# Depth from Asymmetric Frame-Event Stereo: A Divide-and-Conquer Approach

Xihao Chen Wenming Weng Yueyi Zhang Zhiwei Xiong*
University of Science and Technology of China
{xhchen10, wmweng}@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn

## Abstract

*Event cameras asynchronously measure brightness changes in a scene without motion blur or saturation, while frame cameras capture images with dense intensity and fine details at a fixed rate. The exclusive advantages of the two modalities make depth estimation from Stereo Asymmetric Frame-Event (SAFE) systems appealing. However, due to the inevitable information absence of one modality in certain challenging regions, existing stereo matching methods lose efficacy for asymmetric inputs from SAFE systems. In this paper, we propose a divide-and-conquer approach that decomposes depth estimation from SAFE systems into three sub-tasks, i.e., frame-event stereo matching, frame-based Structure-from-Motion (SfM), and event-based SfM. In this way, the above challenging regions are addressed by monocular SfM, which estimates robust depth with two views belonging to the same functioning modality. Moreover, we propose a dual sampling strategy to construct cost volumes with identical spatial locations and depth hypotheses for different sub-tasks, which enables sub-task fusion at the cost volume level. To tackle the occlusion issue raised by the sampling strategy, we further introduce a temporal fusion scheme to utilize long-term sequential inputs with multi-view information. Experimental results validate the superior performance of our method over existing solutions.*

## 1. Introduction

Event cameras, based on bio-inspired neuromorphic sensors, output an asynchronous stream of *events*. An event is triggered by a pixel intensity change above a certain threshold and characterized by the corresponding pixel location, timestamp, and polarity. Due to the unique working principle, event cameras present several attractive advantages, including high dynamic range (>120 dB), high temporal resolution (in the order of microsecond), etc [6]. These advantages make event cameras a promising alternative to conventional frame cameras in challenging scenarios, such as challenging illumination or high-speed situations. However, event cameras primarily report events at the edges of objects, where brightness changes typically occur. They are incapable of outputting dense intensity, which makes it challenging to extract sufficient contextual information from event streams alone. On the other hand, frame cameras capture images with abundant information, such as color and texture, but they suffer from motion blur and severe saturation in challenging environments. Considering the complementary characteristics of two imaging principles, there is emerging research interest in constructing Stereo Asymmetric Frame-Event (SAFE) camera systems, consisting of *an event camera and a frame camera*, to solve long-standing challenges in various applications, including deblur [23], HDR imaging [10], SLAM [30], etc.

Recently, depth estimation from SAFE systems [32, 42] has also been explored, which aims to estimate accurate depth in various conditions. To conduct stereo matching from a pair of frame and event images (converted from event streams) with significant asymmetry, existing methods propose to normalize [32] or transform [15] different modalities to a unified form. However, modality asymmetry can not be eliminated or even mitigated in certain challenging regions due to the inevitable information absence of one modality, *e.g.*, high dynamic range regions for frame cameras and regions inside object contours for event cameras. Therefore, these methods are prone to fail in such conditions. We argue that depth in these regions can be inferred by monocular Structure-from-Motion (SfM)[1] in two consecutive views with the same functioning modality where high-quality signals are ready to estimate correspondence.

In this paper, we propose a divide-and-conquer approach to robustly estimate depth in various scenarios. Instead of only relying on frame and event images at the current time step, we utilize past information and decompose depth estimation from SAFE systems into three sub-tasks, including Frame-Event Stereo Matching (FE-StM), Frame-based SfM (F-SfM), and Event-based SfM (E-SfM). Concretely, for regular regions (*i.e.*, symmetric features can be extracted),

*Corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant 62131003.

[1]Given that SAFE systems aim at robotics and autonomous driving applications, it is reasonable to assume sequential inputs.

we resort to FE-StM in current stereo inputs. For challenging regions with information absence, we solve them by the SfM sub-task based on the functioning modality.

Moreover, we propose a dual sampling strategy to fuse different sub-tasks. Instead of re-projecting and merging the depth estimates of different sub-tasks, we fuse their aggregated cost volumes, which are deliberately aligned in terms of both spatial locations and depth hypotheses. In the strategy, the cost volumes of different sub-tasks are all constructed in the spatial locations and depth hypothesis planes of the same reference view. Specifically, to fuse the sub-task that does not involve the reference view, we dual sample pairs of candidates from two source views with the locations and depth planes of the reference view to construct a cost volume aligned to the others. The dual sampling strategy intuitively works for most regions without occlusion. For regions with occlusion, given that currently occluded regions may be correctly matched previously, we propose to utilize long-term sequential inputs. To this end, we introduce a temporal fusion scheme that caches the previous cost volume embeddings of different sub-tasks and propagates them to the current ones through 3D ConvLSTM cells.

To evaluate the performance of our method, we process the widely used stereo event camera dataset DSEC to formulate a SAFE dataset. Our method demonstrates distinct improvements compared with (i) the variants of representative stereo matching methods, (ii) existing asymmetric frame-event depth estimation methods, and (iii) potential solutions using the same sequential inputs as ours.

Contributions of this paper are summarized as follows:

- A novel divide-and-conquer approach to adequately utilize the complementary characteristics of SAFE systems.
- A dual sampling strategy to fuse different sub-tasks at the cost volume level without re-projection.
- The SoTA performance of asymmetric frame-event depth estimation on the DSEC dataset.

## 2. Related Works

**Symmetric Stereo.** As a classical computer vision task, stereo matching, conducted with symmetric stereo frame cameras by default, has been extensively studied for decades [11, 26] and substantially advanced by deep learning techniques recently [3, 14, 19]. Given the widely known limits of frame cameras in challenging scenarios, *e.g.* textureless or HDR scenarios, researchers have initiated exploration into alternative camera systems, such as active or passive stereo systems with near infrared (NIR) cameras [37, 39] and stereo event camera systems [1, 8, 28, 36, 40]. However, these systems with a single modality are still limited in certain scenarios. More recently, cross-modal symmetric stereo systems with cameras of different modalities on both sides have been proposed and demonstrate distinct performance in various scenarios [4, 20, 21].

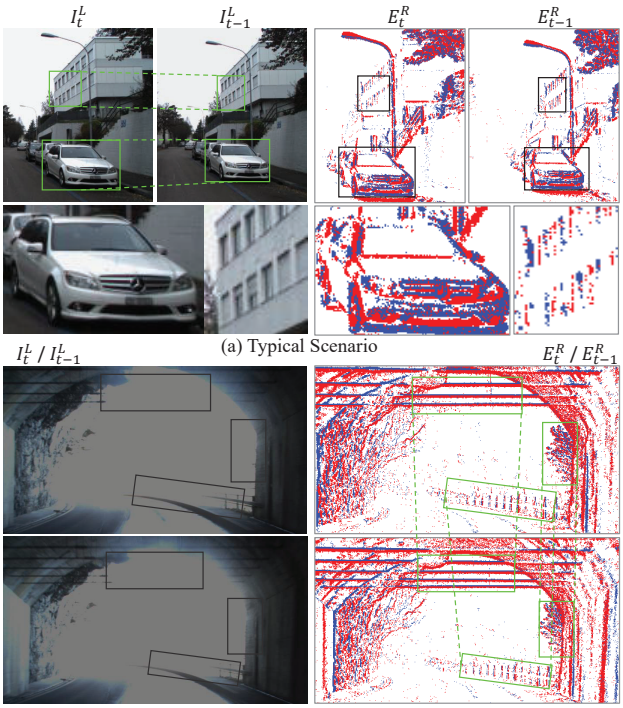


Figure 1. The complementary characteristics of SAFE systems. Frame cameras capture fine details in typical scenarios (a) while event cameras report high-quality signals in HDR scenarios (b). Depth in these regions should be inferred by the correspondence of two consecutive views, *i.e.*, $\mathbf{I}_t^L$ and $\mathbf{I}_{t-1}^L$ or $\mathbf{E}_t^R$ and $\mathbf{E}_{t-1}^R$.

**Asymmetric Stereo.** Compared with cross-modal symmetric stereo systems, asymmetric stereo systems with a single modality on one side, *e.g.,* frame-event [9, 32, 42] and RGB-NIR [16, 38] systems, possess the same sensing capabilities and come with half costs. For stereo matching based on asymmetric stereo, one key challenge is to handle asymmetry in different modalities in either handcrafted or learning-based ways. For example, edge images and temporal gradient images are adopted to normalize frame and event images [15, 32], while transformation networks are proposed to make up the photometric inconsistency of RGB and NIR images [16, 38]. However, there always is cross-modality information in asymmetric stereo that can not be normalized or translated and is thus overlooked. In this paper, we introduce monocular SfM to exploit the complementary information of SAFE systems, which exists in consecutive views with the same functioning modality.

**Multi-View Stereo.** In MVS, multiple images from different views are used to estimate the geometry of a scene or an object [27]. Typically, to boost the estimation of one reference view, multiple source views are matched with the reference jointly [34, 35] or respectively [12, 18]. Our method estimates depth based on the reference view and its depth plane hypotheses similar to MVS, but we exclusively match two source views with a dual sampling strategy to utilize the complementary information of different modalities.

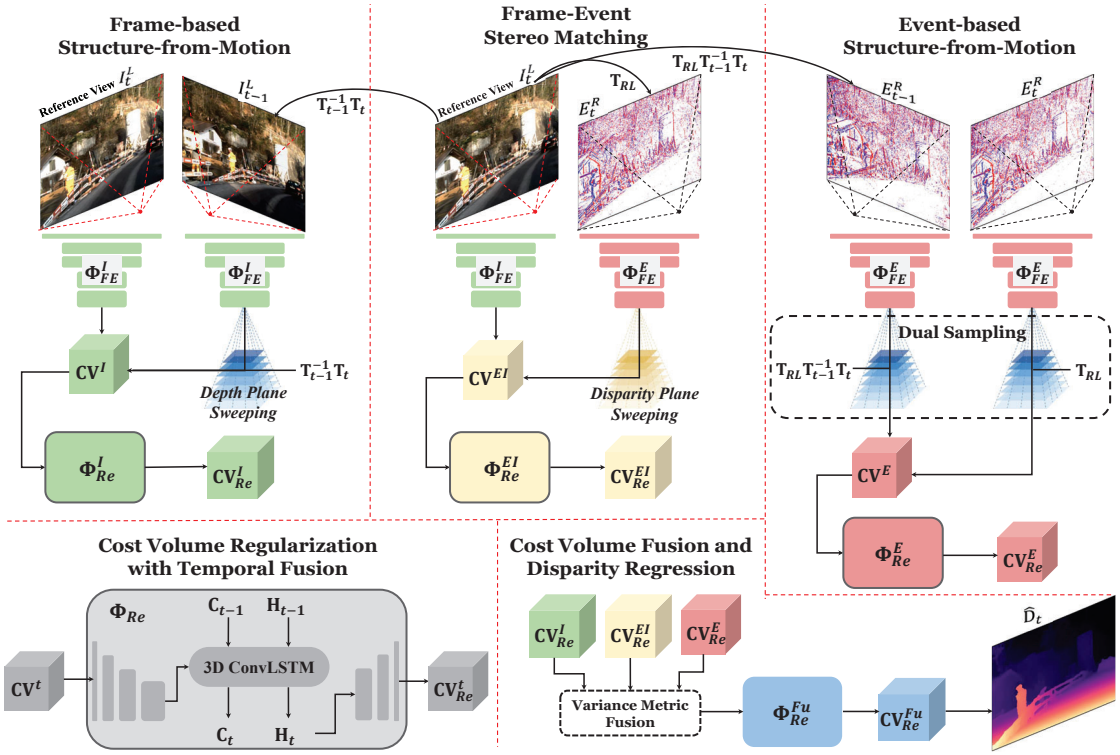


Figure 2. The proposed depth estimation method with a divide-and-conquer approach that decomposes depth from SAFE systems into three sub-tasks, *i.e.*, FE-StM, F-SfM, and E-SfM. With a dual sampling strategy, we fuse different sub-tasks at the cost volume level (*i.e.*, $\mathbf{CV}^{EI}_{Re}$, $\mathbf{CV}^{E}_{Re}$, and $\mathbf{CV}^{I}_{Re}$). Long-term sequential inputs are propagated by the temporal fusion scheme with 3D ConvLSTM.

## 3. Motivation

In a SAFE system, a frame camera and an event camera are used to perceive scenes with different modalities. Depth estimation from SAFE systems is expected to be accurate in various scenarios, because the event camera provides high-quality signals even in high dynamic range or high-speed regions while the frame camera provides clear intensity signals in most regions. However, these high-quality outputs may exist in only one modality due to the inherent limits of respective imaging principles. For example, in typical scenarios, the frame camera captures diverse information, *e.g.*, color, brightness, and texture, while the event camera only reports at object contours (see Fig. 1(a)). Under challenging illumination, the event camera maintains the output quality, while the frame camera suffers from severe saturation (see Fig. 1(b)). In these challenging regions with information absence, existing stereo matching methods may lose efficacy since symmetric features can not be extracted from such extremely asymmetric inputs from SAFE systems.

Instead of only relying on the current frame and event images, previous information from SAFE systems should be utilized. As can be seen in Fig. 1, high-quality signals exist in two views with the same functioning modality yet at different time steps, *e.g.*, the cables and the trees are consistently reported by the event camera in HDR scenarios. Given that SAFE systems aim at applications with frequent movement, *intra-modality* SfM, which infers depth by the correspondence of two consecutive views, should be a promising remedy for fragile *inter-modality* stereo matching in the above challenging regions.

## 4. Depth from SAFE Systems

Fig. 2 illustrates the proposed method working on SAFE systems. Without loss of generality, we take the frame camera as the left view and the event camera as the right view. Our depth estimation network $\Phi$ predicts depth $\widehat{\mathbf{D}}_t$ at the current time step $t$, *i.e.*,

$$\widehat{\mathbf{D}}_t = \Phi(\mathbf{I}^L_t, \mathbf{I}^L_{t-1}, \cdots, \mathbf{I}^L_{t-\delta}, \mathbf{T}_t, \mathbf{T}_{t-1}, \cdots, \mathbf{T}_{t-\delta}, \mathbf{E}^R_t, \mathbf{E}^R_{t-1}, \cdots, \mathbf{E}^R_{t-\delta}, \mathbf{T}_{RL}, \mathbf{K}; \theta), \tag{1}$$

with sequential inputs, including the camera intrinsic matrix $\mathbf{K}$, the transformation matrix from the left view to the right $\mathbf{T}_{RL} \in T(3)$, left view frame images $\mathbf{I}^L \in \mathbb{R}^{H\times W\times 3}$ and camera poses $\mathbf{T} \in SE(3)$, and the right view event stream $\xi^t_{t-\delta} = \{(\mathbf{x}_i, t_i, p_i) | t-\delta \le t_i \le t\}$. The event stream is converted into event images $\mathbf{E}^R \in \mathbb{R}^{H\times W\times B}$ according to the event representation proposed by Zhu *et al.* [41]. The supervised learning problem for depth estimation from SAFE systems can be formulated as

$$\theta^\star = \text{argmin}_\theta \, l\left(\widehat{\mathbf{D}}_t, \mathbf{D}_t\right), \tag{2}$$

where $l(\cdot,\cdot)$ is the loss between $\widehat{\mathbf{D}}_t$ and the ground truth $\mathbf{D}_t$.

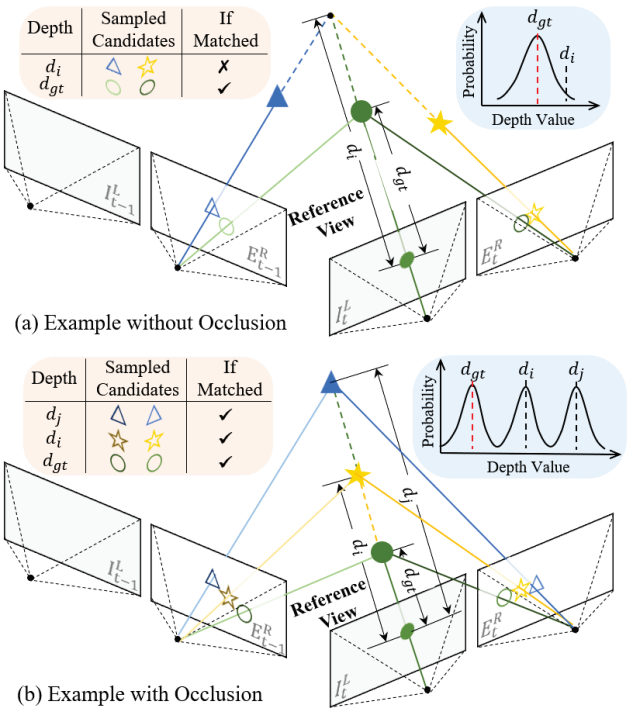


Figure 3. Dual sampling strategy produces uni-modal depth probability distributions to infer correct depth at most regions (a). The occlusion regions (b) with multi-modal distributions are handled by the temporal fusion scheme with long-term sequential inputs.

### 4.1. Dual Sampling

Originally, there are multiple reference views for different sub-tasks, *i.e.,* $\mathbf{I}_t^L$ for FE-StM and F-SfM and $\mathbf{E}_t^R$ for E-SfM. In other words, the outcomes of different sub-tasks are misaligned in terms of both spatial locations and depth (or disparity) plane hypotheses, and thus vulnerable re-projection with roughly estimated depth is required to fuse sub-tasks. To tackle this issue, we propose a dual sampling strategy to fuse sub-tasks at the cost volume level without any re-projection. Specifically, we choose the frame image at the current time step $\mathbf{I}_t^L$ as the only reference view and adopt a unified set of regularly discretized disparity plane hypotheses $\mathbb{P} = \{1, 2, \cdots, N\}$, where N is the maximum disparity value. For SfM sub-tasks, we convert $\mathbb{P}$ into equivalent depth plane hypotheses $\mathbb{D} = \{d_i = bf/p_i | p_i \in \mathbb{P}\}$, where $b$ and $f$ are the baseline distance and focal length of the SAFE system respectively. To construct the cost volumes of different sub-tasks, we sample candidates from source views, *i.e.*, $\mathbf{E}_t^R$, $\mathbf{E}_{t-1}^R$, and $\mathbf{I}_{t-1}^L$, according to their camera pose transformations $\mathbf{T}_{src \leftarrow ref}$ with regard to the reference view $\mathbf{I}_t^L$ and the hypothesized depth $d_i$ (or disparity $p_i$). Given a point in the reference view $\mathbf{x}^{ref} \doteq [u, v, 1]^T$, its corresponding candidate in a certain source view $\mathbf{x}_i^{src}$ is determined as

$$\mathbf{x}_i^{src} \sim \mathbf{K}[\mathbf{R} \mid \mathbf{t}] \begin{bmatrix} \left(\mathbf{K}^{-1}\mathbf{x}^{ref}\right) d_i \\ 1 \end{bmatrix}, \tag{3}$$

where $\mathbf{R}$ and $\mathbf{t}$ are the corresponding rotation matrix and translation vector of $\mathbf{T}_{src \leftarrow ref}$.

In MVS, whether the hypothesized $d_i$ is the actual depth or not is inferred by the similarity of $\mathbf{x}^{ref}$ and $\mathbf{x}_i^{src}$. Instead, our dual sampling strategy infers depth by the similarity of two candidate source points $\mathbf{x}_i^{src1}$ and $\mathbf{x}_i^{src2}$ for SfM sub-tasks. The strategy intuitively works in most regions as illustrated in Fig. 3(a). Specifically, for a point $\mathbf{x}^{ref}$ (a green circle), its two pairs of candidate points at the $i^{th}$ and ground-truth depth planes are $\mathbf{x}_i^{src1}$ (a yellow star) with $\mathbf{x}_i^{src2}$ (a blue triangle) and $\mathbf{x}_{gt}^{src1}$ with $\mathbf{x}_{gt}^{src2}$ (two green circles), respectively. Apparently, only one unique pair of candidate points on the ground-truth depth planes (*i.e.*, $\mathbf{x}_{gt}^{src1}$ and $\mathbf{x}_{gt}^{src2}$) that can be correctly matched because they correspond to the same scene point. In other words, the similarity of candidate pairs sampled according to the dual sampling strategy produces a uni-modal depth probability distribution to infer the correct depth. One exception is where more than one object is located in the ray from $\mathbf{x}^{ref}$, *i.e.*, scenarios with occlusion. In these regions, there are "pseudo" matched candidate pairs to interfere with depth estimation, *e.g.*, $\mathbf{x}_i^{src1}$ with $\mathbf{x}_i^{src2}$ (two yellow stars) and $\mathbf{x}_j^{src1}$ with $\mathbf{x}_j^{src2}$ (two blue triangles), as can be seen in Fig. 3(b). We solve these regions by infusing long-term temporal information, which is explained in Sec. 4.3.

### 4.2. Divide-and-Conquer

As aforementioned, to enable robust stereo depth estimation in challenging regions by utilizing the complementary information of SAFE systems, we explicitly decompose the task into three branches, including FE-StM between $\mathbf{I}_t^L$ and $\mathbf{E}_t^R$, F-SfM between $\mathbf{I}_t^L$ and $\mathbf{I}_{t-1}^L$, and E-SfM between $\mathbf{E}_t^R$ and $\mathbf{E}_{t-1}^R$. Before constructing the cost volumes for different branches, we extract features $\mathbf{F} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$ from frame and event images by two feature extractors $\Phi_I^{FE}$ and $\Phi_E^{FE}$ with the same architecture and independent weights,

$$\mathbf{F}^I = \Phi_{FE}^I(\mathbf{I}^L), \ \mathbf{F}^E = \Phi_{FE}^E(\mathbf{E}^R) \tag{4}$$

**Frame-Event Stereo Matching.** FE-StM conducts stereo matching using the reference view $\mathbf{I}_t^L$ and a source view $\mathbf{E}_t^R$. After feature extraction, the 4D cost volume $\mathbf{CV}^{EI} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \frac{N}{4} \times C_2}$ of FE-StM is constructed by the disparity plane sweeping, *i.e.*, concatenating and matching $\mathbf{F}_t^I$ with corresponding shifted $\mathbf{F}_t^E$ for each disparity $p_i$ in $\mathbb{P}$ ($N$ is scaled by 4 since we sweep at the downsampled features):

$$\mathbf{CV}^{EI}(u, v, p_i, \cdot) = \Phi_M^{EI}(\oplus\{\mathbf{F}_t^I(u, v, \cdot), \mathbf{F}_t^E(u - p_i, v, \cdot)\}), \tag{5}$$

where $\oplus$ stands for concatenation at the feature channel dimension and $\Phi_M^{EI}$ is a 2D convolutional matching module

to extract compact 3D features from the concatenated input, similar to those of [3,29]. Then, $\mathbf{CV}^{EI}$ is aggregated by an hourglass-like cost regularization module $\Phi_{Re}^{EI}$ consisting of multiple 3D convolutional layers, *i.e.*,

$$\mathbf{CV}_{Re}^{EI} = \Phi_{Re}^{EI}(\mathbf{CV}^{EI}). \tag{6}$$

**Event-based Structure-from-Motion.** E-SfM with two source views ($\mathbf{E}_t^R$ and $\mathbf{E}_{t-1}^R$) could be recognized as a two-view SfM problem [31] with known camera poses (*e.g.*, from the inertial measurement unit, IMU). According to the proposed dual sampling strategy, both $\mathbf{E}_t^R$ and $\mathbf{E}_{t-1}^R$ are required to be sampled by Eq. (3) with transformation matrices $\mathbf{T}_{RL}$ and $\mathbf{T}_{RL}\mathbf{T}_{t-1}^{-1}\mathbf{T}_t$ respectively. Note that the camera intrinsic matrix $\mathbf{K}$ is required to be scaled. With the depth plane sweeping, the cost volume $\mathbf{CV}^E$ of E-SfM is formulated as:

$$\mathbf{CV}^E(\mathbf{x}_t^I, p_i, \cdot) = \Phi_M^E(\oplus\{\mathbf{F}_t^E(\widetilde{\mathbf{x}}_t^E, \cdot), \mathbf{F}_{t-1}^E(\widetilde{\mathbf{x}}_{t-1}^E, \cdot)\}), \tag{7}$$

where $\widetilde{\mathbf{x}}_t^E$ and $\widetilde{\mathbf{x}}_{t-1}^E$ is the sampled source points of $\mathbf{E}_t^R$ and $\mathbf{E}_{t-1}^R$ corresponding to the reference point $\mathbf{x}_t^I$ of $\mathbf{I}_t^L$ at disparity plane $d_i = bf/p_i$ ($f$ is required to be scaled), and $\Phi_M^E$ is the matching module of the E-SfM branch. We use differentiable bilinear interpolation [13] to sample points since the transformed coordinates are not integers. Similar to the FE-StM branch, a regularization module $\Phi_{Re}^E$ aggregates the cost volume and output $\mathbf{CV}_{Re}^E$.

**Frame-based Structure-from-Motion.** To construct the cost volume $\mathbf{CV}^I$, F-SfM samples points $\widetilde{\mathbf{x}}_{t-1}^I$ from $\mathbf{I}_{t-1}^L$ with $\mathbf{T}_{t-1}^{-1}\mathbf{T}_t$ according to Eq. (3), *i.e.*,

$$\mathbf{CV}^I(\mathbf{x}_t^I, p_i, \cdot) = \Phi_M^I(\oplus\{\mathbf{F}_t^I(\mathbf{x}_t^I, \cdot), \mathbf{F}_{t-1}^I(\widetilde{\mathbf{x}}_{t-1}^I, \cdot)\}), \tag{8}$$

and matches them with reference points $\mathbf{x}_t^I$ from $\mathbf{I}_t^L$ by a matching module $\Phi_M^I$. The aggregated cost volume $\mathbf{CV}_{Re}^I$ is then generated by $\Phi_{Re}^I$.

**Cost Volume Level Fusion and Disparity Estimation.** After aggregation by regularization modules, the 4D aggregated cost volumes of different branches, *i.e.*, $\mathbf{CV}_{Re}^{EI}$, $\mathbf{CV}_{Re}^E$, and $\mathbf{CV}_{Re}^I$, not only contain abundant contextual information in different modalities but also implicitly encode depth probability distributions in different branches [14]. Although depth (disparity) can already be estimated from these cost volumes and then merged, a more flexible and reasonable fusion (*i.e.,* considering the pros and cons of different modalities in various regions) could be realized by a learning-based module with the information and depth hints in the cost volumes. Therefore, instead of clumsily merging their depth estimates, we propose to fuse at the cost volume level. In specific, we adopt the variance-based cost metric [34] to obtain a compact cost volume $\mathbf{CV}_{var}$. Then $\mathbf{CV}_{var}$ is further processed by a fusion module $\Phi_{Re}^{Fu}$ to obtain the final cost volume $\mathbf{CV}_{Re}^{Fu}$, *i.e.*,

$$\mathbf{CV}_{Re}^{Fu} = \Phi_{Re}^{Fu}(\mathbf{CV}_{var}), \tag{9}$$

where $\Phi_{Re}^{Fu}$ consists of an hourglass-like 3D convolutional network to fuse and two 3D transposed convolutional layers to upsample. We obtain depth estimate $\widehat{\mathbf{D}}_t$ with the sub-pixel maximum a posteriori approximation proposed by [29] that computes the expectation around the hypothesized depth with minimum matching cost as the final depth.

### 4.3. Temporal Fusion

In our method, the outcomes of different branches are fused without depth re-projection thanks to the proposed dual sampling strategy. In non-occlusion regions that occupy the majority, the effectiveness of the strategy is demonstrated intuitively. In regions with occlusion, there are "pseudo" matched candidate pairs to interfere with depth estimation. Specifically, although these "pseudo" matched candidate pairs indeed reveal the exact geometry of certain 3D scene points (*e.g.*, the yellow star and the blue triangle in Fig. 3(b)), they can not infer the ground-truth depth for the reference view since the 3D scene points they represent are occluded by the scene points closer to the reference camera plane (*e.g.*, the green circle). In this way, the depth probability distributions in these regions are inherently multi-modal and may not indicate the correct depth.

To address this issue, we propose a temporal fusion scheme to utilize the long-term sequential inputs of SAFE systems given that currently occluded regions might be matched in previous time steps. Instead of explicitly propagating the previously estimated depth $\widehat{\mathbf{D}}_{t-1}$, we choose to propagate the intermediate features of different branches at the previous time step $t-1$ to those at the current time step $t$ by ConvLSTM cells following [5]. Specifically, we use the bottleneck features $\mathbf{X} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times\frac{N}{32}\times C_3}$ of regularization modules as the inputs of 3D ConvLSTM cells, which are the variants with 3D convolutional layers to process the 4D features. The 3D ConvLSTM cells fuse the past scene geometry encoded at the previous hidden state $\mathbf{H}_{t-1}$ and cell state $\mathbf{C}_{t-1}$ with the current geometry encoded at $\mathbf{X}_t$ and output the current states $\mathbf{H}_t$ and $\mathbf{C}_t$:

$$\mathbf{H}_t, \mathbf{C}_t = \text{cell}(\mathbf{X}_t, \mathbf{H}_{t-1}, \mathbf{C}_{t-1}). \tag{10}$$

The detailed logic inside 3D ConvLSTM cells is as follows:

$$\begin{aligned}
\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t &= \text{split}(\text{sigmoid}\,(\mathbf{w}_x * \mathbf{X}_t + \mathbf{w}_h * \mathbf{H}_{t-1})) \\
\mathbf{g}_t &= \text{ELU}\,(\text{layernorm}\,(\mathbf{w}_{xg} * \mathbf{X}_t + \mathbf{w}_{hg} * \mathbf{H}_{t-1})) \\
\mathbf{C}_t &= \text{layernorm}\,(\mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t) \\
\mathbf{H}_t &= \mathbf{o}_t \odot \text{ELU}\,(\mathbf{C}_t)\,,
\end{aligned} \tag{11}$$

where $*$ and $\odot$ denote 3D convolution and Hadamard product while $\mathbf{i}_t, \mathbf{f}_t$, and $\mathbf{o}_t$ are the gates of the 3D ConvLSTM cells. After temporal fusion, the hidden states $\mathbf{H}_t$ with enhanced scene geometry are fed into the decoder part of regularization modules as illustrated in Fig. 2. Such a temporal fusion scheme is conducted in the regularization modules of different sub-tasks and the final fusion module.

Table 1. Comparison of different methods on the DSEC dataset. The best results are highlighted with **bold** fonts.

| Method | Disparity Metrics | | | | | Depth Metrics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE ↓ | RMSE ↓ | 1PE ↓ | 2PE ↓ | 3PE ↓ | MAE ↓ | $MAE_{rel}$ ↓ | RMSE ↓ | $\Delta < 1.05^1$ ↑ | $\Delta < 1.05^2$ ↑ | $\Delta < 1.05^3$ ↑ |
| RAFT-Stereo [17] | 0.6263 | 1.2120 | 15.38 | 3.59 | 1.64 | 1.0689 | 0.0395 | 2.1439 | 77.83 | 92.38 | 96.39 |
| AANet [33] | 0.6555 | 1.3243 | 15.28 | 3.92 | 2.01 | 1.1119 | 0.0416 | 2.2804 | 77.62 | 91.94 | 95.85 |
| PSMNet [2] | 0.5886 | 1.1968 | 13.15 | 2.87 | 1.40 | 1.0063 | 0.0374 | 2.0437 | 79.38 | 93.44 | 96.91 |
| E2VID [24] + RAFT-Stereo [17] | 0.7788 | 1.4640 | 21.95 | 5.78 | 2.56 | 1.3165 | 0.0484 | 2.5197 | 70.03 | 88.7 | 94.41 |
| E2VID [24] + AANet [33] | 0.7757 | 1.4376 | 22.08 | 5.70 | 2.44 | 1.2987 | 0.0480 | 2.4529 | 70.30 | 88.74 | 94.65 |
| E2VID [24] + PSMNet [2] | 0.6971 | 1.3085 | 18.55 | 4.38 | 1.92 | 1.1997 | 0.0446 | 2.3105 | 73.25 | 90.54 | 95.52 |
| HDES [42] | 0.6978 | 1.3074 | 19.02 | 4.97 | 2.14 | 1.1617 | 0.0432 | 2.2675 | 74.38 | 80.48 | 95.43 |
| DCNet [32] | 0.5869 | 1.1927 | 13.05 | 2.85 | 1.38 | 1.0052 | 0.0373 | 2.0378 | 79.43 | 93.47 | 96.93 |
| Ours | **0.5434** | **1.1751** | **10.97** | **2.23** | **1.09** | **0.9548** | **0.0353** | **2.0028** | **81.54** | **94.03** | **97.26** |

# 5. Experiments

## 5.1. Dataset and Evaluation Metrics

To evaluate the performance of our method, we opt for the DSEC dataset [7], which comprises a pair of frame and event cameras on both left and right sides. DSEC is a large scale stereo dataset collected by driving in various challenging scenarios. It provides high-quality ground-truth disparity for the development and evaluation of different stereo methods. We choose the left frame camera and the right event camera to formulate a SAFE system. The left frame images with the resolution of $1440 \times 1080$ are downsampled to the resolution of the event camera ($640 \times 480$). We conduct stereo rectification for the SAFE system with the provided camera intrinsic and extrinsic matrices and distortion coefficients. The disparity of DSEC is provided in the coordinate of the original left frame camera. We re-project the disparity data with *higher* resolution into the rectified left camera of the SAFE system. Note that we do not submit estimated disparity to the benchmark website of DSEC, because it requires estimated disparity maps in the view of the original left camera but re-projecting disparity maps from those with *the same* resolution would suffer from grid-like artifacts. Moreover, the disparity of the original test set is not released. Therefore, we randomly split 35 sequences of the original training set (41 sequences) on DSEC as the new training set and the others as the new test set.

Our method is evaluated by standard metrics in terms of both disparity and depth. The standard disparity metrics include mean absolute error (MAE), root mean squared error (RMSE), and N-Pixel-Error (NPE). The standard depth metrics include MAE, mean absolute relative error ($MAE_{rel}$), RMSE, and inlier ratios with threshold $1.05$, $1.05^2$, and $1.05^3$ ($\Delta < 1.05^n$). All these metrics are the lower the better except for the inlier ratios.

## 5.2. Implementation Details

We set the resolution of inputs as 640×480 and the maximum disparity as 96. We use the event representation in [41] to convert events within a duration of 50 ms into an event image with the number of time bins $B = 15$. We first train our network $\Phi$ without the temporal fusion scheme with a learning rate of 0.0001 and use pre-trained weights to initialize $\Phi$ except for 3D ConvLSTM cells. $\Phi$ is then fine-tuned with an initial learning rate of 0.0001 and a decreased learning rate of 0.00005 after the $50,000^{th}$ iteration. The loss of both stages $l(\cdot, \cdot)$ is the sub-pixel cross-entropy proposed in [29] with the diversity of the Laplace distribution $b = 2$. We use the ADAM solver ($\beta_1$=0.9, $\beta_1$=0.99) and set batch size as 2 and subsequence length $\delta$ as 3.

## 5.3. Comparison on SAFE systems

**Methods.** We adopt three categories of comparison methods. The first category includes three representative stereo matching networks with different cost volume aggregation ideas, *i.e.,* PSMNet [2] with 3D convolutional layers, AANet [33] with 2D deformable convolutional layers, and RAFT-Stereo [17] with 2D convolutional GRUs. We use two feature extractors with the same architecture and independent weights for them to extract features from frame and event images. The second category uses E2VID [24] to reconstruct intensity images from event streams and then conduct stereo matching with the methods adopted in the first category. The methods in the third category are specifically designed for SAFE systems, including DCNet [32] (a depth completion network that combines the dense estimate of PSMNet and the sparse estimate computed with the *edge* maps of frame and event images), HDES [42] (a depth estimation network with pyramid attention), and ours.

**Quantitative Results.** Table 1 shows the quantitative results of different methods on the DSEC datasets. Among the methods from the first category, PSMNet demonstrates the best performance. It is different from the results of symmetric frame-based stereo matching where AANet and RAFT-Stereo are the more advanced network architectures than PSMNet. We attribute the discrepancy to the difficulty of asymmetric stereo matching. The 3D cost volumes of AANet and RAFT-Stereo constructed by the correlation of asymmetric features are insufficient to reveal the actual geometry of scenes, while the 4D cost volume of PSMNet contains more contextual information and benefits the cost aggregation with asymmetric inputs. The performance of the second category degrades severely since the E2VID framework loses efficacy and generates unsatisfactory in-

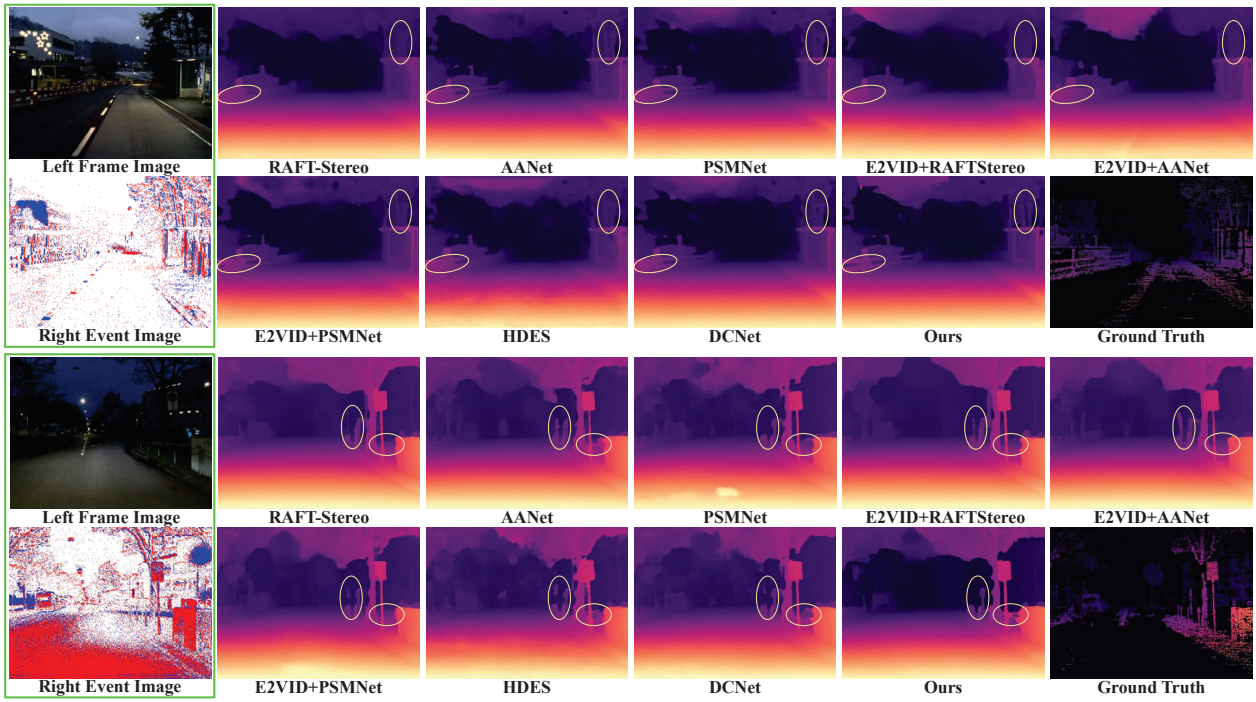


Figure 4. Disparity maps of two exemplar scenes on the DSEC dataset. Our method produces the best results in regions with complex geometry (*e.g.*, the railings and trunks in the first scene) and regions with challenging illumination (*e.g.*, the chains in the second scene).

tensity images, exacerbating the difficulty of matching. For the third category, HDES does not construct cost volume to capture correspondence and thus presents inferior performance, while DCNet obtains negligible improvements over PSMNet indicating that the asymmetry caused by *the information absence of a certain modality* can not be mitigated by the edge maps. In contrast, our method introduces SfM sub-tasks to estimate reliable depth from two consecutive views with *the same functioning modality*, thus outperforming all comparison methods by a large margin in all metrics.

**Visual Results.** The visual results of two exemplar scenes are shown in Fig. 4. In these scenes, comparison methods that only rely on the inputs at the current time steps can not present the correct disparity, since a certain modality fails to report high-quality signals in some challenging regions, *e.g.*, the railing and the trunks in the first scene for event cameras and the chains in the second one for frame cameras. In contrast, our method obtains more robust results, indicating the effectiveness of our divide-and-conquer approach that solves these regions by SfM sub-tasks.

## 5.4. Comparison with Symmetric Systems

To demonstrate the advantages of SAFE systems over stereo symmetric frame-based (FF) and event-based (EE) systems, we adopt two representative methods proposed for these systems, including PSMNet [2] and DDES [28] (denoted as FF-PSMNet and EE-DDES). Instead of a quantitative comparison, we conduct a qualitative comparison on the DSEC dataset, because three system setups on DSEC have considerably different baseline distances and focal

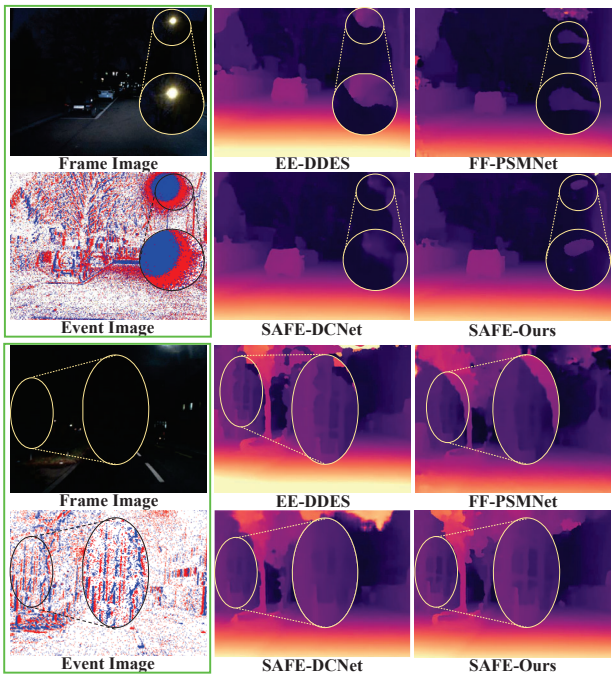


Figure 5. Visual comparison of SAFE systems with stereo symmetric event-based (EE-DDSE) and frame-based (FF-PSMNet) systems. SAFE-Our estimates robust depth in various scenarios while the others fail in certain scenarios (*e.g.*, EE-DDSE in the flickering light region and FF-PSMNet in the low-light region).

Table 2. Comparison of the dual sampling strategy with other possible solutions using the same sequential inputs of SAFE systems.

| Method | Disparity Metrics | | | | | Depth Metrics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE ↓ | RMSE ↓ | 1PE ↓ | 2PE ↓ | 3PE ↓ | MAE ↓ | $MAE_{rel}$ ↓ | RMSE ↓ | $\Delta < 1.05^1$ ↑ | $\Delta < 1.05^2$ ↑ | $\Delta < 1.05^3$ ↑ |
| Depth Re-projection (Forward Warping) | 0.6555 | 1.3734 | 15.18 | 3.98 | 2.08 | 1.1111 | 0.0410 | 2.2281 | 76.96 | 91.54 | 95.68 |
| Depth Re-projection (Backward Warping) | 0.6628 | 1.3744 | 15.80 | 4.09 | 2.04 | 1.1088 | 0.0411 | 2.2291 | 76.82 | 91.41 | 95.73 |
| MVSNet [34] | 0.5993 | 1.3176 | 12.85 | 2.69 | 1.33 | 1.0288 | 0.0392 | 2.1748 | 79.95 | 93.28 | 96.88 |
| Dual Sampling (w/o Temporal Fusion) | **0.5627** | **1.2307** | **11.56** | **2.40** | **1.17** | **0.9710** | **0.0364** | **2.0382** | **81.15** | **93.85** | **97.10** |

Table 3. Ablation study of different components in our method.

| | E-SfM | F-SfM | Temporal Fusion | Disparity Metrics | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | MAE ↓ | RMSE ↓ | 1PE ↓ | 2PE ↓ | 3PE ↓ |
| (A) | ✗ | ✗ | ✗ | 0.6084 | 1.3556 | 13.26 | 2.97 | 1.50 |
| (B) | ✓ | ✗ | ✗ | 0.5802 | 1.2934 | 12.05 | 2.70 | 1.41 |
| (C) | ✗ | ✓ | ✗ | 0.5780 | 1.2560 | 11.91 | 2.49 | 1.22 |
| (D) | ✓ | ✓ | ✗ | 0.5627 | 1.2307 | 11.56 | 2.40 | 1.17 |
| (E) | ✓ | ✓ | ✓ | **0.5434** | **1.1751** | **10.97** | **2.23** | **1.09** |

lengths. As can be seen in Fig. 5, our method in SAFE systems overcomes the limits of both modalities and presents high-quality results in these scenes. On the one hand, excess events are triggered in the flickering light region and event images suffer from "saturation" (*i.e.*, the first scene in Fig. 5). EE-DDES predicts over-smooth disparity, while our method in SAFE systems obtains sharper results similar to those of FF-PSMNet. On the other hand, frame cameras suffer in low-light scenarios and present rare details in dark regions compared with event cameras (*i.e.*, the second scene in Fig. 5). Therefore, FF-PSMNet can not reveal the 3D geometry in these regions as accurately as SAFE-Ours and EE-DDES. In short, our method realizes robust depth estimation in challenging scenarios by effectively leveraging the complementary characteristics of two modalities.

### 5.5. Ablation Study

**Dual Sampling**. To demonstrate the effectiveness of the dual sampling strategy, we compare our method with other possible solutions using the same sequential inputs of SAFE systems. The first solution is to re-project disparity or depth from the FE-StM, F-SfM, and E-SfM branches (without the sampling strategy). Specifically, the re-projection solution uses $\mathbf{I}_t^L$ as the reference view of FE-StM and F-SfM while using $\mathbf{E}_t^R$ as the reference of E-SfM. The estimate from E-SfM is re-projected to the coordinate of $\mathbf{I}_t^L$ by either differentiable forward [22] or backward warping [13]. The estimates of FE-StM and F-SfM and the re-projected estimate of E-SfM are fed into a U-Net [25] to obtain the final estimate. The second solution follows the spirit of MVSNet [34] which uses $\mathbf{I}_t^L$ as the reference view and the others as the source views and constructs a cost volume to aggregate by matching the reference view with the source views separately. As can be seen in Table 2, our method without the temporal fusion scheme (to guarantee equal input information) consistently outperforms these solutions in terms of all metrics. It clearly demonstrates the distinct advantage of the dual sampling strategy which avoids depth re-projection and enables the cost volume level fusion. In contrast, the re-projection solutions conduct independent matching in different branches and can only fuse the final depth estimates, thus overlooking the abundant contextual information and encoded depth or disparity probability distributions in the cost volumes. Although MVSNet aggregates information in a unified cost volume, it does not conduct matching between $\mathbf{E}_t^R$ and $\mathbf{E}_{t-1}^R$ to make full use of the exclusive information in event streams, indicating the inferior performance compared with our method.

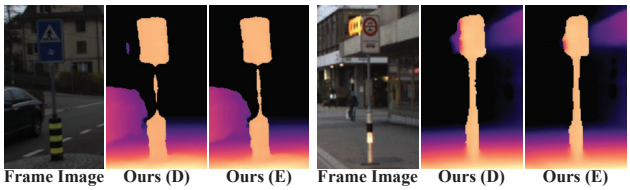


Figure 6. Visual comparison of our method with (E) and without (D) the temporal fusion scheme in regions with occlusion.

**Divide-and-Conquer**. To validate the effectiveness of different SfM sub-tasks in our method, we conduct ablation studies as can be seen in Table 3. When both SfM sub-tasks are ablated (*i.e.*, (A) in Table 3), our method degrades to asymmetric stereo matching and presents performance similar to PSMNet due to the consistency of cost aggregation. Both F-SfM ((B) *vs.* (A)) and E-SfM ((C) *vs.* (A)) contribute to the distinct performance of our method, demonstrating the effectiveness of our divide-and-conquer approach to exploit the exclusive advantages of both modalities.

**Temporal Fusion**. Compared with our method without the temporal fusion scheme (D), our full method (E) obtains considerate performance gains in all metrics (see Tab. 3), indicating the benefit of utilizing long-term temporal information. It is also demonstrated by the visual results in Fig. 6 where our full method estimates better geometry than that without temporal fusion for occlusion regions.

## 6. Conclusion

This paper aims to utilize the complementary characteristics of SAFE systems and realize robust depth estimation in scenarios that challenge the stereo symmetric systems with a single modality. As validated by experiments, our divide-and-conquer approach with a dual sampling strategy and a temporal fusion scheme demonstrates superior performance over various comparison methods. We expect our method could generalize to other stereo asymmetric systems and leave it as a future work.

## 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*, volume 35, pages 882–890, 2021. 2

[2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 6, 7

[3] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *NeurIPS*, 2020. 2, 5

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

[5] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *CVPR*, pages 15324–15333, 2021. 5

[6] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1

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

[8] Suman Ghosh and Guillermo Gallego. Multi-event-camera depth estimation and outlier rejection by refocused events fusion. *Advanced Intelligent Systems*, 4(12):2200221, 2022. 2

[9] Jinjin Gu, Jinan Zhou, Ringo Sai Wo Chu, Yan Chen, Jiawei Zhang, Xuanye Cheng, Song Zhang, and Jimmy S Ren. Self-supervised intensity-event stereo matching. *arXiv:2211.00509*, 2022. 2

[10] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on pattern analysis and machine intelligence*, 2023. 1

[11] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 2

[12] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *ICLR*, 2019. 2

[13] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 28, 2015. 5, 8

[14] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017. 2, 5

[15] Haram Kim, Sangil Lee, Junha Kim, and H Jin Kim. Real-time hetero-stereo matching for event and frame camera with aligned events using maximum shift distance. *IEEE Robotics and Automation Letters*, 8(1):416–423, 2022. 1, 2

[16] Mingyang Liang, Xiaoyang Guo, Hongsheng Li, Xiaogang Wang, and You Song. Unsupervised cross-spectral stereo matching by learning to synthesize. In *AAAI*, 2019. 2

[17] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227, 2021. 6

[18] Jiachen Liu, Pan Ji, Nitin Bansal, Changjiang Cai, Qingan Yan, Xiaolei Huang, and Yi Xu. Planemvs: 3d plane reconstruction from multi-view stereo. In *CVPR*, pages 8665–8675, 2022. 2

[19] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 2

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

[21] 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. 2

[22] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, pages 5437–5446, 2020. 8

[23] 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

[24] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 6

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 8

[26] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002. 2

[27] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006. 2

[28] 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. 2, 7

[29] Stepan Tulyakov, Anton Ivanov, and François Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. In *NeurIPS*, 2018. 5, 6

[30] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018. 1

[31] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *CVPR*, pages 8953–8962, 2021. 5

[32] Ziwei Wang, Liyuan Pan, Yonhon Ng, Zheyu Zhuang, and Robert Mahony. Stereo hybrid event-frame (shef) cameras for 3d perception. In *IROS*, pages 9758–9764, 2021. 1, 2, 6

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

[34] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018. 2, 5, 8

[35] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, pages 5525–5534, 2019. 2

[36] 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. 2

[37] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. In *ECCV*, pages 784–801, 2018. 2

[38] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *CVPR*, 2018. 2

[39] Jiageng Zhong, Ming Li, Xuan Liao, and Jiangying Qin. A real-time infrared stereo matching algorithm for rgb-d cameras' indoor 3d perception. *ISPRS International Journal of Geo-Information*, 9(8):472, 2020. 2

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

[41] 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. 3, 6

[42] Yi-Fan Zuo, Li Cui, Xin Peng, Yanyu Xu, Shenghua Gao, Xia Wang, and Laurent Kneip. Accurate depth estimation from a hybrid event-rgb stereo setup. In *IROS*, pages 6833–6840, 2021. 1, 2, 6