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 de-blur [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 cannot 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),

---

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

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

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.
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 even in high dynamic range or high-speed 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 cannot 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 \( \hat{D}_t \) at the current time step \( t \), i.e.,

\[
\hat{D}_t = \Phi(I^L_{t-\delta}, I^L_{t-1}, \cdots, I^L_{t-\delta}, T_t, T_{t-1}, \cdots, T_{t-\delta}, E^R_{t}, E^R_{t-1}, \cdots, E^R_{t-\delta}, T_{RL}, K; \theta),
\]

(1)

with sequential inputs, including the camera intrinsic matrix \( K \), the transformation matrix from the left view to the right \( T_{RL} \in T(3) \), left view frame images \( I^L \in \mathbb{R}^{H \times W \times 3} \) and camera poses \( T \in SE(3) \), and the right view event stream \( t'_{t-\delta} = \{(x_i, t_i, p_i)|t - \delta \leq t_i \leq t\} \). The event stream is converted into event images \( E^R \in \mathbb{R}^{H \times W \times B} \) 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

\[
\theta^* = \text{argmin}_\theta l \left( \hat{D}_t, D_t \right),
\]

(2)

where \( l(\cdot, \cdot) \) is the loss between \( D_t \) and the ground truth \( D_t \).
4.1. Dual Sampling

Originally, there are multiple reference views for different sub-tasks, i.e., $I_i^L$ for FE-StM and F-SfM and $E_i^R$ 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 $I_t^L$ 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 $\mathbb{D} = \{d_i = b f / p_i | p_i \in \mathbb{P}\}$, 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., $B_i^L$, $E_i^R$, $E_{i-1}^L$, and $I_{t-1}^L$, according to their camera pose transformations $T_{src-ref}$ with regard to the reference view $I_t^L$ and the hypothesized depth $d_i$ (or disparity $p_i$). Given a point in the reference view $x_i^{ref} = [u, v, 1]^T$, its corresponding candidate in a certain source view $x_i^{src}$ is determined as

$$x_i^{src} \sim K[R \cdot t] \left[ \frac{1}{(K^{-1}x_i^{ref}) d_i} \right],$$

where $R$ and $t$ are the corresponding rotation matrix and translation vector of $T_{src-ref}$.

In MVS, whether the hypothesized $d_i$ is the actual depth or not is inferred by the similarity of $x_i^{ref}$ and $x_i^{src}$. Instead, our dual sampling strategy infers depth by the similarity of two candidate source points $x_i^{src1}$ and $x_i^{src2}$ for SfM sub-tasks. The strategy intuitively works in most regions as illustrated in Fig. 3(a). Specifically, for a point $x_i^{ref}$ (a green circle), its two pairs of candidate points at the $i^{th}$ and ground-truth depth planes are $x_i^{src1}$ (a yellow circle) and $x_i^{src2}$ (a blue triangle) and $x_i^{gt1}$ with $x_i^{gt2}$ (two green circles), respectively. Apparently, only one unique pair of candidate points on the ground-truth depth planes (i.e., $x_i^{gt1}$ and $x_i^{gt2}$) 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 $x_i^{ref}$, i.e., scenarios with occlusion. In these regions, there are “pseudo” matched candidate pairs to interfere with depth estimation, e.g., $x_i^{src1}$ with $x_i^{src2}$ (two yellow stars) and $x_i^{gt2}$ with $x_i^{gt1}$ (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 $I_t^L$ and $E_i^R$, F-SfM between $I_t^L$ and $I_{t-1}^L$, and E-SfM between $E_i^L$ and $E_{i-1}^L$. Before constructing the cost volumes for different branches, we extract features $F \in \mathbb{R}^{l_f \times l_f \times c_f}$ from frame and event images by two feature extractors $\Phi_{FE}^f$ and $\Phi_{FE}^E$ with the same architecture and independent weights, $F^I = \Phi_{FE}^f(I_t^L)$, $F^E = \Phi_{FE}^E(E_i^R)$

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

$$C^E(u, v, p_i, \cdot) = \Phi_{M}^E(\oplus\{F^I_j(u, v, \cdot), F^E(u - p_i, v, \cdot)\}),$$

where $\oplus$ stands for concatenation at the feature channel dimension and $\Phi_{M}^E$ is a 2D convolutional matching module
to extract compact 3D features from the concatenated input, similar to those of [3, 29]. Then, \( \text{CV}^{E_1} \) is aggregated by an hourglass-like cost regularization module \( \Phi^{E_1}_{Re} \) consisting of multiple 3D convolutional layers, i.e.,

\[
\text{CV}^{E_1}_{Re} = \Phi^{E_1}_{Re}(\text{CV}^{E_1}).
\]

\( \text{CV}^{E_1}_{Re} \) is further processed by a fusion module \( \Phi^{F_u}_{Re} \) to obtain the final cost volume \( \text{CV}^{F_u} \).

Event-based Structure-from-Motion. E-SfM with two source views (\( E_t^R \) and \( E_t^{R-1} \)) 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 \( E_t^R \) and \( E_t^{R-1} \) are required to be sampled by Eq. (3) with transformation matrices \( T_{RL} \) and \( T_{RL}^{-1}T_t \), respectively. Note that the camera intrinsic matrix \( K \) is required to be scaled. By using the depth plane sweeping, the cost volume \( \text{CV}^E \) of E-SfM is formulated as:

\[
\text{CV}^E(x_t^r, p_t^r, \cdot) = \Phi^{E_M}(\{F_t^E(\tilde{x}_t^E, \cdot), F_{t-1}^E(\tilde{x}_{t-1}^E, \cdot)\}),
\]

where \( \tilde{x}_t^E \) and \( \tilde{x}_{t-1}^E \) are the sampled source points of \( E_t^R \) and \( E_t^{R-1} \), respectively. Since the transformed coordinates are not integers. Similar to the FE-StM branch, a regularization module \( \Phi^{E}_{Re} \) aggregates the cost volume and output \( \text{CV}^{E}_{Re} \).

Frame-based Structure-from-Motion. To construct the cost volume \( \text{CV}^I \), F-SfM samples points \( \tilde{x}_{t-1}^I \) from \( I_{t-1}^I \) with \( T_{t-1}^{-1}T_t \) according to Eq. (3), i.e.,

\[
\text{CV}^I(x_t^I, p_t^I, \cdot) = \Phi^{I_M}(\{F_t^I(\tilde{x}_t^I, \cdot), F_{t-1}^I(\tilde{x}_{t-1}^I, \cdot)\}),
\]

and matches them with reference points \( x_t^I \) from \( I_t^I \) by a matching module \( \Phi^{I_M} \). The aggregated cost volume \( \text{CV}^I_{Re} \) is then generated by \( \Phi^{I}_{Re} \).

Cost Volume Level Fusion and Disparity Estimation. After aggregation by regularization modules, the 4D aggregated cost volumes of different branches, i.e., \( \text{CV}^{E_1}, \text{CV}^E, \text{CV}^I, \text{CV}^{F_u} \), 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 \( \text{CV}_{var} \). Then \( \text{CV}_{var} \) is further processed by a fusion module \( \Phi^{F_u}_{Re} \) to obtain the final cost volume \( \text{CV}^{F_u}_{Re} \), i.e.,

\[
\text{CV}^{F_u}_{Re} = \Phi^{F_u}_{Re}(\text{CV}_{var}),
\]

where \( \Phi^{F_u}_{Re} \) consists of an hourglass-like 3D convolutional network to fuse and two 3D transposed convolutional layers to upsample. We obtain depth estimate \( \hat{D}_t \) 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 \( \hat{d}_{t-1} \), 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 \( X \in \mathbb{R}^{H \times W \times C} \) 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 \( H_{t-1} \) and cell state \( C_{t-1} \) with the current geometry encoded at \( X_t \) and output the current states \( H_t \) and \( C_t \):

\[
H_t, C_t = \text{cell}(X_t, H_{t-1}, C_{t-1}).
\]

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

\[
i_t, f_t, o_t = \text{split}(\text{sigmoid}(w_x^t \cdot X_t + w_h^t \cdot H_{t-1}))
\]

\[
g_t = \text{ELU}(\text{layernorm}(w_x^g \cdot X_t + w_h^g \cdot H_{t-1}))
\]

\[
C_t = \text{layernorm}(f_t \odot C_{t-1} + i_t \odot g_t)
\]

\[
H_t = o_t \odot \text{ELU}(C_t),
\]

where \( \odot \) denotes 3D convolution and Hadamard product while \( i_t, f_t, o_t \) are the gates of the 3D ConvLSTM cells. After temporal fusion, the hidden states \( H_t \) 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Disparity Metrics</th>
<th>Depth Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE↓, RMSE↓, 1PE↓, 2PE↓, 3PE↓</td>
<td>MAE↓, MAErel↓, RMSE↓, Δ &lt; 1.05°↓, Δ &lt; 1.05°↑, Δ &lt; 1.05°↑</td>
</tr>
<tr>
<td>RAFT-Stereo [17]</td>
<td>0.6263, 1.2120, 15.38, 3.59, 1.64</td>
<td>1.0689, 0.0395, 2.1439, 77.83, 92.38, 96.39</td>
</tr>
<tr>
<td>AANet [33]</td>
<td>0.6555, 1.3243, 15.28, 3.92, 2.01</td>
<td>1.1119, 0.0416, 2.2804, 77.62, 91.94, 95.85</td>
</tr>
<tr>
<td>PSMNet [2]</td>
<td>0.5886, 1.1968, 13.15, 2.87, 1.40</td>
<td>1.0639, 0.0380, 2.1439, 77.83, 92.38, 96.39</td>
</tr>
<tr>
<td>E2VID [24] + RAFT-Stereo [17]</td>
<td>0.5757, 1.4376, 22.08, 5.70, 2.44</td>
<td>1.3165, 0.0484, 2.5197, 70.03, 88.7, 94.41</td>
</tr>
<tr>
<td>E2VID [24] + AANet [33]</td>
<td>0.5757, 1.4376, 22.08, 5.70, 2.44</td>
<td>1.2987, 0.0480, 2.5297, 70.30, 88.74, 94.65</td>
</tr>
<tr>
<td>E2VID [24] + PSMNet [2]</td>
<td>0.6971, 1.3805, 18.55, 4.38, 1.92</td>
<td>1.1997, 0.0446, 2.3105, 73.25, 90.54, 95.52</td>
</tr>
<tr>
<td>HDES [42]</td>
<td>0.6978, 1.3074, 19.02, 4.97, 2.14</td>
<td>1.1617, 0.0432, 2.2675, 74.38, 80.48, 95.43</td>
</tr>
<tr>
<td>DCNet [32]</td>
<td>0.5869, 1.1927, 13.05, 2.85, 1.38</td>
<td>1.0052, 0.0373, 2.0378, 79.43, 93.47, 96.93</td>
</tr>
<tr>
<td>Ours</td>
<td>0.5434, 1.1751, 10.97, 2.23, 1.09</td>
<td>0.9548, 0.0353, 2.0208, 81.54, 94.03, 97.26</td>
</tr>
</tbody>
</table>

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 × 1080 are downsampled to the resolution of the event camera (640 × 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 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, 1.05², and 1.05³ (Δ < 1.05³). 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 Φ without the temporal fusion scheme with a learning rate of 0.0001 and use pre-trained weights to initialize Φ except for 3D ConvLSTM cells. Φ is then fine-tuned with an initial learning rate of 0.0001 and a decreased learning rate of 0.00005 after the 50, 000th iteration. The loss of both stages l(·, ·) is the sub-pixel cross-entropy proposed in [29] with the diversity of the Laplace distribution b = 2. We use the ADAM solver (β₁=0.9, β₂=0.99) and set batch size as 2 and subsequence length δ 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 ideas, are downsam-
tensity images, exacerbating the difficulty of matching. For
the third category, HDES does not construct cost volume
to capture correspondence and thus presents inferior perfor-
mance, while DCNet obtains negligible improvements over
PSMNet indicating that the asymmetry caused by the informa-
tion 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 outperform-
ing 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 cam-
eras. In contrast, our method obtains more robust results,
indicating the effectiveness of our divide-and-conquer ap-
proach that solves these regions by SfM sub-tasks.

**5.4. Comparison with Symmetric Systems**

To demonstrate the advantages of SAFE systems over
er stereoscopic frame-based (FF) and event-based (EE)
systems, we adopt two representative methods proposed for
these systems, including PSMNet [2] and DDES [28] (de-
noted as FF-PSMNet and EE-DDES). Instead of a quanti-
tative comparison, we conduct a qualitative comparison on

**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).

**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).
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 cannot 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 $I^E_t$ as the reference view of FE-StM and F-SfM while using $E^R_t$ as the reference of E-SfM. The estimate from E-SfM is re-projected to the coordinate of $I^F_t$ 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 $I^E_t$ 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 $E^R_t$ and $E^R_{t-1}$ to make full use of the exclusive information in event streams, indicating the inferior performance compared with our method.

#### Divide-and-Conquer

To validate the effectiveness of different SiM sub-tasks in our method, we conduct ablation studies as can be seen in Table 3. When both SiM 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 considerable 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.

---

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

<table>
<thead>
<tr>
<th>Method</th>
<th>E-SfM</th>
<th>F-SfM</th>
<th>Temporal Fusion</th>
<th>Disparity Metrics</th>
<th>Depth Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MAE $\downarrow$</td>
<td>RMSE $\downarrow$</td>
</tr>
<tr>
<td>Depth Re-projection (Forward Warping)</td>
<td>0.6555</td>
<td>1.3734</td>
<td>15.18</td>
<td>3.98</td>
<td>2.08</td>
</tr>
<tr>
<td>Depth Re-projection (Backward Warping)</td>
<td>0.6628</td>
<td>1.3744</td>
<td>15.80</td>
<td>4.09</td>
<td>2.04</td>
</tr>
<tr>
<td>MVSNet [34]</td>
<td>0.5993</td>
<td>1.3176</td>
<td>12.85</td>
<td>2.69</td>
<td>1.33</td>
</tr>
<tr>
<td>Dual Sampling (w/o Temporal Fusion)</td>
<td>0.5627</td>
<td>1.2307</td>
<td>11.56</td>
<td>2.40</td>
<td>1.17</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Temporal Fusion</th>
<th>Disparity Metrics</th>
<th>MAE $\downarrow$</th>
<th>RMSE $\downarrow$</th>
<th>1PE $\downarrow$</th>
<th>2PE $\downarrow$</th>
<th>3PE $\downarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-SfM</td>
<td><strong>✓</strong></td>
<td>0.6084</td>
<td>13.556</td>
<td>2.97</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-SfM</td>
<td><strong>✓</strong></td>
<td>0.5802</td>
<td>12.934</td>
<td>2.70</td>
<td>1.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal Fusion</td>
<td><strong>✓</strong></td>
<td>0.5780</td>
<td>12.560</td>
<td>2.49</td>
<td>1.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-SfM</td>
<td><strong>✓</strong></td>
<td>0.5627</td>
<td>12.307</td>
<td>2.40</td>
<td>1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-SfM</td>
<td><strong>✓</strong></td>
<td>0.5434</td>
<td>11.751</td>
<td>2.23</td>
<td>1.09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Figure 6](image_url)
References


[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


