event image ${}^{E}\mathbf{I}_n^{\Delta\tau}$ and the aligned event image ${}^{E}\mathbf{I}_n^{\Delta\mathbf{x}}$ with the estimated disparity d^* represented as $\mathbf{P}(\mathbf{f}+[d^*,0]^\intercal, {}^{E}\mathbf{I}_n^{\Delta\tau})$, $\mathbf{P}(\mathbf{f}+[d^*,0]^\intercal, {}^{E}\mathbf{I}_n^{\Delta\mathbf{x}})$, respectively.

Significant progress has been made in the fields of feature tracking and high-frame-rate scene reconstruction with certain event cameras that also can capture conventional frame images from active pixel sensor (APS). However, the cameras are placed in the same position in such configuration, the stereo vision could not be applied, even though two vision sensors are used. Here, we conduct a study using a frame camera and an event camera simultaneously as a stereo camera that can utilize the existing vision algorithm, and can be used in harsh environments as Fig. 2. We propose a method to estimate the disparity by using an event camera and a frame camera as a hetero-stereo camera which have different camera parameters. Our contributions can be summarized as follows:

- We provide the hetero-stereo matching methods in order to associate different data types of event-frame cameras and estimate disparities.
  1) Initial matching method with time-gradient images for fast initialization of the system (Section III-A)
  2) Aligned-event-based matching method with edge images for accurate stereo matching (Section III-C)
- We provide an accurate and intuitive event aligning method to describe the edge-like images, which utilizes internally estimated camera poses (Section III-B).
- We suggest the concept of maximum shift distance to align events efficiently (Section III-D).

Manuscript received: September 2, 2022; Revised November 6, 2022; Accepted November 7, 2022. This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers' comments.

This research was supported by Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea(NRF) and Unmanned Vehicle Advanced Research Center(UVARC) funded by the Ministry of Science and ICT, the Republic of Korea(NRF-2020M3C1C1A01086411).

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 202013D05)

[1] Lab for Autonomous Robotics Research (LARR), Automation and Systems Research Institute (ASRI) and Seoul National University (SNU), Seoul 08826, South Korea, [2] Samsung Electronics, Hwaseong, South Korea {rlgkfka614, sangil07, wnsgk02, hjinkim}@snu.ac.kr

Digital Object Identifier (DOI): see top of this page.

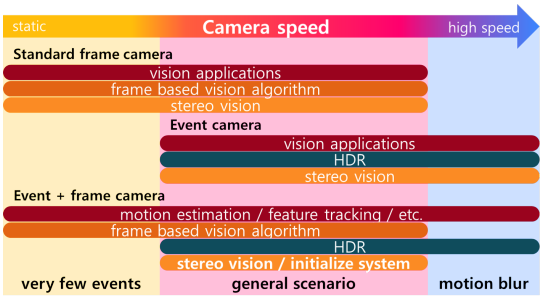


Fig. 2: Hetero-stereo matching algorithm operating range according to camera speed. The proposed method enables hetero-stereo camera system to apply vision applications (motion estimation, feature tracking, etc.) regardless of camera speed and to have high dynamic range (HDR).

The proposed methods estimate the disparity and depth of events by associating the event and frame data. We obtain camera poses from the 3D reconstructed points which are computed from the initial matching method. Then, we estimate accurate disparity and depth by matching the frame edge image to the aligned events using the camera poses. For aligning events, we extend the event alignment module in [6], additionally considering the translation motion and an arbitrary depth in the disparity range. Rather than considering all disparities in range, we efficiently compute the aligned events with maximum shift distance method by considering the representative disparities which produces distinct aligned event images.

II. Related Works

A significant number of vision algorithms employ 3D scene information to demonstrate more diverse applications. Similarly, in order to extend the limited application of a single event camera, stereo event studies [7]–[16] aimed to estimate the depth. These studies were conducted to utilize event cameras as stereo, and several attempts have been made to extract meaningful features from events. In [8], semi-dense 3d reconstruction was performed employing a stereo event camera. The method estimated the depth on edges where events frequently spiked, considering multiple viewpoints. The authors of [7] estimated the disparity after aligning events with optical flow in consideration of camera motion. Since it is difficult to accurately estimate the camera pose with event data only, the authors used ground truth poses. Even though stereo event cameras can utilize the depth information, there is a fundamental issue that the events do not spike in static situations and the advantage is revealed only in specific scenarios.

In order to extend the usability of event cameras in general scenarios, several attempts [17]–[21] have recently been made to combine the advantages of frame and event, which can also enable the various applications of frame cameras.

In [17], the author attempted to use stereo cameras and event cameras together. When combining an event camera with a stereo camera or a RGB-D camera, the individual depth of an event is still unknown due to asynchronous characteristic. To combine events and frames, the author conducted the dense and continuous disparity estimation method using camera motion. To accomplish the same goal as [17] in a monocular camera, the authors of [18] estimated depth map in high frame rate from the monocular event camera which also provides frame images. Studies [17]–[20] use two types of vision data, but cannot operate as stereo because frames and events are acquired from a monocular camera.

So far, there has been a lack of study on stereo event frame studies that can estimate the depth of events in general scenarios. In [22], [23], the authors conveyed a study to perceive 3D scene by firstly configuring an event and a frame camera as stereo. By using the event-and-frame camera attached to the robot arm, the author estimated the disparity through stereo matching with binary frame edge and binary event edge images. The binary frame edge images are generated from frame images, and the binary event edge images are from the event data with a high-pass filter and non-maximal suppression.

In order to extend the applications of the event camera, we tackle the problem of how to associate event data and frame images in stereo event frame camera settings. We present an accurate and intuitive way to align asynchronous event data. In order to accurately describe the edge of the scene, the proposed method warps events by only considering camera motion and disparity. Thus, the proposed event aligning module is parameter-free regardless of the speed of camera motion and data domain. In [22], binary event edge images are obtained from the high-pass filter of [24] with fixed parameters. Such binary edge images do not describe the scene properly on evaluation dataset of Section IV. Thus, [22] shows low accuracy on disparity estimation as in Table I.

While the method [7] utilized the ground truth camera pose and assumed that the depth of all events in a short time window only depends on the pixel coordinate (not the temporal coordinate), we estimate the camera pose from the initial matching method and assume that events have the same depth value only at the reference time. For aligning events with 6-DOF camera motion, we extend the accurate warping method of [6] rather than using the first order approximation of the warping function.

The proposed method can produce an aligned event image that looks similar to an edge image as in Fig. 1.(e), and the edge features are appropriately represented even when we use uniformly sampled events at 10% of the total as in Fig. 4. Additionally, we introduce the concept of maximum shift distance to efficiently compute aligned event images. We expect to contribute to the event camera community by suggesting an intuitive and efficient way to present edge images from event data.

III. Hetero-Stereo Matching

There are two concepts to associate the event camera with the frame camera. The event data should be reconstructed into frame images [25]–[28], or frame images should be converted into the event camera-like images. Frame reconstruction methods still suffer from unwanted artifacts such as bleeding and local reconstruction error amplification problems. We confirmed that the artifacts degrade the performance of

disparity estimation in the attached video. Thus, we will cover the matching method by converting frame images to event images between the two concepts.

Each event $e_k = (\mathbf{x}_k, t_k, p_k)$ consists of position $\mathbf{x}_k$ in pixel coordinate, timestamp t_k and polarity $p_k \in \{+1, -1\}$. We construct an event set between the time of $(n-1)$th frame τ_{n-1} and the time of n-th frame τ_n as $\mathbf{E}|_{\tau_{n-1}}^{\tau_n} = \{e_k | \tau_{n-1} \le t_k \le \tau_n\}$. Here, k is the index of an event in the event set and N_n is the cardinality of $\mathbf{E}|_{\tau_{n-1}}^{\tau_n} = \{e_k\}_{k=1}^{N_n}$.

For stereo matching, we should undistort and rectify the event and frame cameras. In general for frame images, the inverse mapping technique is used during rectification and undistortion. Likewise, we utilize the inverse mapping to rectify the raw event image for the initial stereo matching of Section III-A. The inverse mapping is an *image-to-image* operation. On the other hand, for aligning events in the Section III-B, we should warp events *individually* considering their time and camera motion after rectification. This means that warping of individual events cannot be performed after the inverse mapping (because the time information of individual events get lost to perform the image-to-image operation). Thus, we undistort and rectify events in a direct way, before warping the events.

In the heterogeneous camera case, the resolution of the event camera and the frame camera may be different. If the rectification resolution is set to be different from the resolution of the event camera, the image projected from events will suffer from web-shaped artifacts, because some pixels contain overlapped events or nothing. Likewise, the focal length should not be significantly different from the focal length of the event camera. Thus, we set the resolution and the focal length of rectification coordinate to those of the event camera. In this paper, all the event points and frame images are rectified.

A. Initial Stereo Matching

Event cameras record polarities of change of log intensity and have a different dynamic range from frame cameras. For associating the frame image ${}^F\mathbf{I}_n$ with the event set $\mathbf{E}|_{\tau_{n-1}}^{\tau_n}$ in the initial phase, we use the temporal gradient image ${}^F\mathbf{I}_n^{\Delta\tau} = {}^F\mathbf{I}_n - {}^F\mathbf{I}_{n-1}$ and the event image ${}^E\mathbf{I}_n^{\Delta\tau}$ which is computed by accumulating the raw events of $\mathbf{E}|_{\tau_{n-1}}^{\tau_n}$ considering polarity. The frame image ${}^F\mathbf{I}_n^{\Delta\tau}$ and the event image ${}^E\mathbf{I}_n^{\Delta\tau}$ contain information about intensity changes. Because the above two images have the same tendency but different dynamic range, we utilize the normalized cross correlation (NCC) cost that is robust to the difference in dynamic range, rather than applying residual cost such as the sum of absolute distance (SAD) and the sum of squared distance (SSD).

Then, we extract the edge pixel $\mathbf{f} \in \mathbb{R}^{2\times 1}$ from the frame image with Sobel filter and conduct patch matching methods on the temporal gradient image and the event image. We shift the event patch $\mathbf{P}(\mathbf{f}, {}^E\mathbf{I}_n^{\Delta\tau})$ along the x coordinate (parallel to epipolar line) and compare $\mathbf{P}(\mathbf{f}, {}^E\mathbf{I}_n^{\Delta\tau})$ with the frame patch $\mathbf{P}(\mathbf{f}, {}^F\mathbf{I}_n^{\Delta\tau})$ using NCC cost $\mathrm{C}(\cdot,\cdot)$. Additionally, we apply 2D Gaussian-smoothing on NCC cost along the image coordinate to alleviate the noise effect. Then, we estimate the disparity $\hat{d}$ where the Gaussian-smoothed NCC $\mathrm{G}(\mathscr{E}_\tau(\mathbf{f}, d))$ is maximized as

$$\mathscr{E}_\tau(\mathbf{f}, d) = \mathrm{C}\left(\mathbf{P}(\mathbf{f}, {}^F\mathbf{I}_n^{\Delta\tau}), \mathbf{P}(\mathbf{f} + [d, 0]^\intercal, {}^E\mathbf{I}_n^{\Delta\tau})\right), \tag{1}$$

$$\hat{d} = \underset{d}{\operatorname{argmax}}\ \mathrm{G}(\mathscr{E}_\tau(\mathbf{f}, d)). \tag{2}$$

For the sub-pixel accuracy, we interpolate the disparity with quadratic interpolation as follows:

$$d^* = \hat{d} + 0.5\frac{\mathscr{E}_\tau(\mathbf{f}, \hat{d}-1) - \mathscr{E}_\tau(\mathbf{f}, \hat{d}+1)}{2\mathscr{E}_\tau(\mathbf{f}, \hat{d}) - \mathscr{E}_\tau(\mathbf{f}, \hat{d}-1) - \mathscr{E}_\tau(\mathbf{f}, \hat{d}+1)}. \tag{3}$$

The presented initialization of the system can be performed with two frames if sufficient events are spiked.

B. Event Alignment

We can expect more accurate matching by using edge images instead of temporal gradient images. In order to match the event data with edge images, we align events with the camera motion. We estimate the frame camera pose ${}^F\mathbf{T}$ by applying APnP [29] from the constructed 3D points with the initial disparity, and then compute the event camera pose ${}^E\mathbf{T}$ with extrinsic parameters.

When accurately warping event points, it is computationally expensive to consider the time and depth of each event. In order to reduce the computation load of warping function, we assume that the events in $\mathbf{E}|_{\tau_{n-1}}^{\tau_n}$ share the same depth at reference time τ_n, which means that the depth values of events are all the same after warping with the motion ${}^E\mathbf{T}_{n-1}^n$. Here, the depth at reference time is a given value and is covered in Section III-C, which allows the event to have different depth at the event time t_k. This condition is a more relaxed than [7] which assumes that the depth of the event is the same regardless of the event time.

Even if the events have the same depth at reference time τ_n, an event e_k can have different depth $z_k(t_k)$ before warping. The warping function for the k-th event can be presented as follows.

$$\mathbf{X}_k(\tau_n) = \mathbf{R}_n(\delta t_k)\mathbf{X}_k(t_k) + \mathbf{t}_n(\delta t_k), \tag{4}$$

$$\mathbf{X}_k(\tau_n) = z_k(t_k)\mathbf{R}_n(\delta t_k)\bar{\mathbf{X}}_k(t_k) + \mathbf{t}_n(\delta t_k) \tag{5}$$

where $\mathbf{X}_k(t) = [x_k(t), y_k(t), z_k(t)]^\top$ is the inverse-projected 3D point of the event e_k in the camera coordinate at time t and $\bar{\mathbf{X}}_k(t) = [\bar{x}_k(t), \bar{y}_k(t), 1]^\top$ is the inverse-projected point from the event pixel $\mathbf{x}_k$, which satisfying $\mathbf{X}_k(t) = z_k(t)\bar{\mathbf{X}}_k(t)$. $z_k(t_k)$ is the exact depth of the event and $z_k(\tau_n)$ is the depth at the reference time. The reference depths $z_k(\tau_n)$ of events in $\mathbf{E}|_{\tau_{n-1}}^{\tau_n}$ are given and have all the same value, as we assumed, while $z_k(t_k)$ differs. δt_k is the time difference between the time of an event t_k and the reference time τ_n i.e. $\delta t_k = \tau_n - t_k$, $\mathbf{R}_n(\delta t_k)$ and $\mathbf{t}_n(\delta t_k)$ is the rotation and the translation matrix of camera motion ${}^E\mathbf{T}_{n-1}^n$ considering the time difference δt_k.

Then, the depth of the event $z_k(t_k)$ can be inversely computed with given $z_k(\tau_n)$ as follows.

$$z_k(\tau_n) = \begin{bmatrix}0 & 0 & 1\end{bmatrix}\left(z_k(t_k)\mathbf{R}_n(\delta t_k)\bar{\mathbf{X}}_k(t_k) + \mathbf{t}_n(\delta t_k)\right), \tag{6}$$

$$z_k(t_k) = \frac{z_k(\tau_n) - \begin{bmatrix}0 & 0 & 1\end{bmatrix}\mathbf{t}_n(\delta t_k)}{\begin{bmatrix}0 & 0 & 1\end{bmatrix}\mathbf{R}_n(\delta t_k)\bar{\mathbf{X}}_k(t_k)}, \tag{7}$$

We can get aligned event points with the same depth at reference time by putting $z_k(t_k)$ back into Eq. (5).

Since each event has different δt_k, we need to lighten the computational load when computing $\mathbf{R}_n(\delta t_k)$ and $\mathbf{t}_n(\delta t_k)$ for each event. We convert the existing warping function into a matrix operation using the second-order approximation as in [6]. We employ the twist coordinate representation $[\nu;\omega] \in \mathbb{R}^6$ provided by the Lie algebra $\mathfrak{se}(3)$ associated with the group SE(3), where ν is the linear velocity and ω is the angular velocity. The exact warping can be achieved with Rodrigues' formula as follows.

$$\mathbf{R}_n(\delta t_k)\bar{\mathbf{X}}_k(t_k) = \bar{\mathbf{X}}_k(t_k) + \frac{\hat{\omega}}{|\omega|}\bar{\mathbf{X}}_k(t_k)\sin(|\omega|\delta t_k) + \frac{\hat{\omega}^2}{|\omega|^2}\bar{\mathbf{X}}_k(t_k)\left(1-\cos(|\omega|\delta t_k)\right), \quad (8)$$

$$\mathbf{t}_n(\delta t_k) = \nu\delta t_k + \frac{\hat{\omega}\nu}{|\omega|^2}\left(1-\cos(|\omega|\delta t_k)\right) + \frac{\hat{\omega}^2\nu}{|\omega|^3}\left(|\omega|\delta t_k - \sin(|\omega|\delta t_k)\right), \quad (9)$$

where $\hat{\omega}$ is the cross product matrix of ω. By substituting $\cos(|\omega|\delta t_k) \approx 1-|\omega|^2\delta t_k^2/2$ and $\sin(|\omega|\delta t_k) \approx |\omega|\delta t_k$ in Eq. (8) and Eq. (9), the rotation and translation can be simplified as follows.

$$\mathbf{R}_n(\delta t_k)\bar{\mathbf{X}}_k(t_k) \approx \bar{\mathbf{X}}_k(t_k) + \hat{\omega}\bar{\mathbf{X}}_k(t_k)\delta t_k + \frac{1}{2}\hat{\omega}^2\bar{\mathbf{X}}_k(t_k)\delta t_k^2, \quad (10)$$

$$\mathbf{t}_n(\delta t_k) \approx \nu\delta t_k + \frac{1}{2}\hat{\omega}\nu\delta t_k^2. \quad (11)$$

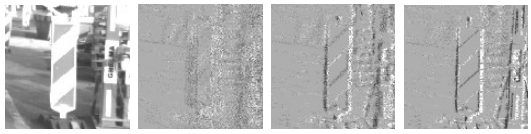
Then, the aligned event point $\mathbf{X}_k(\tau_n)$ which has the same depth at reference time can be obtained by substituting Eq. (10) and Eq. (11) into Eq. (7) and Eq. (5).

C. Stereo Matching with Aligned Events

We perform stereo matching using the aligned events to increase the number of inlier disparity and to estimate disparity more accurately than the initial phase. To align the event, the camera pose ${}^{E}\mathbf{T}^{n}_{n-1}$ and the reference depth $z_k(\tau_n)$ are required. Since we slide the patch images in stereo matching, there are several disparity and depth candidates. We set the reference depth candidates by converting the disparity range as $z = f\cdot b/d$ where f is the focal length and b is the length of the baseline, and compute the bunch of aligned events for each depth candidates by assuming that the entire events of $\mathbf{E}|^{\tau_n}_{\tau_{n-1}}$ have the same reference depth. By projecting the aligned events $\mathbf{X}(\tau_n)$ into the image plane, we obtain the aligned event image ${}^{E}\mathbf{I}^{\Delta\mathbf{x}}_n(d)$ with disparity d as a parameter as shown in Fig. 3. Then, we perform the stereo matching with the aligned event image ${}^{E}\mathbf{I}^{\Delta\mathbf{x}}_n(d)$ and the edge image ${}^{F}\mathbf{I}^{\Delta\mathbf{x}}_n = \left|\nabla {}^{F}\mathbf{I}_n\right|$ as follows.

$$\mathscr{E}_{\mathbf{x}}(\mathbf{f},d) = \mathrm{C}\left(\mathbf{P}(\mathbf{f},{}^{F}\mathbf{I}^{\Delta\mathbf{x}}_n), \mathbf{P}\left(\mathbf{f}+[d,0]^{\intercal},{}^{E}\mathbf{I}^{\Delta\mathbf{x}}_n(d)\right)\right), \quad (12)$$

$$\hat{d} = \underset{d}{\operatorname{argmax}}\ \mathrm{G}\left(\mathscr{E}_{\mathbf{x}}(\mathbf{f},d)\cdot\mathscr{E}_{\tau}(\mathbf{f},d)\right). \quad (13)$$



(a) Reference (b) $d=0$ (c) $d=15$ (d) $d=30$

Fig. 3: Illustrations of (a) frame image ${}^{F}\mathbf{I}_n$ and (b - d) aligned event images ${}^{E}\mathbf{I}^{\Delta\mathbf{x}}_n(d)$ with varying disparity d. For aligned event images, polarity is considered only in the figures for visibility.

(a) Frame (b) Edge (c) 100% (d) 50% (e) 10%

Fig. 4: Illustrations of sampling effect on the aligned event image. (a) frame image, (b) edge of frame image, (c) aligned event image without sampling, (d) aligned event image with half of the events and (e) aligned event image with 10% events. Even with a small number of events, the proposed method can describe edges for stereo matching.

The aligned event images improve stereo matching accuracy by representing clear edges, but similar edges can be mismatched because such matching does not take into account polarity. We alleviate this ambiguity problem by multiplying the initial matching cost that considers polarity as in Eq. (13).

As in the initial phase, the interpolated disparity d^* is computed as in Eq. (3) for sub-pixel accuracy. Due to the aligning events using camera motion, the proposed method can reliably represent edge features even when we use the smaller number of events as in Fig. 4.

D. Maximum Shift Distance (MSD) by Translational Motion

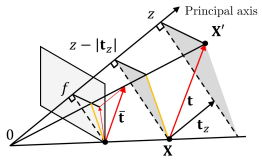
If there is no translational motion ($\mathbf{t}_n = 0$), the aligned event images are identical, regardless of the varying disparity. In this case, it is inefficient to compute the aligned event image for each disparity, because we only need one aligned event image. We mitigate this inefficiency by grouping disparities that produce similar aligned events. The disparities are grouped based on the maximum shift distance of events which is related to the magnitude of translation.

In this section, we assume that all points are already rotated in order to focus on the effect of translational motion on the maximum shift distance. When points are shifted by the translational motion, the most shifted point exists at the corner of the image due to the characteristics of the pinhole camera model.

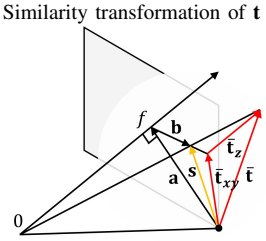
In the presence of the translational motion $\mathbf{t}$, the 3D point $\mathbf{X}$ which corresponds to the image corner will be warped to $\mathbf{X}'$ with depth z as depicted in Fig. 5. The trajectory of the 3D point $\mathbf{X}$ is projected onto the image plane is called the event shift trajectory, denoted as $\mathbf{s}$. It is the orange colored vector in Fig. 5.(a).

$$\bar{\mathbf{t}} = \frac{f}{z-|\mathbf{t}_z|}\mathbf{t}, \quad (14)$$

where $\bar{\mathbf{t}}$ is the scaling transformation of translation $\mathbf{t}$ for the image plane. $\mathbf{t}_{xy}, \mathbf{t}_z$ are the decomposed vectors of $\mathbf{t}$,



(a) Similarity transformation of $\mathbf{t}$ to $\bar{\mathbf{t}}$



(b) Computing maximum shift distance $\mathbf{s}$

Fig. 5: Illustrations of maximum shift distance. The corner point $\mathbf{X}$ is moved to $\mathbf{X}'$ by translation. The maximum shift distance is the orange colored arrow s, and translation is the red colored arrow $\mathbf{t}$.

perpendicular to the image plane and parallel to the principal axis, respectively.

As shown in Fig. 5.(b), the shift $\mathbf{s}$ is computed as the sum of $\mathbf{a}$ and $\mathbf{b}$, where $\mathbf{a}$ is the vector from the corner to the principal point. $\mathbf{b}$ is parallel to $\bar{\mathbf{t}}_{xy}-\mathbf{s}$, and has a similarity ratio of f and $|\bar{\mathbf{t}}_z|$ in the plane perpendicular to $\mathbf{t}_{xy}$. Then, $\mathbf{b}$ is composed of the two vectors $\mathbf{t}_{xy}$ and $\mathbf{s}$ by substituting $\bar{\mathbf{t}}$ with $\mathbf{t}$ as in the second part of Eq. (15).

$$\mathbf{b} = \frac{f}{|\bar{\mathbf{t}}_z|}(\bar{\mathbf{t}}_{xy}-\mathbf{s}) = \frac{f}{|\mathbf{t}_z|}\mathbf{t}_{xy} - \left(\frac{z}{|\mathbf{t}_z|}-1\right)\mathbf{s} \tag{15}$$

$$\mathbf{s} = \mathbf{a}+\mathbf{b} = \mathbf{a}+\frac{f}{|\mathbf{t}_z|}\mathbf{t}_{xy}+\left(1-\frac{z}{|\mathbf{t}_z|}\right)\mathbf{s}, \tag{16}$$

By substituting $\mathbf{b}$ in Eq. (16) with the second part of Eq. (15) and rearranging Eq. (16) for $\mathbf{s}$, the shift $\mathbf{s}$ can be obtained as Eq. (17).

$$\mathbf{s} = \frac{|\mathbf{t}_z|\mathbf{a}+f\mathbf{t}_{xy}}{z} \tag{17}$$

By the trigonometric inequality, $|\mathbf{s}| \le (|\mathbf{a}||\mathbf{t}_z|+f|\mathbf{t}_{xy}|)/z$ is satisfied, where $|\mathbf{a}|$ is the half of the diagonal image size. We set the maximum shift distance as $(|\mathbf{a}||\mathbf{t}_z|+f|\mathbf{t}_{xy}|)/z$. If the depth z is expressed again as disparity d, the maximum shift distance satisfies $s_{\max}(d,\mathbf{t}) := (|\mathbf{a}||\mathbf{t}_z|+f|\mathbf{t}_{xy}|)/(f\cdot b)\cdot d$ which indicates that it is proportional to the disparity d.

We can quantize the aligned event image without loss by grouping disparities whose s_{max} values do not differ by more than 1 px. In addition, we can significantly reduce the number of aligned event bunches by adjusting the interval of pixels where the s_{max} value differs (MSD interval).

IV. Experimental Results

We evaluated the proposed method on the DSEC dataset [30]. DSEC is the disparity and optical-flow evaluation dataset that records the city driving scenario. It employs high resolution stereo event (640×480) and frame (1440×1080) cameras (4 cameras in total), and provides the ground truth disparity converted from LIDAR.

We compared the proposed method with the initial stereo matching method described in Section III-A, E2VID [25] based method and the implemented version of SHEF [22]. The reconstruction performance of [25] affected by the number of events. We evaluate [25] into the two event grouping manners. E2VID-τ reconstructs frame images from the events in a fixed temporal window ($\tau_n-\tau_{n-1}\approx 50$ ms), $\mathbf{E}|_{\tau_{n-1}}^{\tau_n}$, which we used. E2VID-$N$ uses the fixed number of events ($N=10^5$). Then, we perform stereo matching on the reconstructed frame with NCC cost as in Eq. (2) (E2VID-N/τ). Also, we applied the standard semi-global matching method (SGM) which utilizes residual costs (E2VID-SGM-N/τ).

We set disparity range to 100 px, MSD interval to 10 px, std of Gaussian filter $\mathrm{G}(\cdot)$ to 2 px and the kernel radius of all methods to 12 px.

A. Performance of Stereo Matching

We verified the matching accuracy with the disparity error in Table I. We used the following metrics:

- root mean squared error (RMSE): $\sqrt{\frac{1}{T}\sum_{\mathrm{p}}(d_{\mathrm{gt}}-d_{\mathrm{p}})^2}$
- mean absolute error (MAE): $\frac{1}{T}\sum_{\mathrm{p}}|d_{\mathrm{gt}}-d_{\mathrm{p}}|$
- percentage of absolute error for the all edge pixels (recall) with threshold δ^*: percentage of d_{p} s.t. $|d_{\mathrm{gt}}-d_{\mathrm{p}}|=\delta<\delta^*$

For depth evaluation, we used RMSE and the following metrics:

- absolute relative distance (ARD): $\frac{1}{T}\sum_{\mathrm{p}}\frac{|z_{\mathrm{gt}}-z_{\mathrm{p}}|}{z_{\mathrm{gt}}}$
- percentage of relative error (recall) with threshold δ^*: percentage of z_{p} s.t. $\max(\frac{z_{\mathrm{gt}}}{z_{\mathrm{p}}},\frac{z_{\mathrm{p}}}{z_{\mathrm{gt}}})=\delta<\delta^*$

All error metrics except percentages (i.e. disparity RMSE, MAE, depth RMSE and ARD) are only evaluated for pixels with disparity error within 3 px. This is to evaluate the error for inlier matchings.

The proposed method outperformed other stereo event frame methods (HSM-Init, E2VID-τ [25] and SHEF [22]) for inliers and disparity RMSE and MAE. The standard stereo matching method (E2VID-SGM-N/τ) failed due to the different dynamic range of the reconstructed frame image. NCC cost based E2VID-N [25] methods showed comparable disparity RMSE to the proposed method. The reconstructed images from E2VID methods describe detailed features. Thus, RMSE of E2VID methods is small for the inlier disparity matches. However, the percentage of absolute error of E2VID-N is 10% to 15% less than the proposed method due to the reconstruction failed regions.

For other data sequences, we evaluated the disparity RMSE and percentage of absolute error within 3 px as in Table II. In E2VID methods, especially E2VID-τ, there exist improperly reconstructed local regions for some data sequences, as already reported in [31]. These local region artifacts appear as black region or squiggle pattern when the events are detected either too few over a long period of time or too many over a short period of time. For this reason, in Fig. 6, E2VID-N method estimates the disparity of the foreground better than the background. E2VID-τ often fails on reconstruction, because there are too many events in a single input tensor to E2VID network. The performance of E2VID methods drops

Disparity		RMSE	MAE	percentage of absolute error (δ^*)			Depth		RMSE	ARD	percentage of relative error (δ^*)		
				1 px	2 px	3 px					1.05	1.05^2	1.05^3
HSM	Prop.	**1.036**	**0.796**	**0.560**	**0.743**	**0.800**	HSM	Prop.	3.092	0.060	**0.444**	**0.664**	**0.743**
	Init.	1.131	0.895	0.501	0.723	0.791		Init.	3.118	0.066	0.350	0.634	0.739
E2VID-N		1.048	0.807	0.451	0.595	0.644	E2VID-N		**2.950**	**0.058**	0.363	0.534	0.599
E2VID-τ		1.768	1.558	0.079	0.175	0.268	E2VID-τ		7.490	0.148	0.048	0.100	0.154
E2VID-SGM-N		-	-	0.009	0.116	0.120	E2VID-SGM-N		-	-	0.079	0.108	0.116
E2VID-SGM-τ		-	-	0.002	0.005	0.009	E2VID-SGM-τ		-	-	0.001	0.003	0.006
SHEF		1.630	1.386	0.088	0.163	0.227	SHEF		6.156	0.112	0.063	0.127	0.170

TABLE I: Disparity and depth estimation results for *interlaken_c*. We evaluate the disparity estimation results on edges with root mean squared error, mean absolute error and percentage of absolute error (recall) and evaluate the depth with root mean squared error, absolute relative distance and percentage of relative error (recall). '-' indicates that the algorithm failed with less than 15 % inliers. E2VID-N/τ apply the NCC cost on stereo matching, and E2VID-SGM-N/τ utilize the standard semi-global matching method.

Dataset		*interlaken_c*		*interlaken_d*		*interlaken_e*		*interlaken_f*		*interlaken_g*	
Disparity		RMSE	recall/prec.	RMSE	recall/prec.	RMSE	recall/prec.	RMSE	recall/prec.	RMSE	recall/prec.
HSM	Prop	1.036	0.800/0.806	1.106	0.834/0.838	1.179	0.737/0.742	1.152	0.783/0.788	1.057	0.846/0.850
	Init	1.131	0.791/0.797	1.232	0.838/0.842	1.205	0.748/0.753	1.173	0.791/0.795	1.112	0.837/0.839
E2VID-N		1.048	0.644/0.655	1.093	0.546/0.561	1.348	0.569/0.582	1.384	0.629/0.637	1.267	0.738/0.744
E2VID-τ		1.768	0.268/0.279	1.720	0.167/0.183	1.765	0.271/0.286	1.770	0.372/0.387	1.720	0.371/0.378
SHEF		1.630	0.227/0.230	1.658	0.190/0.196	1.564	0.255/0.278	1.529	0.244/0.282	1.624	0.276/0.299

TABLE II: Disparity estimation results for *interlaken* sequences. We evaluate the disparity with root mean squared error (px) and the percentage of absolute error within 3 px for ground truth edges (recall manner) and disparity estimated edges (precision manner).

in the other data sequences due to the reconstruction artifacts. Meanwhile, the proposed method is free from this artifact issue and always shows reliable disparity estimation.

B. Qualitative Evaluation of Stereo Matching

The results of semi-dense reconstruction is depicted in Fig. 6. The proposed method estimates the disparity on edges. For evaluation, we display the disparity which has less than 10 px error on edges. The proposed method can estimate disparity as densely as the ground truth and perform better than E2VID [25] and SHEF [22]. SHEF [22] cannot properly estimate disparity on unclear edge regions where it is difficult to generalize a high-pass filter to build binary edge maps. We display the detailed patch matching results of the proposed method HSM-Prop and HSM-Init in Fig. 7. The last columns of Fig. 7.(a) and Fig. 7.(b) show examples where HSM-Init failed with the disparity error of more than 10 px.

C. Computation time

We ran the proposed method on NVIDIA GeForce RTX 3080 Ti GPU and Intel Core i9-12900KF @ 3.20GHz CPU. We compute disparity at 640×480 resolution and events are not scaled or sampled at all. The analysis at '*interlaken_c*' is shown in Fig. 8 and Table III. We estimate computation time by averaging the computation time of the modules processing 10 times. The proposed stereo matching method achieves real-time performance by taking 21.13 ms per frame, which is less than 50 ms per frame. Such performance has become possible by utilizing the concept of the maximum shift distance (MSD) which lessen the computational load. The maximum shift distance depends on the size of the translation and the pixel interval. We can compute fewer aligned event images by adjusting the MSD interval. Since most pixels are not located on the corner of the image, most events are shifted much less than $s_{\max}(d, \mathbf{t})$. Thus, even if we aligned the events with more sparse disparity values, it does not significantly affect the stereo matching performance as in Table IV. Rather, the performance becomes better than default, since MSD supports Gaussian smoothing to work better. When a camera moves fast, $s_{\max}(d, \mathbf{t})$ can be greater than the maximum disparity. In this case, it is necessary to compute the aligned event image for all disparity. We implemented SHEF [22] without computation time optimization. SHEF reconstructs edge images using the high-pass filter of [24] and non-maximal suppression, which is a similar concept to the Canny edge detection with the time complexity $\mathcal{O}(n \log n)$ where n is the number of image pixels. SHEF requires relatively light computation on reconstruction than the E2VID methods and HSM-Prop. Thus, SHEF can sufficiently operate in real-time.

HSM-Prop		E2VID-N/τ	
Compute MSD	2.66	Reconstruction	93.26 / 16.36
Compute AEI	3.13		
Stereo NCC AEI	5.00		
Stereo NCC Init	4.79	Stereo NCC	6.04
Stereo postproc	3.12	Stereo postproc	2.42
Etc.	2.41	Etc.	2.08
Total	21.13	Total	103.82 / 26.90

TABLE III: Computation time per frame (ms). MSD is maximum shift distance, AEI is aligned event images. 'Stereo NCC AEI, Init' correspond to construct cost volume $\mathscr{E}_{\mathbf{x}}(\mathbf{f}, d), \mathscr{E}_{\tau}(\mathbf{f}, d)$, respectively, and 'Stereo postproc' is the computation process of Eq. (13). Except computing MSD, all the processes are computed on GPU. E2VID-N takes 93.26 ms for reconstruction, while E2VID-τ takes 16.36 ms.

MSD interval	1 px	2 px	3 px	5 px	10 px	Avg. events per frame
Compute AEI	15.96	10.41	7.06	5.24	3.13	0.8 million
RMSE	1.08	1.07	1.07	1.06	1.04	

TABLE IV: Computation time (ms) to construct aligned event images (AEI) and disparity RMSE (px) with varying maximum shift distance interval.

V. Conclusion

We performed hetero-stereo matching using frame cameras and event cameras, which have different characteristics.

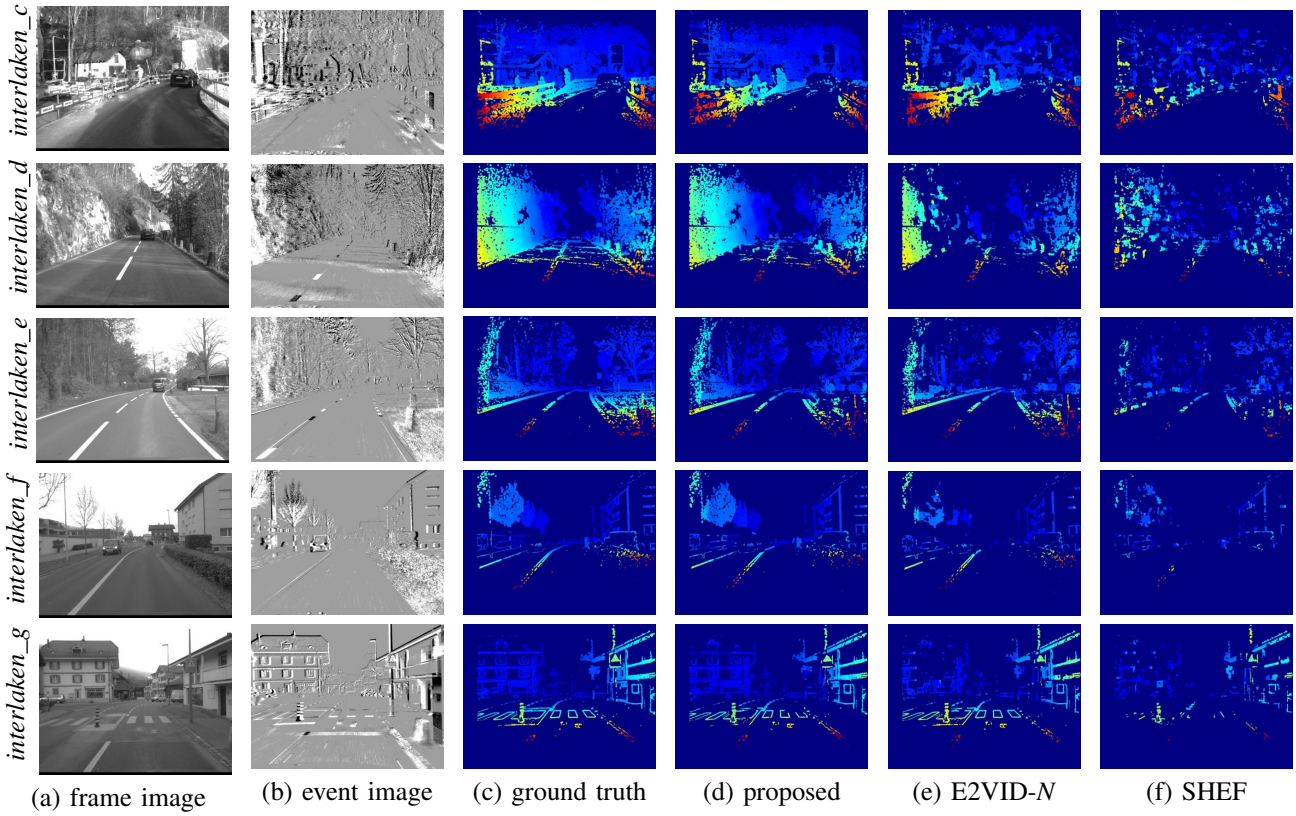


Fig. 6: Snapshots of the semi-dense disparity results.

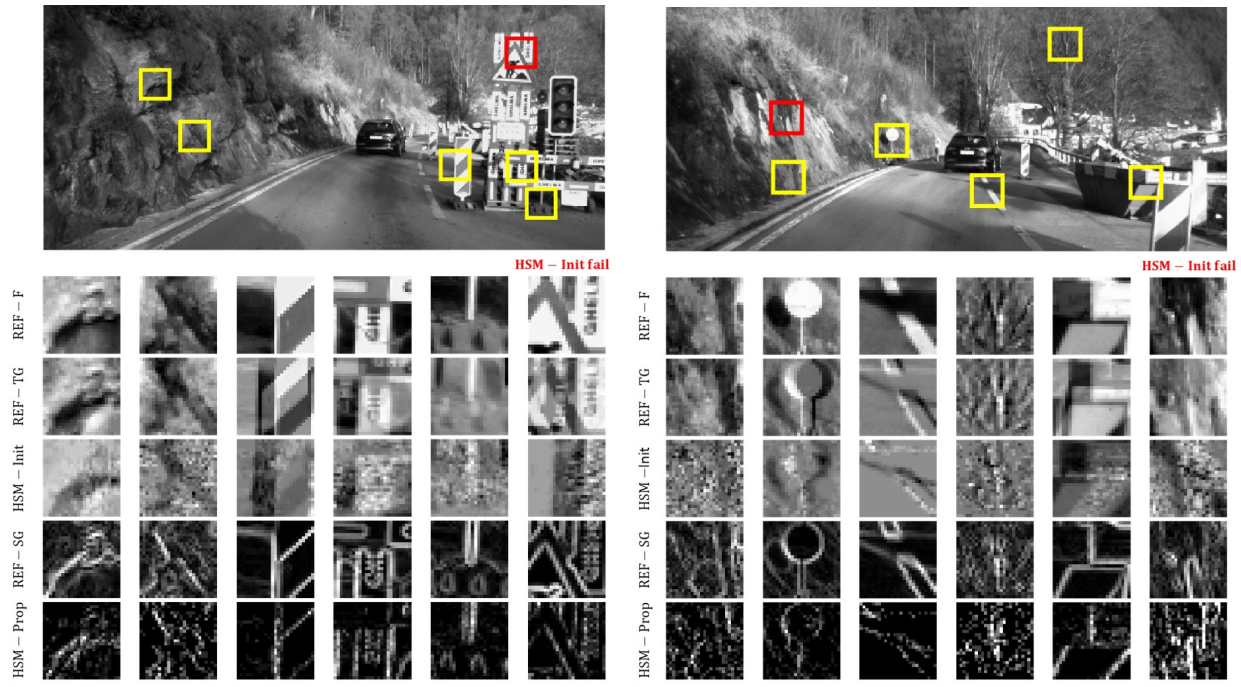


Fig. 7: Snapshots of the hetero-stereo matching results. The reference frame images in the first row (REF-F) are magnified images of the yellow squared patch, and the red squared patches indicate the mismatched patches in HSM-Init method (the last column). The second row (REF-TG) indicates the corresponding reference temporal gradient of the frame images ${}^{F}\mathbf{I}_n^{\Delta\tau}$, and the third row represents the matched event images ${}^{E}\mathbf{I}_n^{\Delta\tau}$. The fourth row (REF-SG) and the last row are showing the edge images ${}^{F}\mathbf{I}_n^{\Delta\mathbf{x}}$ (the norm of spatial gradient image) obtained from the frame images and the aligned event images ${}^{E}\mathbf{I}_n^{\Delta\mathbf{x}}$ of the proposed method, respectively.

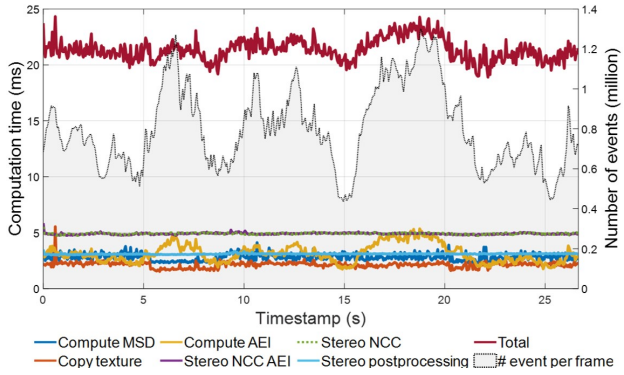


Fig. 8: Graph of computation time per frame (ms). Computation time to construct aligned event images is proportional to the number of events.

We presented a method using a temporal gradient image to perform initialization within a few frames to estimate the camera pose, and proposed an accurate, efficient and intuitive method for aligning events utilizing camera motion. We proposed the warping method considering the different depth of asynchronous events, and the maximum shift distance method to use fewer aligned event images for real-time performance. The proposed method describes edges using a much smaller number of events through the aligning events with camera motion. We verified the method with several experiments, which confirm that the proposed method outperforms the other method for the inlier percentage and matching accuracy. We expect that our approach will improve the capability of using frame and event camera and the provided code will contribute to the event camera community.

REFERENCES

[1] G. Gallego, H. Rebecq, and D. Scaramuzza, "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3867–3876.

[2] G. Gallego, M. Gehrig, and D. Scaramuzza, "Focus is all you need: Loss functions for event-based vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 280–12 289.

[3] T. Stoffregen and L. Kleeman, "Event cameras, contrast maximization and reward functions: an analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 300–12 308.

[4] J. Hidalgo-Carrió, G. Gallego, and D. Scaramuzza, "Event-aided direct sparse odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5781–5790.

[5] S. Shiba, Y. Aoki, and G. Gallego, "Event collapse in contrast maximization frameworks," *Sensors*, vol. 22, no. 14, p. 5190, 2022.

[6] H. Kim and H. J. Kim, "Real-time rotational motion estimation with contrast maximization over globally aligned events," *IEEE Robotics and Automation Letters*, 2021.

[7] A. Z. Zhu, Y. Chen, and K. Daniilidis, "Realtime time synchronized event-based stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 433–447.

[8] Y. Zhou, G. Gallego, H. Rebecq, L. Kneip, H. Li, and D. Scaramuzza, "Semi-dense 3d reconstruction with a stereo event camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 235–251.

[9] A. Hadviger, I. Marković, and I. Petrović, "Stereo event lifetime and disparity estimation for dynamic vision sensors," in *2019 European Conference on Mobile Robots (ECMR)*. IEEE, 2019, pp. 1–6.

[10] A. Andreopoulos, H. J. Kashyap, T. K. Nayak, A. Amir, and M. D. Flickner, "A low power, high throughput, fully event-based stereo system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7532–7542.

[11] D. Zou, P. Guo, Q. Wang, X. Wang, G. Shao, F. Shi, J. Li, and P.-K. Park, "Context-aware event-driven stereo matching," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1076–1080.

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

[13] S.-H. Ieng, J. Carneiro, M. Osswald, and R. Benosman, "Neuromorphic event-based generalized time-based stereovision," *Frontiers in neuroscience*, vol. 12, p. 442, 2018.

[14] D. Zou, F. Shi, W. Liu, J. Li, Q. Wang, P.-K. Park, C.-W. Shi, Y. J. Roh, and H. E. Ryu, "Robust dense depth map estimation from sparse dvs stereos," in *British Mach. Vis. Conf.(BMVC)*, vol. 1, 2017.

[15] L. A. Camunas-Mesa, T. Serrano-Gotarredona, S. H. Ieng, R. B. Benosman, and B. Linares-Barranco, "On the use of orientation filters for 3d reconstruction in event-driven stereo vision," *Frontiers in neuroscience*, vol. 8, p. 48, 2014.

[16] S. Ghosh and G. Gallego, "Multi-event-camera depth estimation and outlier rejection by refocused events fusion," *Advanced Intelligent Systems*, p. 2200221, 2022.

[17] A. Hadviger, I. Marković, and I. Petrović, "Stereo dense depth tracking based on optical flow using frames and events," *Advanced Robotics*, vol. 35, no. 3-4, pp. 141–152, 2021.

[18] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2822–2829, 2021.

[19] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "Eklt: Asynchronous photometric feature tracking using events and frames," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 601–618, 2020.

[20] Z. Wang, Y. Ng, C. Scheerlinck, and R. Mahony, "An asynchronous kalman filter for hybrid event cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 448–457.

[21] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens: Event-based video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 155–16 164.

[22] Z. Wang, L. Pan, Y. Ng, Z. Zhuang, and R. Mahony, "Stereo hybrid event-frame (shef) cameras for 3d perception," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 9758–9764.

[23] Y.-F. Zuo, L. Cui, X. Peng, Y. Xu, S. Gao, X. Wang, and L. Kneip, "Accurate depth estimation from a hybrid event-rgb stereo setup," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6833–6840.

[24] 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.

[25] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[26] X. Zhang, W. Liao, L. Yu, W. Yang, and G.-S. Xia, "Event-based synthetic aperture imaging with a hybrid network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 235–14 244.

[27] Y. Zou, Y. Zheng, 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*, 2021, pp. 2024–2033.

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

[29] T. Ke and S. I. Roumeliotis, "An efficient algebraic solution to the perspective-three-point problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7225–7233.

[30] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.

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