Asymmetric Event-Image Stereo with Temporal Feature Gating and Iterative Structure-Detail Refinement

Hao Zhuang[1]
haozhuang@bit.edu.cn
Yan Yang[2]
yan.yang@anu.edu.au
Liyuan Pan [1 †]
liyuan.pan@bit.edu.cn

[1] School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

[2] Australian National University
Canberra, ACT, Australia

Abstract

This paper studies disparity estimation using asymmetric stereo inputs, combining event streams and conventional frame-based images. Past works directly calculate the similarity between features from the two modalities for making predictions, neglecting the inherent domain shift between them. First, event data are both temporally and spatially aware, while a frame is spatially aware. Second, event data are well-suited for capturing scene structure, whereas frame-based images excel at preserving fine-grained local visual details. To address these challenges, our method introduces two key components: (1) We enhance frame features by gating them with temporal cues extracted from consecutive frames, which improves their temporal awareness for more accurate similarity computation. (2) We perform disparity estimation using an iterative refinement scheme that alternates between updating structure and local features. This process starts with reliable structure features and progressively refining the disparity prediction by integrating fine detail. Extensive experiments on the MVSEC and DSEC datasets show that our method significantly outperforms prior state-of-the-art approaches.

1 Introduction

Event cameras asynchronously detect pixel-level intensity changes by recording the pixel location, polarity (direction of change), and timestamp. They offer a high dynamic range and microsecond-level temporal resolution, making them effective in latency-sensitive or challenging lighting conditions [25, 26, 40]. However, events are typically concentrated along object edges in time, leading to spatial sparsity [9, 12]. In contrast, traditional frame-based cameras produce dense intensity images that are rich in visual detail, but they suffer from limitations such as motion blur, saturation, and poor performance in low-light environments. To exploit the complementary advantages of both modalities, researchers have developed hybrid stereo asymmetric frame-event (SAFE) vision systems, which integrate data from both

© 2025. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
† Corresponding author.

sensor types and have shown promise in disparity estimation [4, 30, 35, 41]. This paper studies the SAFE for disparity estimation.

Existing stereo systems are commonly based on RAFT-Stereo [20], where a cost volume is calculated by measuring the similarities between event- and frame-based inputs. Disparity estimates are then progressively refined through iterative updates, with local costs retrieved based on the current disparity prediction. Some SAFE methods further exploit the high temporal resolution of event data to enhance event feature representations, thereby improving the cost volume. Meanwhile, due to the spatial sparsity, the resulting disparity predictions can also be spatially sparse. To address this, other works have attempted to refine the predictions by applying depth completion techniques, filling in missing or unreliable regions. However, none of these approaches explicitly address the domain shift between the two modalities: event data, which are spatio-temporal and dominated by structural edges, and frame data, which are spatial and rich in local visual detail. This mismatch degrade feature similarity between the two modalities for the cost volume, and limit the disparity estimation.

In this paper, motivated by the observation that structure information is relatively easier to capture from frames, which could be well aligned with event data, we propose a SAFE-based method for disparity estimation that iteratively refines local detail predictions based on a reliable structure information. Our method follows the prior work [20], but overcomes the domain shift limitation with two key components. To improve the alignment between event and frame features for cost volume computation, we enhance frame features by gating them with temporal cues extracted from consecutive frames, which improves their temporal awareness for more accurate similarity computation.

Using the improved frame features along with event features, we compute a cost volume and perform iterative disparity estimation. In each iteration, we retrieve a local cost volume based on the previous disparity estimate and use it to predict structure-aware disparity features. Meanwhile, we apply cross-attention between the previous disparity prediction and the local cost volume to refine the disparity features with fine-grained local detail. The structure-aware features and detail-improved features are then used to produce an updated disparity prediction. This iterative process continues, progressively improving both structure consistency and local accuracy in the disparity estimation. Our contributions are:

- We propose a novel SAFE framework to address the domain shift between event and frame modalities in disparity estimation.
- We improve the alignment between event and frame features by improving the temporal representation of frame, and introduce a method that jointly refines structural and local details of disparity estimation.
- Extensive experiments on the MVSEC and DSEC datasets demonstrate that our method significantly outperforms prior state-of-the-art approaches.

2 Related Work

2.1 Single-Modality Stereo Matching

Image-Based Stereo. Image-based stereo matching is a fundamental task in computer vision. Recent advances in deep learning [34, 37] have significantly improved stereo matching performance, with existing methods generally categorized into three main types: aggregation-based approaches, iterative-based methods, and transformer-based methods. Aggregation-based methods such as AANet [33] utilize adaptive aggregation and deformable convolutions for efficient matching, while PSMNet [3] introduces spatial pyramid pooling combined with 3D hourglass networks to effectively capture contextual information. GWC-Net [14] further enhances matching accuracy by employing group-wise correlation strategies. Iterative-based methods, notably RAFT-Stereo [20], employ multi-scale correlation volumes and recurrent refinement modules for precise disparity estimation, and SStereo [29] introduces a novel update operator that adaptively fuses multi-frequency disparity information to preserve both structural detail and global consistency. Transformer-based methods, exemplified by STTR [19], reformulate stereo matching as a sequence-to-sequence correspondence problem, eliminating the need for cost volume construction. These approaches collectively demonstrate superior accuracy and robustness across various stereo matching benchmarks.

Event-Based Stereo. Although frame-based methods achieve high accuracy under ideal conditions, they often fail in extreme scenarios such as high dynamic range or motion blur. Recent methods address these challenges by aligning event streams with image-based stereo through frame-like representations [40]. Early work such as DDES [27] introduced learning-based embeddings using fully connected layers. To handle spatially uneven event densities, Se-cff [23] proposed multi-density stacked representations. DTC-SPADE [36] further integrated discrete-time convolution modules within a spatio-temporal recurrent structure. More recently, StereoFlow-Net [7] utilized flow-based temporal aggregation to enhance consistency in disparity estimation. However, the inherent sparsity of events still poses challenges for extracting fine-grained structural details.

2.2 Multi-Modality Stereo Matching

Cross-Modal Symmetric Stereo. Cross-modal symmetric stereo systems are designed with heterogeneous sensors mounted symmetrically on both sides. EIS [22] initiated research in this direction by developing a recurrent recycling structure [8] that transforms event stacks and grayscale frames into blur-free representations. Subsequently, SMC-Net [5] further refined event-image fusion by introducing an attention-based fusion module [28], adaptively integrating event and image features according to their spatial relevance. More recently, SCS-Net [6] advanced beyond previous methods by designing a differentiable event selection network, explicitly extracting relevant event features and effectively combining them with images via a neighbor cross-similarity feature (NCSF) module. These systems demonstrate superior performance in modality-specific scenarios.

Asymmetric Stereo. Unlike symmetric stereo setups, asymmetric systems combine different modalities on only one side of the stereo pair. The primary challenge lies in establishing direct pixel-wise correspondence across modalities. Prior methods address this either by directly transforming modalities or by mapping them into a unified intermediate representation. In the first category, SSIE-Stereo [13] reconstructs event data into intensity-like frames for stereo matching. The second category unifies modalities either spatially or temporally. Spatially-unified methods like DCNet [30] compute sparse disparities by correlating edge maps from event and intensity data. Temporally-unified methods, such as ZiE-Stereo [21], leverage visual prompting to produce aligned intermediate representations, while AFES-dc [4] employs camera pose estimation for improved temporal alignment. Sparse disparity estimates are further refined into dense depth maps using various strategies: direct CNN-based mappings (e.g., HDES [41]), explicit depth completion (e.g., DCNet [30]), and cost-volume-based aggregation (e.g., SSIE-Stereo [13], ZiE-Stereo [21], AFES-dc [4]). Transformer-based approaches such as TDANet [35] forgo explicit cost-volume construction, instead directly establishing dense correspondences through transformer modules.

3 Methodology

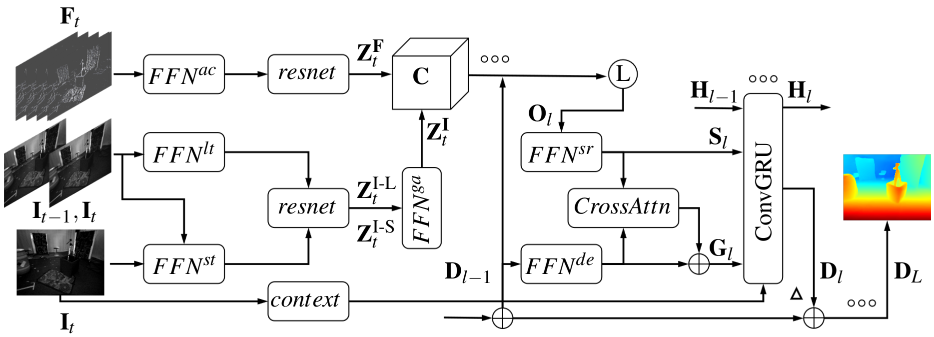


Figure 1: Overview of the proposed framework. We first extract event features $\mathbf{Z}_t^{\mathbf{F}}$ from the stacks of event frames $\mathbf{F}_t$, and the corresponding aligned frame features $\mathbf{Z}_t^{\mathbf{I}}$ from consecutive frames $\mathbf{I}_t$ and $\mathbf{I}_{t-1}$. A cost volume $\mathbf{C}$ is then calculated between $\mathbf{Z}_t^{\mathbf{F}}$ and $\mathbf{Z}_t^{\mathbf{I}}$ to iteratively predict disparity $\mathbf{D}$ over L stages. In each iteration, $1 \le l \le L$, we retrieve a local cost patch $\mathbf{O}_l$ based on the previous disparity $\mathbf{D}_{l-1}$, and together with the context features compute the structure-aware disparity features $\mathbf{S}_l$ and fine-grained disparity features $\mathbf{G}_l$, which are concatenated and fed into the ConvGRU update operator to update the hidden state $\mathbf{H}_l$ and predict the disparity estimate $\mathbf{D}_l$.

Overview. Given an event sequence $\{\mathbf{e}_n\}_{n=1}^{N}$ and a frame sequence $\{\mathbf{I}_t\}_{t=1}^{T}$ from a stereo setup, our goal is to estimate a disparity map $\mathbf{D}$ corresponding to frame $\mathbf{I}_t$, where $\mathbf{D}$ represents the pixel-wise horizontal displacement between the event and frame, N is the number of events, and T is the total number of time steps. Each event $\mathbf{e}_n = (\mathbf{u}_i, m_i, p_i)$ records the timestamp m_i, pixel location $\mathbf{u}_i = (x_i, y_i)$, and polarity p_i indicating the sign of brightness change. To predict the disparity, we select a window of events that occur up to the timestamp of frame $\mathbf{I}_t$, and convert this subset into stacks of event frames $\mathbf{F}_t$ to extract features $\mathbf{Z}_t^{\mathbf{F}}$. Meanwhile, the aligned frame features $\mathbf{Z}_t^{\mathbf{I}}$ are extracted from the frame $\mathbf{I}_t$ along with a nearby frame $\mathbf{I}_{t-1}$. We then calculate a cost volume $\mathbf{C}$ between them. By initializing disparity estimation $\mathbf{D}_0$ with zeros, we perform L iterative updates. In each iteration l, we retrieve a local cost patch $\mathbf{O}_l$ based on $\mathbf{D}_{l-1}$ and compute the structure-aware disparity features $\mathbf{S}_l$ and fine-grained disparity features $\mathbf{G}_l$ to refine $\mathbf{D}_{l-1}$ in $\mathbf{D}_l$. The final disparity map $\mathbf{D}_L$ is our output. An overview is in Fig. 1.

3.1 Network

Feature Extraction. We follow [23] to extract the event features $\mathbf{Z}_t^{\mathbf{F}}$ from stacks of event frames $\mathbf{F}_t$ with a ResNet encoder $resnet(\cdot)$ [15],

$$\mathbf{Z}_t^{\mathbf{F}} = resnet((FFN^{ac}(\mathbf{F}_t)) \ . \tag{1}$$

To improve the alignment between event and frame features for cost volume computation, we enhance frame features by gating them with temporal cues extracted from consecutive frames, which improves their temporal awareness for more accurate similarity computation. In a short-term temporal window around the current frame, brightness change can be estimated from spatial gradients and image-plane motion [10]. To capture such short-term cues, we compute frame features $\mathbf{Z}_t^{\text{I-S}}$ using a feedforward convolutional network $FFN^{st}(\cdot)$,

$$\mathbf{Z}_t^{\text{I-S}} = FFN^{st}(\mathbf{I}_t, \mathbf{I}_{t-1}) = resnet\left(\text{Edge}(\mathbf{I}_t) \cdot FFN^{mot}(\mathbf{I}_t, \mathbf{I}_{t-1})\right), \tag{2}$$

where $\text{Edge}(\cdot)$ applies a Sobel operator to extract spatial gradients from the current frame, and $FFN^{mot}(\cdot)$ adopts a lightweight U-Net-like structure to capture motion-aware features. The dot product encodes the edge-induced event generation mechanism, in which motion across intensity gradients triggers brightness change. Conversely, across the temporal interval between consecutive frames, the brightness change is modeled by intensity difference. Accordingly, we apply a log brightness difference (LBD) operator [21] to extract long-term temporal cues, which are then encoded into frame features $\mathbf{Z}_t^{\text{I-L}}$ using a feedforward convolutional network $FFN^{lt}(\cdot)$,

$$\mathbf{Z}_t^{\text{I-L}} = FFN^{lt}(\mathbf{I}_t, \mathbf{I}_{t-1}) = resnet\left(\text{LBD}(\mathbf{I}_t, \mathbf{I}_{t-1})\right), \tag{3}$$

where $\text{LBD}(\cdot)$ denotes the pixel-wise log intensity difference between consecutive frames. Finally, the short-term and long-term frame features are adaptively fused via a temporal gating module $FFN^{ga}(\cdot)$ to produce the final temporally aligned frame representation $\mathbf{Z}_t^{\mathbf{I}}$,

$$\mathbf{Z}_t^{\mathbf{I}} = FFN^{ga}\left(\mathbf{Z}_t^{\text{I-S}}, \mathbf{Z}_t^{\text{I-L}}\right). \tag{4}$$

This design enables the model to integrate motion-sensitive and temporally discriminative signals across different time scales, enhancing its ability to align with the event domain.

Cost Volume Calculation. We calculate the cost volume $\mathbf{C}$ by measuring the similarity between the event features $\mathbf{Z}_t^{\mathbf{F}}$ and the frame features $\mathbf{Z}_t^{\mathbf{I}}$. For each pixel (i, j) in $\mathbf{Z}_t^{\mathbf{F}}$, we evaluate the similarity between its feature and features at different horizontal disparities d in $\mathbf{Z}_t^{\mathbf{I}}$, i.e., at position $(i-d, j)$. The similarity is stored in the cost volume at position $\mathbf{C}(i, j, d)$,

$$\mathbf{C}(i, j, d) = \frac{\mathbf{Z}_t^{\mathbf{F}}(i, j)^\top \mathbf{Z}_t^{\mathbf{I}}(i-d, j)}{const}, \tag{5}$$

where *const* is a normalization factor for numerical stability, usually set to the square root of the feature dimension [20].

Disparity Prediction. We perform L iterative updates starting from an initial disparity estimation $\mathbf{D}_0$ (initialized as zeros) to arrive at the final disparity prediction $\mathbf{D}_L$. To provide global priors for the update process, we employ a context encoder to extract context features from $\mathbf{I}_t$. In each update step l, we retrieve local cost $\mathbf{O}_l$ from the cost volume $\mathbf{C}$ based on the disparity estimation $\mathbf{D}_{l-1}$ [32]. As the cost from event features and aligned frame features is more reliable in sparse edge regions, we use $\mathbf{O}_l$ as structural guidance and project it to the structure disparity features $\mathbf{S}_l$ with a feedforward network $FFN^{sr}(\cdot)$ composed of two convolutional layers,

$$\mathbf{S}_l = FFN^{sr}(\mathbf{O}_l). \tag{6}$$

To calculate fine-grained local detail disparity features $\mathbf{G}_l$, the previous disparity estimate $\mathbf{D}_{l-1}$ is first transformed by a feedforward network $FFN^{de}(\cdot)$ with two convolutional layers, and then used as the query in a cross-attention $CrossAttn(\cdot, \cdot)$ with structural disparity features $\mathbf{S}_l$. The output of the cross-attention is combined with the transformed disparity embedding through a residual addition, forming the final $\mathbf{G}_l$. This operation allows the model to selectively attend to informative regions within the local cost, and get refined disparity features $\mathbf{G}_l$ with local details,

$$\mathbf{G}_l = FFN^{de}(\mathbf{D}_{l-1}) + CrossAttn(FFN^{de}(\mathbf{D}_{l-1}), \mathbf{S}_l). \tag{7}$$

Finally, we use the update operator from [20], implemented as a ConvGRU, concatenates the context features with the structure-aware features $\mathbf{S}_l$ and the detail-enhanced features $\mathbf{G}_l$, updates the hidden state $\mathbf{H}_l$, and uses it to predict the disparity estimate $\mathbf{D}_l$. Since feature extraction is performed at 1/4 resolution, we predict the disparity field at 1/4 of the input resolution and recover the full-resolution map using the convex upsampling method [20].

3.2 Loss

To train our model, we employ an ℓ_1 loss that measures the deviation between the predicted disparity maps $\{\mathbf{D}_l\}_{l=1}^{L}$ and the ground truth disparity $\mathbf{D}^{gt}$. We progressively assign greater importance to later-stage predictions by applying an exponential decay weighting [32]. The final training loss is defined as:

$$\mathcal{L} = \sum_{l=1}^{L} \lambda^{L-l} \ell_1 \left(\mathbf{D}_l - \mathbf{D}^{gt}\right), \tag{8}$$

where λ is an exponential decay factor that down-weights earlier predictions.

Modality	Method	MDE [cm] ↓		MdDE [cm] ↓		MDispE [pix] ↓		1PA [%] ↑	
		Split 1	Split 3	Split 1	Split 3	Split 1	Split 3	Split 1	Split 3
2E+2F	EIS [22]	13.7	22.4	–	–	–	–	89.0	88.1
	SMC-Net [5]	11.2	14.5	4.5	6.3	0.37	0.52	94.3	92.0
	EGEI-Net [38]	9.7	11.1	3.9	4.7	0.35	0.38	95.1	93.4
	Ours	10.1	11.3	3.9	4.5	0.35	0.38	94.9	93.4
2F	PSMNet [3]	15.9	18.3	8.0	10.2	0.57	0.68	88.6	83.1
	GWC-Net [14]	15.0	17.4	7.5	9.3	0.53	0.64	89.9	85.8
	SStereo [29]	14.0	16.4	6.4	7.8	0.48	0.56	91.5	88.5
2E	DDES [27]	16.7	27.8	6.8	14.7	0.59	0.94	89.8	74.8
	EITNet [1]	14.2	19.4	5.9	10.4	0.55	0.75	92.1	89.6
	DTC-SPADE [36]	13.5	17.1	5.2	<u>7.9</u>	0.46	0.60	<u>93.0</u>	89.7
	StereoFlow-Net* [7]	<u>13.0</u>	<u>15.0</u>	<u>5.0</u>	**5.8**	<u>0.46</u>	**0.49**	92.9	**92.6**
1E+1F	HDES [41]	16.0	18.0	–	–	–	–	86.4	80.1
	DCNet [30]	16.9	17.8	7.0	8.6	0.66	0.75	85.0	83.3
	TDANet [35]	15.8	19.7	–	–	0.75	0.87	88.1	77.4
	Ours*	**12.1**	**14.2**	**4.8**	**5.8**	**0.41**	<u>0.54</u>	**94.1**	<u>91.4</u>

Table 1: Comparison of our method with SOTA disparity estimation methods on the MVSEC dataset [39], where methods marked with * utilize temporal information. The Methods are categorized by input modality configuration. **Bold** indicates the best, and <u>underlined</u> indicates the second-best.

4 Experiments

4.1 Experiment Setup

Dataset. We evaluate our method on the two established and appropriate benchmarks for asymmetric stereo vision: MVSEC [39] and DSEC [11], following established protocols to ensure a fair comparison with previous works. For the MVSEC dataset, we focus on the Indoor Flying scenes, which provide event streams from calibrated stereo event cameras with a resolution of 346×260. We follow [7, 27, 36], to evaluate on two different train-test splits. For the DSEC dataset[1], we adhere to the official training and testing splits [11]. DSEC consists of 53 outdoor driving scenes captured by a sensor suite including stereo event cameras with a resolution of 640×480, standard RGB cameras, and a LiDAR system.
Evaluation Metric. For the MVSEC dataset[39], we use the evaluation metrics: mean depth error (MDE), median depth error (MdDE), mean disparity error (MDispE), and 1-pixel accuracy (1PA) [7, 27, 36]. To assess performance on the DSEC dataset [11], we use root mean squared error (MSE) , mean absolute error (MAE), 1-pixel error (1PE), and 2-pixel error (2PE) [7, 23, 27].
Implementation Details. We implement the proposed method using the PyTorch framework and train it on RTX4090 GPUs. All models are trained for 100 epochs using the Adam optimizer with a weight decay 1e-4. We use a One-Cycle learning rate scheduler with a peak learning rate of 0.0004. A batch size of 8 is used for both the MVSEC and DSEC datasets. For all experiments, we perform 12 update iterations during both training and inference. To account for disparity variations, we set the maximum disparity to 96 for MVSEC and 192 for DSEC.
Comparison Methods. We explore state-of-the-art disparity estimation approaches with three modality settings: 1E+1F (one event and one frame camera) [30], 2E (two event cameras) [27], 2F(two frame cameras) [14] and 2E+2F (two event and two frame cameras) [22].

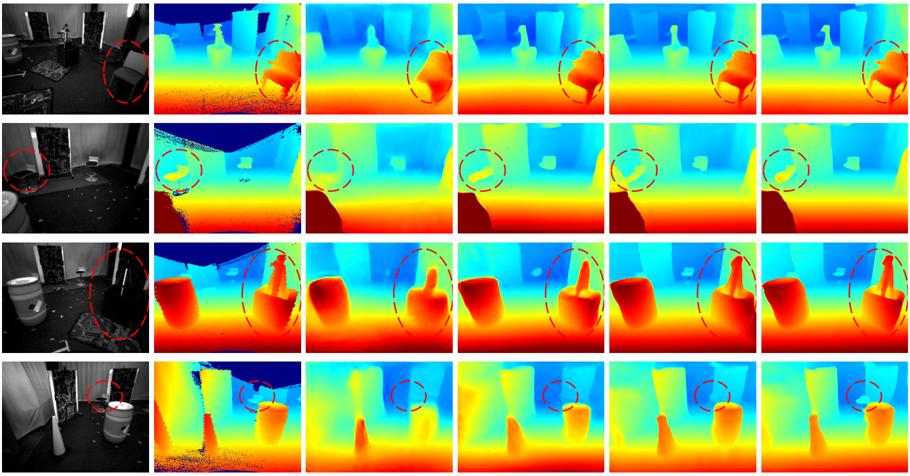
4.2 Results

Quantitative Results. Tab. 1 and Tab. 2 summarize the performance of our method and other state-of-the-art disparity estimation approaches on the MVSEC and DSEC datasets. On the two dataset, our method outperforms all existing '1E+1F' state-of-the-art methods, except for a marginal 0.005 decrement in 2-pixel error compared to AFES-dc [7] on the DSEC dataset. Furthermore, we surpass all '2E' and '2F' methods, and achieves performance comparable to the state-of-the-art '2E+2F' EGEI-Net [38] on the MVSEC dataset.
Qualitative Results. In Fig. 2 and Fig. 3, we qualitatively evaluate the proposed method against several state-of-the-art approaches on the MVSEC and DSEC datasets, respectively. Our method demonstrates best disparity estimation performance, particularly in preserving structural integrity and fine local details. For example, in Fig. 2, our method is the only one capable of capturing the fine-grained details of the small chair positioned behind the garbage bin.

4.3 Ablation Study

We conduct an ablation study on the MVSEC dataset [39] to assess the contribution of each component in our method.

[1] https://dsec.ifi.uzh.ch/

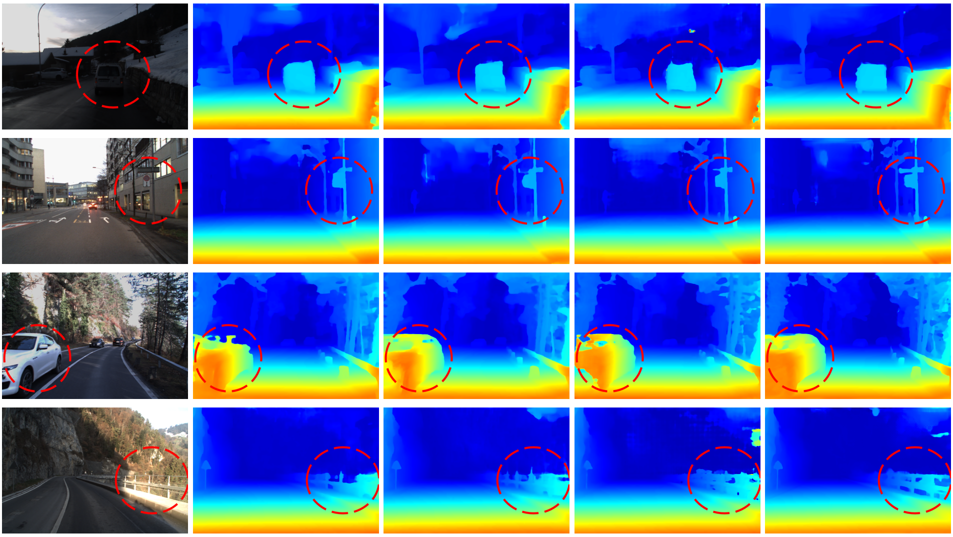


(a) Frame (b) Ground Truth (c) DCNet (d) DCT-SPADE (e) StereoFlow-Net (f) Ours

Figure 2: Qualitative comparison on MVSEC dataset. From left to right, we show the (a) frame, (b) ground truth disparity, (c) prediction of DCNet [30], (d) prediction of DCT-SPADE [36], (e) prediction of StereoFlow-Net [7] and (f) prediction of Ours.

Modality	Method	MAE ↓	1PE ↓	2PE ↓	RMSE ↓
2E+2F	EIS [22]	0.396	5.814	1.055	0.906
	SCSNet [6]	0.387	5.673	0.992	0.852
2E	DDES [27]	0.576	10.915	2.905	1.381
	EIS-E [22]	0.529	9.958	2.645	1.222
	DTC-SPADE [36]	0.527	9.277	2.416	1.291
	Se-cff [23]	0.519	9.583	2.620	1.231
	StereoFlow-Net* [7]	<u>0.493</u>	**8.662**	<u>2.259</u>	<u>1.172</u>
	MGC-Net [16]	0.514	9.625	2.512	1.187
1E+1F	HDES [41]	0.698	19.020	4.970	1.307
	DCNet [30]	0.587	13.050	2.850	1.193
	AFES-dc* [4]	0.543	10.970	**2.230**	1.175
	Ours*	**0.488**	<u>8.881</u>	2.335	**1.168**

Table 2: Comparison of our method with SOTA disparity estimation methods on the DSEC dataset [11], where methods marked with * utilize temporal information. The Methods are categorized by input modality configuration. **Bold** indicates the best, and <u>underined</u> indicates the second-best.



(a) Frame (b) Se-cff (c) MGC-Net (d) StereoFlow-Net (e) Ours

Figure 3: Qualitative comparison on DSEC dataset. From left to right, we show the (a) frame, (b) prediction of Se-cff [23], (c) prediction of MGC-Net [16], (d) prediction of StereoFlow-Net [7], and (e) prediction of Ours.

Effectiveness of the Proposed Modules. We evaluate the impact of the event-aligned frame encoder and the structure-detail refinement module in Tab. 4. Our RAFT-Stereo baseline [20] for asymmetric stereo matching utilizes two independent feature extractors with the same architecture to extract features from frame and event data separately. The results show that the full model, which incorporates all these features, achieves the best overall performance. Moreover, our ablation studies against the baseline further underscore the crucial role of temporal cues from consecutive frames.

Generalization to Existing Frameworks. We evaluate the effectiveness of our proposed event-aligned frame encoder in Tab. 5 and the structure-detail refinement module in Tab. 6, by integrating them into existing stereo matching frameworks. In Tab. 5, we fix the refinement strategy and explore different cross-modal alignment schemes, including E2VID [24], DCNet [30], and ZiE-Stereo [21]. In Tab. 6, we fix the alignment module and instead explore different refinement strategies, including GWC-Net [14], SStereo [29], and EGEI-Net [38]. In addition, Tab. 7 compares different feature extraction modules, including LinAttn [17], DefAttn [31], and our ResNet-based encoder, while Tab. 8 shows the generalization performance on the M3ED [2] dataset, comparing HSM [18], TDANet [35], and our proposed method. In all cases, the inclusion of our modules leads to consistent performance improvements, demonstrating the generalizability and effectiveness of our method across different architectures.

Discussion. To show the performance of our method in capturing structural and fine-grained local details, we employ the Sobel operator to segment the frame for edge detection. We then compare our method with other '1E+1F' and '2E' approaches in terms of structural accuracy and local detail preservation in Table 3. Our method significantly outperforms DCNet [30], and has a 3.16% improvement compared to StereoFlow-Net [7].

Method	Edge Region				Non-Edge Region			
	MDE ↓	MdDE ↓	MDispE ↓	1PA ↑	MDE ↓	MdDE ↓	MDispE ↓	1PA ↑
DCNet [30]	19.7	7.5	0.60	86.3	17.2	7.8	0.63	87.7
DTC-SPADE [36]	17.7	6.7	0.56	90.1	14.5	6.8	0.56	91.2
StereoFlow-Net [7]	<u>16.6</u>	<u>6.0</u>	<u>0.50</u>	**91.4**	<u>13.1</u>	<u>5.6</u>	<u>0.48</u>	**92.9**
Ours	**15.9**	**5.5**	**0.49**	<u>91.3</u>	**12.6**	**5.2**	**0.47**	<u>92.5</u>

Table 3: Performance comparison on the MVSEC dataset [39]. For each sample, we divide it into edge and non-edge regions by using the Sobel filter. **Bold** denotes the best, and <u>underlined</u> the second-best performance.

Baseline	$\mathbf{Z}_t^{\text{I-L}}$	$\mathbf{Z}_t^{\text{I-S}}$	$\mathbf{Z}_t^{\mathbf{I}}$	$\mathbf{S}_l$	$\mathbf{G}_l$	MDE ↓	MdDE ↓	MDispE ↓	1PA ↑
✓						16.5	8.0	0.61	87.4
✓	✓					15.7	7.3	0.60	88.5
✓	✓	✓				14.4	6.2	0.56	89.7
✓	✓	✓	✓			14.0	5.9	0.53	90.5
✓	✓	✓	✓	✓		13.7	5.7	0.51	91.3
✓	✓	✓	✓	✓	✓	**13.4**	**5.4**	**0.49**	**92.5**

Table 4: Ablation studies of the proposed modules.

Method	MDE ↓	MdDE ↓	MDispE ↓	1PA ↑
E2VID [24]	18.2	9.5	0.81	81.7
DCNet [30]	16.0	8.0	0.71	85.7
ZiE-Stereo [21]	15.4	6.8	0.63	88.1
Ours	**14.0**	**5.9**	**0.53**	**90.1**

Table 5: Ablation studies of the event-intensity alignment module.

Method	MDE ↓	MdDE ↓	MDispE ↓	1PA ↑
GWC-Net [14]	14.7	6.1	0.57	89.6
SStereo [29]	14.1	5.5	0.53	90.2
EGEI-Net [38]	14.5	5.8	0.56	90.8
Ours	**13.4**	**5.4**	**0.49**	**92.5**

Table 6: Ablation studies of the refinement modules.

Method	MDE ↓	MdDE ↓	MDispE ↓	1PA ↑
LinAttn [17]	18.0	8.5	0.67	86.9
DefAttn [31]	16.5	8.4	0.63	87.1
Ours	**16.5**	**8.0**	**0.61**	**87.4**

Table 7: Ablation studies of the feature extraction modules.

Method	MAE ↓	RMSE ↓	2PE ↓	3PE ↓
HSM [18]	12.39	**14.27**	92.58	90.87
TDANet [35]	20.07	22.19	95.47	93.12
Ours	**9.24**	15.94	**78.74**	**66.32**

Table 8: Ablation studies of the generalization performance on the M3ED dataset [2].

5 Conclusions

In this paper, we study disparity estimation from asymmetric stereo by exploiting the complementarity of event and frame data. Recognizing the inherent domain shift between these two modalities, first, we enhance the temporal awareness of frame features by incorporating temporal cues extracted from consecutive frames, making them more compatible with the temporally rich event data. Second, noting the spatial sparsity of event data often leads to bad disparity predictions in fine-grained details, we introduce an iterative refinement scheme that alternates between estimating structural information and fine-grained local details. Extensive experiments on the MVSEC and DSEC datasets demonstrate that our method significantly outperforms state-of-the-art approaches.

Acknowledgment. This work was supported by the National Natural Science Foundation of China under grant No. 62302045, and the Beijing Institute of Technology Special-Zone.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 882–890, 2021.

[2] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4023, 2023.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018.

[4] Xihao Chen, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Depth from asymmetric frame-event stereo: A divide-and-conquer approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3045–3054, 2024.

[5] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 454–462, 2022.

[6] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022.

[7] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pages 294–314. Springer, 2024.

[8] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2768–2776, 2020.

[9] 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.

[10] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018.

[11] 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. doi: 10.1109/LRA.2021.3068942.

[12] Suman Ghosh and Guillermo Gallego. Event-based stereo depth estimation: A survey. *arXiv preprint arXiv:2409.17680*, 2024.

[13] 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 preprint arXiv:2211.00509*, 2022.

[14] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Junjie Jiang, Hao Zhuang, Xinjie Huang, Delei Kong, and Zheng Fang. Event-based stereo depth estimation with motion guidance and left-right consistency. *IEEE Transactions on Instrumentation and Measurement*, pages 1–1, 2025. doi: 10.1109/TIM.2025.3566813.

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[18] 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.

[19] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021.

[20] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[21] Hanyue Lou, Jinxiu Liang, Minggui Teng, Bin Fan, Yong Xu, and Boxin Shi. Zero-shot event-intensity asymmetric stereo via visual prompting from image domain. *Advances in Neural Information Processing Systems*, 37:13274–13301, 2024.

[22] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021.

[23] Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6114–6123, 2022.

[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.

[25] Richard Shaw, Sibi Catley-Chandar, Ales Leonardis, and Eduardo Pérez Pellitero. Hdr reconstruction from bracketed exposures and events. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0603.pdf.

[26] Yi Tian and Juan Andrade-Cetto. Event transformer flownet for optical flow estimation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0577.pdf.

[27] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024.

[30] Ziwei Wang, Liyuan Pan, Yonhon Ng, Zheyu Zhuang, and Robert Mahony. Stereo hybrid event-frame (shef) cameras for 3d perception. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9758–9764. IEEE, 2021.

[31] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.

[32] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21919–21928, 2023.

[33] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1959–1968, 2020.

[34] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24902–24912, 2024.

[35] Dehao Zhang, Qiankun Ding, Peiqi Duan, Chu Zhou, and Boxin Shi. Data association between event streams and intensity frames under diverse baselines. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022.

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8676–8686, 2022.

[37] Ruikun Zhang, Hao Yang, Yan Yang, Ying Fu, and Liyuan Pan. Lmhaze: intensity-aware image dehazing with a large-scale multi-intensity real haze dataset. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1, 2024.

[38] Fengan Zhao, Qianang Zhou, and Junlin Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 190–205, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73464-9.

[39] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. doi: 10.1109/LRA.2018.2800793.

[40] Hao Zhuang, Zheng Fang, Xinjie Huang, Kuanxu Hou, Delei Kong, and Chenming Hu. Ev-mgrflownet: Motion-guided recurrent network for unsupervised event-based optical flow with hybrid motion-compensation loss. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.

[41] 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 *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6833–6840. IEEE, 2021.