

EV-CGNet: Co-visible Focused 3D-guided 2D Event Keypoint Detection Network

Yuan Gao¹, Tianle Ding¹, Yuqing Zhu¹, Tianzhu Zhang^{1,2*}

¹ University of Science and Technology of China

² National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory

{wazs98, dtl98, yuqingzhu}@mail.ustc.edu.cn, {tzzhang}@ustc.edu.cn

Abstract

Event keypoint detection has garnered significant attention due to its ability to establish spatial associations through matching, which is fundamental for various computer vision tasks. However, achieving robust event keypoint detection remains challenging due to the difficulty in balancing the exploitation of event data and compatibility with established algorithms. Moreover, the limited use of co-visible information often results in excessive keypoint detection in non-matching regions, leading to incorrect matches. To address these challenges, we propose a novel Co-visible Focused 3D-guided 2D Event Keypoint Detection Network (EV-CGNet), which mainly consists of a 3D-guided 2D feature prototype learning (G2PL) module and a co-visible region-focused detector and descriptor learning (CDDL) module. The proposed method enjoys several merits. First, the proposed G2PL module can leverage fine-grained spatio-temporal cues from event points to guide the learning of event frame feature prototypes. Second, the proposed CDDL module can focus keypoint detection on co-visible regions, ensuring accurate matches. Comprehensive experimental evaluations on six challenging benchmarks show that our method significantly outperforms state-of-the-art event keypoint detection methods.

1. Introduction

Event cameras asynchronously capture pixel-level brightness changes and produce event streams $\{\mathbf{e}_i\}_{i \in \mathbb{N}}$, where each event $\mathbf{e}_i = (x_i, y_i, t_i, p_i)$ encodes pixel position (x_i, y_i) , timestamp t_i and brightness change polarity p_i . Compared to traditional frame-based cameras, event cameras excel with large dynamic range, low bandwidth, high temporal resolution, and minimal power usage [1, 14, 31, 34, 39]. These properties enable diverse applications, particularly in high-speed motion and extreme lighting conditions, such as 3D reconstruction [6, 9, 22, 45], depth es-

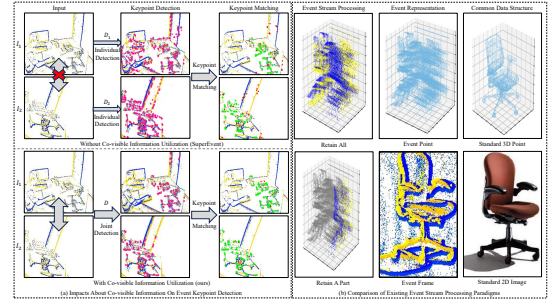


Figure 1. Illustration of our motivation. (a) presents keypoint detection and matching results of our method and SuperEvent [4]. Without co-visible information, SuperEvent detects keypoints in non-matching regions, leading to numerous incorrect matches (red stars). While our method restricts keypoints to co-visible regions, yielding correct matches (green stars). (b) compares the point-based and frame-based paradigms in event loss (gray points) and their correlation with the corresponding traditional data structures.

timization [36, 40, 41], and Simultaneous Localization and Mapping (SLAM) [5, 21, 26, 29]. Among these applications, event keypoint detection [1, 4, 18, 20, 27, 31, 32, 46] has received increasing attention due to its ability to establish spatial associations via matching, which is fundamental for various computer vision tasks. However, achieving robust event keypoint detection remains challenging, primarily due to the complex data structure, the difficulty in fully exploiting fine-grained spatio-temporal information inherent in event data, and the need for reliable matching under significant scene appearance variations.

To address these issues, event keypoint detection methods typically convert event data into common data structures, mainly categorized into point-based methods [8, 25] and frame-based methods [1, 4, 18, 20, 27, 32, 33, 43, 46]. Point-based methods convert event streams into spatio-temporal event points in (X, Y, T) space and then adapt standard 3D (X, Y, Z) point cloud algorithms to detect keypoints. For example, the state-of-the-art point-based method eCDT [25] adapts DBSCAN [12] to build clus-

*Corresponding Author

ters for event points and defines the centroid of each cluster as a keypoint. While these methods fully retain fine-grained spatio-temporal information, the inherent disparity between the spatio-temporal domain (X, Y, T) and the spatial domain (X, Y, Z) limits the efficacy of standard 3D algorithms. Differently, frame-based methods transform event streams into event frames and modify traditional image keypoint detectors [10, 17, 23, 42]. State-of-the-art methods [18, 27] typically segment event streams into fixed-duration windows and create an event frame for each time window with the latest event per pixel. While frame-based methods enhance compatibility with conventional algorithms and consequently achieve higher accuracy than point-based methods, significant event loss limits the effective utilization of fine-grained spatio-temporal cues, thereby limiting robustness to scene appearance changes. To address this issue, recent methods [4, 46] propose fusing multiple event frames within each window to retain more events. However, computational cost limits the number of event frames that can be processed, and discrete sampling can introduce inaccuracies. For example, if a pixel’s intensity fluctuates downward while all sampled events are positive, the model may incorrectly infer a consistent increase in pixel intensity. Besides, keypoint detectors are trained and evaluated independently for each fused event frame, leading to fixed keypoints that lack adaptivity to co-visible information across different event frames during matching. This often results in the detection of numerous keypoints in non-matching areas, ultimately reducing matching accuracy, as shown in Figure 1(a). In summary, two notable limitations remain, including the inadequate utilization of co-visible information, and limited balance between event exploitation and compatibility with established algorithms.

Based on the above discussions, we identify both the utilization of co-visible information and the balance between event exploitation and compatibility with established algorithms as critical for event keypoint detection. Specifically, the following two issues deserve careful attention. (1) **How to leverage co-visible information to enhance keypoint detection for superior matching performance.** State-of-the-art methods learn keypoint detectors independently for individual event frames. Therefore, the lack of inter-frame interaction limits the ability to capture co-visible information. This motivates the use of attention mechanisms to model inter-frame interactions, enabling the capture of co-visible regions and thereby enhancing keypoint detection and matching accuracy. (2) **How to balance compatibility with established algorithms and exploiting fine-grained spatio-temporal cues.** Point-based methods excel in preserving the fine-grained spatio-temporal information inherent in abundant events, yet they struggle with integration into conventional 3D algorithms. In contrast, frame-based methods are more compatible with traditional 2D al-

gorithms, offering higher accuracy, but they suffer from significant loss of events, as shown in Figure 1(b). Notably, these two paradigms exhibit complementary potential: frame-based methods ensure accuracy, while point-based methods enhance them by incorporating fine-grained spatio-temporal cues. Therefore, it is necessary to propose a 3D-guided 2D paradigm that captures the fine-grained spatio-temporal cues in event data while maintaining compatibility with traditional methods, thereby enabling more robust event keypoint detection.

Motivated by the above observations, we propose a novel Co-visible Focused 3D-guided 2D Event Keypoint Detection Network (EV-CGNet), which mainly consists of a 3D-Guided 2D Feature Prototype Learning (G2PL) module and a Co-visible Region-focused Detector and Descriptor Learning (CDDL) module. In the 3D-guided 2D feature prototype learning module, it is proposed to utilize fine-grained spatio-temporal cues encoded in event points to guide the learning of event frame feature prototypes. Specifically, original event point and frame features are extracted by a 1×1 convolutional layer. Event point prototypes then interact with event point features via attention mechanisms, yielding event point agents that effectively capture fine-grained spatio-temporal cues. These agents then interact with frame features to form the 3D-guided 2D feature prototypes. In the co-visible region-focused detector and descriptor learning module, it is proposed to focus keypoint detection on co-visible regions between event frames for robust keypoint matching against significant variations in scene appearance. Specifically, for each event frame pair, 3D-guided 2D feature prototypes extracted from the individual event frames are concatenated into D2 agents and refined through self-attention mechanisms to capture co-visible information. The updated D2 agents serve as co-visible region-focused detectors, which can focus on distinctive keypoint patterns in co-visible regions. Finally, co-visible region-focused descriptors are obtained by fusing keypoint detectors into original event frame features.

The main contributions of this work can be summarized as follows. (1) We propose a novel event keypoint detection method by integrating a distinctive 3D-guided 2D paradigm and an innovative strategy focusing keypoint detection on co-visible regions. Our model can realize robust event keypoint detection against various challenging factors. (2) The proposed G2PL module can leverage fine-grained spatio-temporal cues from event points to guide event frame feature prototype learning, enhancing robustness to scene appearance changes. And the CDDL module can concentrate keypoint detection on co-visible regions and ensure accurate matches. (3) Extensive experiments on six challenging benchmarks show that our method outperforms state-of-the-art event keypoint detection methods by a large margin.

2. Related Work

In this section, we briefly overview point-based and frame-based methods for event keypoint detection.

Point-based Event Keypoint Detection. Point-based methods [8, 25] convert event streams into spatio-temporal event points in (X, Y, T) coordinate space, and then adapt algorithms initially designed for traditional 3D point clouds in (X, Y, Z) space to extract keypoints. An early approach [8] employs plane fitting on spatio-temporal event points to identify the intersections of planes as keypoints. More recently, eCDT [25] leverages DBSCAN [12] clustering to locate cluster centers as keypoints. However, the significant domain gap between spatio-temporal event points in (X, Y, T) space and traditional 3D point clouds in (X, Y, Z) space limits the effectiveness of standard 3D algorithms on event data. Therefore, recent research has shifted to frame-based methods. In contrast, we propose a novel 3D-guided 2D paradigm that retains the advantages of point-based methods and leverages fine-grained spatio-temporal cues to enhance event frame features.

Frame-based Event Keypoint Detection. Frame-based methods [1, 18, 27, 32, 33, 43] transform event streams into image-like event frames and adapt image keypoint detectors [10, 17, 23, 42] for keypoint detection. These methods can be classified into two main categories, including event-by-event methods [1, 32, 33, 43] and sliding-window methods [4, 18, 27, 46]. Event-by-event methods maintain a continuously updated event frame to identify keypoints from each incoming event. For instance, ev-Fast [33] extends the FAST [42] detector to identify keypoints on Time Surface [2]. Specifically, each incoming event updates the corresponding pixel with its timestamp, and the pixel is identified as a keypoint if a specified number of consecutive pixels within its circular neighborhood exhibit higher values than others. Nevertheless, these methods exhibit limited accuracy due to the incompatibility of the event-by-event paradigm with state-of-the-art image keypoint detection methods [10, 11, 17], which jointly optimize descriptors and detectors in a data-driven manner. To overcome these limitations, sliding-window methods [4, 18, 27, 46] segment event streams into fixed-duration windows and construct an event frame within each window for joint learning of descriptors and detectors. For instance, SD2Event [18] leverages the polarity and timestamp of the most recent event at each pixel to form an event frame within each window. Attention mechanisms are then applied to optimize both detectors and descriptors for individual event frames. However, excessive event loss during frame construction prevents it from fully exploiting fine-grained spatio-temporal cues, thereby limiting accuracy. To address this issue, recent methods [4, 46] propose the fusion of multiple event frames within each window to retain more events. Despite these advances, the computa-

tional cost constrains the number of event frames that can be processed, and discrete sampling introduces inaccuracies. Additionally, keypoint detectors are trained and evaluated independently for each fused event frame, leading to fixed keypoints that lack the ability to adaptively adjust based on co-visible information across different event frames during matching. This often leads to numerous keypoints extracted in non-matching areas, ultimately reducing matching accuracy. In contrast, our proposed 3D-guided 2D feature prototype learning module can fully use fine-grained spatio-temporal cues encoded in event points to guide the learning of event frame feature prototypes. Besides, our co-visible region-focused detector and descriptor learning module utilizes inter-frame interactions to capture co-visible regions, thereby ensuring robust keypoint matching even under significant variations in scene appearance.

3. Our Approach

In this section, we present our proposed method by integrating a distinctive 3D-guided 2D paradigm and an innovative strategy focusing keypoint detection on co-visible regions. The overall architecture is presented in Figure 2.

3.1. Overview

As shown in Figure 2, our proposed model contains a 3D-guided 2D feature prototype learning (G2PL) module and a co-visible region-focused detector and descriptor learning (CDDL) module. Two input event streams are first encoded into point and frame representations. In the G2PL module, a 1×1 convolution extracts original event point and frame features. Event point prototypes interact with point features to capture fine-grained spatio-temporal cues, and the resulting event point agents further interact with frame features to produce 3D-guided 2D prototypes. In the CDDL module, prototype pairs are concatenated into D2 (Detector and Descriptor) agents, with interactions modeled via self-attention. The resulting co-visible region-focused detectors are then fused into frame features to obtain co-visible region-focused descriptor pairs. Finally, a cosine similarity constraint on the heatmap pair supervises learning.

3.2. Event Encoder

Event streams $\{(x_i, y_i, t_i, p_i)\}$ are first divided into fixed-duration time windows. For each window $(T, T + \Delta t)$, we generate both an event frame and event points. Specifically, the frame $\mathbf{I}_E \in \mathbb{R}^{3 \times h \times w}$ is created following [18],

$$\mathbf{I}_E[x_i, y_i] = (255, t_{iT}, 0) \leftarrow (x_i, y_i, t, +1), \quad (1)$$

$$\mathbf{I}_E[x_i, y_i] = (0, 255 - t_{iT}, 255) \leftarrow (x_i, y_i, t, -1), \quad (2)$$

where t is the timestamp of the latest event occurred at pixel (x_i, y_i) , and $t_{iT} = 127 \times (T + \Delta t - t)/\Delta t$. Event points are created by ignoring polarities p_j , which encodes pixel

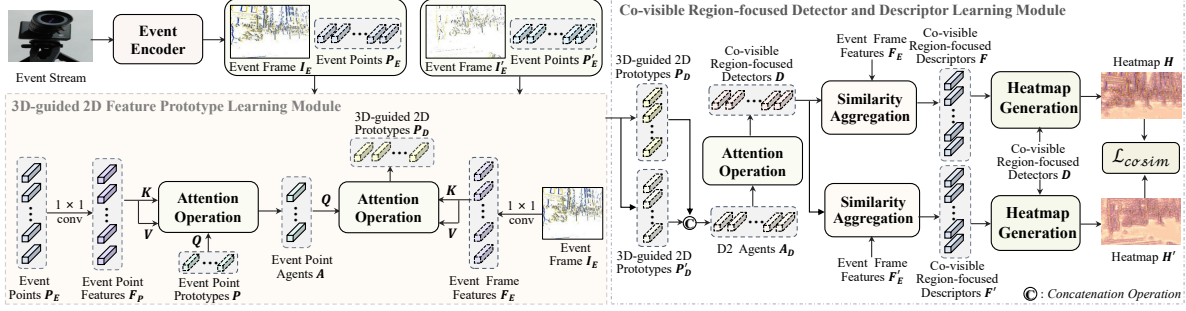


Figure 2. The architecture of our model consists of two primary components, including a 3D-guided 2D Feature Prototype Learning (G2PL) module and a Co-visible Region-focused Detector and Descriptor Learning (CDDL) module. Event streams are encoded into frames and points. In the G2PL module, 1×1 convolution generates event point features $\mathbf{F}_P$ and frame features $\mathbf{F}_E$. Event point prototypes $\mathbf{P}$ interact with $\mathbf{F}_P$ via attention to produce event point agents $\mathbf{A}$, which then interact with $\mathbf{F}_E$ to obtain 3D-guided 2D prototypes $\mathbf{P}_D$. The process is repeated to produce $\mathbf{P}'_D$. In the CDDL module, $\mathbf{P}_D$ and $\mathbf{P}'_D$ are concatenated into D2 agents $\mathbf{A}_D$, with self-attention modeling their interactions. The resulting co-visible region-focused detectors $\mathbf{D}$ are then fused with original frame features to obtain descriptors $\mathbf{F}$ and $\mathbf{F}'$. Finally, a cosine similarity constraint heatmap pairs $\mathbf{H}$ and $\mathbf{H}'$ guides model learning. For more details, please refer to the text.

luminance changes instead of spatio-temporal information, followed by normalization,

$$\mathbf{P}_{Ej} = \left(\frac{x_j}{w}, \frac{y_j}{h}, \frac{t_j - T}{\Delta t} \right). \quad (3)$$

3.3. 3D-Guided 2D Feature Prototype Learning

After obtaining event points and frames, we leverage fine-grained spatio-temporal information from the event points to guide the learning of event frame feature prototypes. First, a 1×1 convolutional layer is applied to extract original event point features $\mathbf{F}_P \in \mathbb{R}^{d \times M}$ and event frame features $\mathbf{F}_E \in \mathbb{R}^{d \times hw}$. We then design N event point prototypes $\mathbf{P} \in \mathbb{R}^{d \times N}$ to interact with $\mathbf{F}_P$ via attention mechanisms. Specifically, keys and values arise from $\mathbf{F}_P$, and queries arise from $\mathbf{P}$. Formally,

$$\mathbf{Q} = \mathbf{W}^Q \mathbf{P}, \mathbf{K} = \mathbf{W}^K \mathbf{F}_P, \mathbf{V} = \mathbf{W}^V \mathbf{F}_P, \quad (4)$$

where $\mathbf{W}^Q \in \mathbb{R}^{d_k \times d}$, $\mathbf{W}^K \in \mathbb{R}^{d_k \times d}$, $\mathbf{W}^V \in \mathbb{R}^{d \times d}$ are linear projections. Then, event point agents $\mathbf{A} \in \mathbb{R}^{d \times N}$ are obtained via multi-head attention mechanisms [44],

$$\mathbf{A} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \text{Softmax}(\mathbf{K}^\top \mathbf{Q}). \quad (5)$$

In this way, event point agents $\mathbf{A}$ can effectively capture fine-grained spatio-temporal cues from event points. To guide the learning of event frame feature prototypes $\mathbf{P}_D \in \mathbb{R}^{d \times N}$, $\mathbf{A}$ interact with frame features $\mathbf{F}_E$ as follows,

$$\mathbf{Q} = \mathbf{W}^Q \mathbf{A}, \mathbf{K} = \mathbf{W}^K \mathbf{F}_E, \mathbf{V} = \mathbf{W}^V \mathbf{F}_E, \quad (6)$$

$$\mathbf{P}_D = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \text{Softmax}(\mathbf{K}^\top \mathbf{Q}). \quad (7)$$

Similarly, $\mathbf{P}'_D \in \mathbb{R}^{d \times N}$ is generated from $\mathbf{I}'_E$ and $\mathbf{P}'_E$. This adaptive interaction mitigates the limitations of the point-based paradigm while effectively exploiting spatio-temporal information to enhance frame feature prototypes.

3.4. Co-visible Region-focused Detector and Descriptor Learning

To focus keypoint detection on co-visible regions between event frames, we use attention mechanisms to model inter-frame interactions. After obtaining the 3D-guided 2D prototype pair $\mathbf{P}_D$ and $\mathbf{P}'_D$, we concatenate them to form D2 agents $\mathbf{A}_D \in \mathbb{R}^{d \times 2N}$. Then, self-attention mechanisms are used to model interactions among multiple D2 agents,

$$\mathbf{Q} = \mathbf{W}^Q \mathbf{A}_D, \mathbf{K} = \mathbf{W}^K \mathbf{A}_D, \mathbf{V} = \mathbf{W}^V \mathbf{A}_D, \quad (8)$$

$$\mathbf{D} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \text{Softmax}(\mathbf{K}^\top \mathbf{Q}). \quad (9)$$

By modeling long-range dependencies within and across event frames, the co-visible region-focused detectors $\mathbf{D} \in \mathbb{R}^{d \times 2N}$ encode co-visible information effectively. The original features $\mathbf{F}_E$ are updated as $\mathbf{F} = \mathbf{F}_E + \mathbf{D}(\mathbf{D}^\top \mathbf{F}_E)$, producing co-visible region-focused adaptive descriptors $\mathbf{F} \in \mathbb{R}^{d \times hw}$. Similarly, $\mathbf{F}' \in \mathbb{R}^{d \times hw}$ is obtained. Multiple score maps $\mathbf{S}_N, \mathbf{S}'_N \in \mathbb{R}^{2N \times hw}$ are computed via dot products between detectors and descriptors, then averaged and reshaped to yield the final heatmaps $\mathbf{H}, \mathbf{H}' \in \mathbb{R}^{h \times w}$.

3.5. Objective Function

To guide model learning effectively, we follow [18] to design supervisory signals. Specifically, for each time window, we generate a random homography as the reference ground truth and warp the event frame $\mathbf{I}_E$ to obtain $\mathbf{I}'_E$. For the event points $\mathbf{P}_E$, the same transformation is applied to the spatial coordinates, while Gaussian noise is added to the temporal dimension, producing the transformed points $\mathbf{P}'_E$. Then, event frame and point groups $(\mathbf{I}_E, \mathbf{P}_E)$ and $(\mathbf{I}'_E, \mathbf{P}'_E)$ are input to our proposed model to generate the heatmap pair $\mathbf{H}$ and $\mathbf{H}'$. We apply the cosine similarity loss $\mathcal{L}_{cosim}$ and peaky loss $\mathcal{L}_{peaky}$ from [18] to the heatmap pair, and

the diversity loss $\mathcal{L}_{div}$ to the event point prototypes. The total loss is thus defined as

$$\mathcal{L}_{total} = \mathcal{L}_{cosim} + \alpha_1 \mathcal{L}_{peaky} + \alpha_2 \mathcal{L}_{div}, \quad (10)$$

where α_1 and α_2 are weight terms to balance these losses.

4. Experiments

In this section, we first outline the implementation details. Then, we present quantitative and qualitative experimental results on six public benchmarks. To further prove the effectiveness of individual components, we conduct extensive ablation studies. Please refer to the **Supplementary Material** for additional discussions and visualization results.

4.1. Implementation Details

We implement our proposed model in Pytorch [37]. Two 1×1 convolutional layers with unshared parameters extract event point and frame features, with a feature dimension of $d = 128$. In the G2PL module, the number of event point prototypes N is set to 16, and cross-attention uses 8 heads with query and key dimensions defined in Eq. (4), Eq. (6) and Eq. (8) $d_k = d$. Keypoints are obtained from the heatmap $\mathbf{H}$ via local maxima filtering and thresholding, followed by mutual nearest neighbor matching. When matching is unnecessary, the two input event streams are set identical for keypoint extraction. For the objective function, we set weight terms α_1 and α_2 to 0.6 and 0.8, respectively. Our model runs at 6 ms per 240×180 event frame pair on a single RTX 3090 GPU. For training, we adopt the same training dataset [19, 34] as [18], where the event streams are divided into multiple time windows with a 50ms interval. All parameters are randomly initialized and trained from scratch using the Adam optimizer, with a learning rate of 10^{-4} and weight decay of 3×10^{-4} . The model converges after 18 hours of training on a single RTX 3090 GPU.

4.2. Datasets and Evaluation Metrics

Event-Camera. The Event-Camera dataset [34] uses a DAVIS240C [3] sensor to capture both real and simulated scenarios, with a peak event rate of up to 8×10^6 events/s. The dataset presents significant challenges including varying motion patterns and extreme viewpoint changes. For a fair comparison, we adopt the same subsets as [1, 4, 18, 32, 46] to evaluate our method. And the *evaluation metric* follows [4, 18, 32, 46], where we report the reprojection error and area under the cumulative curve (AUC) of pose estimation. To further assess the performance of the keypoint detector and descriptor, we adopt the procedure of [10] to evaluate keypoint repeatability, localization error, descriptor discriminability, and matching score.

N-Caltech101. The N-Caltech101 dataset [35] is recorded using an ATIS [38] sensor, capturing saccadic motion in

Table 1. Evaluation results on the Event-Camera dataset. We report the reprojection error in pixels.

Methods	Reprojection Error
evHarris [43]	2.46
SILC [32]	2.16
SD2Event [18]	1.64
STPNet [46]	1.43
SuperEvent [4]	1.22
EV-CGNet (ours)	0.91

Table 2. Evaluation results for pose estimation on the Event Camera dataset. We report the AUC of the pose error at thresholds ($5^\circ, 10^\circ, 20^\circ$).

Methods	AUC@ 5°	AUC@ 10°	AUC@ 20°
SD2Event [18]	16.6	25.3	36.5
STPNet [46]	19.1	31.3	42.9
SuperEvent [4]	22.7	35.8	46.7
EV-CGNet (ours)	31.3	45.7	60.1

front of a screen displaying the Caltech101 dataset [13]. This dataset presents challenges due to significant changes in scale and viewpoint. We follow the same procedure as [18, 46] to evaluate our method. For the *evaluation metric*, we adopt the intersection over union (IoU) matching score, consistent with [18, 39, 46].

HVGA ATIS Corner. HVGA ATIS Corner [32] consists of seven sequences that capture diverse scenes characterized by distinct planar patterns, recorded using an ATIS sensor. The main challenges in this dataset are substantial texture and viewpoint variations. We adopt the same subset as [7, 18, 27, 32, 46] for evaluation. As for the *evaluation metric*, we follow [18, 32, 46] and report the reprojection error.

EDS. EDS [24] utilizes a hand-held DAVIS240C sensor to capture various indoor scenes, which presents challenges due to substantial viewpoint variations and diverse motion patterns. We adopt the evaluation procedure from [4] and use the AUC of pose estimation as the *evaluation metric*.

VECTor. VECTor [16] uses a Prophesee Gen3 stereo camera to capture indoor scenes with an event rate up to 1.3×10^7 events/s. This dataset poses challenges such as extreme texture variations and long trajectories. We follow the same procedure as [4] to evaluate our method and report the Absolute Trajectory Error (ATE) as the *evaluation metric*.

TUM-VIE. TUM-VIE [28] captures indoor and outdoor scenes using a Prophesee Gen4 stereo camera, with an event rate up to 3×10^7 events/s. The primary challenges include diverse motion patterns and long-distance loops. We follow the same procedure as [4] to evaluate our method. As for the *evaluation metric*, we follow [4] and report the ATE.

Table 3. Evaluation results on the Event Camera dataset. We report the Repeatability (Rep.), Mean Localization Error (MLE), Nearest Neighbor mean Average Precision (NN mAP) and Matching Score (M. Score).

Methods	Rep.	MLE	NN mAP	M. Score
SD2Event [18]	0.597	1.413	0.697	0.396
STPNet [46]	0.675	1.215	0.728	0.407
SuperEvent [4]	0.536	1.012	0.745	0.427
EV-CGNet (ours)	0.852	0.636	0.907	0.793

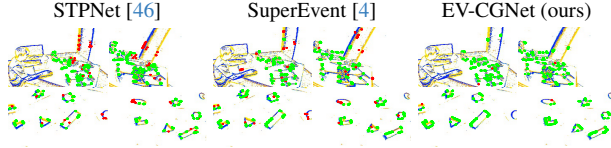


Figure 3. Qualitative comparisons on the Event-Camera dataset. Green and red stars denote correct and incorrect matches, respectively (threshold is 3px).

4.3. Comparison with State-of-the-art Methods

Results on Event-Camera dataset. We compare our method with existing state-of-the-art event keypoint detection methods [4, 18, 27, 32, 43, 46]. As shown in Table 1, our method significantly outperforms all other methods. Specifically, compared to SuperEvent [4], our method reduces the reprojection error by 0.31 pixels. Besides, we conduct a comprehensive evaluation of the keypoint detector and descriptor. As shown in Table 3, our proposed method outperforms SuperEvent [4] by 31.6% in keypoint repeatability (Rep.), 0.384 pixels in keypoint localization error (MLE), 16.2% in descriptor discriminability (NN mAP) and 36.6% in matching score (M. Score). The results show that our matching-oriented design improves both keypoint repeatability and descriptor discriminability while ensuring accurate matching. We further present qualitative comparisons with the state-of-the-art methods [4, 46]. As shown in Figure 3, these methods produce numerous incorrect keypoint matches in non-matching regions under extreme viewpoint variations. In contrast, our proposed method concentrates the keypoint detection on co-visible regions, achieving superior matching accuracy. Finally, we evaluate our method on a downstream pose estimation task. As shown in Table 2, our proposed method outperforms other state-of-the-art methods significantly at all three thresholds. Compared to SuperEvent [4], our model improves by 8.6% in AUC@5°, 9.9% in AUC@10° and 13.4% in AUC@20°. The main reason is that our G2PL module can exploit fine-grained spatio-temporal cues in event points to guide event frame feature prototypes in capturing diverse motion patterns across frames. Moreover, the CDDL module can improve keypoint matching by focusing

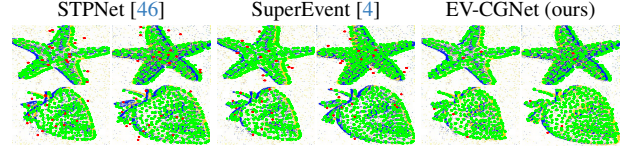


Figure 4. Qualitative comparisons on the N-Caltech101 dataset. Green and red stars denote correct and incorrect matches, respectively (threshold is 3px).

Table 4. Evaluation results on the N-Caltech101 dataset. We report the feature matching IoU in percentage.

Methods	IoU
Eventpoint [27]	0.83
SD2Event [18]	0.89
STPNet [46]	0.93
SuperEvent [4]	0.95
EV-CGNet (ours)	0.98

Table 5. Evaluation results on the HVGA ATIS Corner dataset. We report the reprojection error in pixels across event frame pairs at time intervals of 25ms, 50ms, and 100ms.

Methods	25ms	50ms	100ms
evHarris [43]	2.57	3.46	4.58
SILC [32]	2.45	3.02	3.68
Eventpoint [27]	1.27	1.41	1.72
SD2Event [18]	0.67	0.76	0.93
STPNet [46]	0.54	0.61	0.76
SuperEvent [4]	0.43	0.49	0.63
EV-CGNet (ours)	0.27	0.31	0.42

detection on co-visible regions, thereby facilitating pose estimation, especially under extreme viewpoint variations.

Results on N-Caltech101 dataset. We compare our method with state-of-the-art event keypoint detection methods [4, 18, 27, 46]. As shown in Table 4, our method outperforms SuperEvent [4] by 3% in feature matching IoU. Furthermore, as illustrated in Figure 4, our model can establish reliable keypoint matches, even under extreme scale and viewpoint variations. The superior performance can be attributed to two key innovations. Specifically, our G2PL module can leverage fine-grained spatio-temporal cues encoded in event points to capture keypoint distribution patterns across scales, enabling scale-invariant feature learning for event frame prototypes. Besides, the CDDL module can focus keypoint detection on co-visible regions, ensuring accurate matching even under significant viewpoint changes.

Results on HVGA ATIS Corner dataset. As shown in Table 5, we conduct a comprehensive comparison with state-of-the-art event keypoint detection methods [4, 7, 18, 27, 32, 33, 43, 46]. The results demonstrate that our proposed method outperforms existing methods by a large mar-

Table 6. Evaluation results for pose estimation on the EDS dataset. We report the AUC of the pose error at thresholds (5° , 10° , 20°).

Methods	AUC@ 5°	AUC@ 10°	AUC@ 20°
SD2Event [18]	11.5	22.1	35.7
STPNet [46]	13.9	24.9	38.2
SuperEvent [4]	15.2	26.4	40.1
EV-CGNet (ours)	26.8	40.3	58.7

Table 7. Evaluation results for SLAM on large-scale indoor sequences of the VECtor dataset. We follow the procedure described in [4] to integrate the following event keypoint detection methods into OKVIS2 [30], and report the ATE in cm.

Methods	corr-dolly	corr-walk	units-dolly	units-scooter	school-dolly	school-scooter
SD2Event [18]	50.85	102.71	127.61	132.43	57.25	63.47
STPNet [46]	41.04	85.47	97.03	113.05	47.21	54.59
SuperEvent [4]	33.13	133.16	122.61	59.05	69.96	39.07
EV-CGNet (ours)	10.03	21.06	25.82	31.03	13.93	18.58

Table 8. Evaluation results for SLAM on the small-scale motion capture sequences of the TUM-VIE dataset. We follow the procedure described in [4] to integrate the following event keypoint detection methods into OKVIS2 [30], and report the ATE in cm.

Methods	1d-trans	3d-trans	6dof	desk	desk2	Average
SD2Event [18]	0.62	0.68	0.72	0.71	0.66	0.68
STPNet [46]	0.51	0.65	0.61	0.57	0.52	0.57
SuperEvent [4]	0.44	0.89	0.43	0.58	0.41	0.55
EV-CGNet (ours)	0.21	0.24	0.27	0.25	0.23	0.24

Table 9. Evaluation results for SLAM on the long-distance loop closure sequences of the TUM-VIE dataset. We follow the procedure described in [4] to integrate the following event keypoint detection methods into OKVIS2 [30], and report the ATE in cm.

Methods	loop-floor 0	loop-floor 1	loop-floor 2	loop-floor 3
SD2Event [18]	7.02	6.75	6.93	7.14
STPNet [46]	5.27	5.13	4.97	5.03
SuperEvent [4]	4.96	4.64	8.92	4.74
EV-CGNet (ours)	1.35	1.24	1.03	1.12

gin. Specifically, our approach reduces reprojection error by 0.16, 0.18, and 0.21 pixels compared to SuperEvent [4] for event frame pairs with 25ms, 50ms, and 100ms intervals, respectively. Furthermore, the visualizations in Figure 5 show that our method achieves robust keypoint detection and superior keypoint matching, highlighting the effectiveness of the two proposed modules. Our G2PL module effectively captures fine-grained spatio-temporal information from event points, enhancing the encoding of texture change details and promoting the learning of event frame feature prototypes. And the CDDL module ensures accurate keypoint matching by focusing detection on co-visible regions, even under significant viewpoint variations.

Results on EDS dataset. We compare our method with state-of-the-art event keypoint detection approaches [4, 18,

46]. As shown in Table 6, our proposed method achieves the best pose accuracy among existing methods. Compared with SuperEvent [4], our model improves by 11.6% in AUC@ 5° , 13.9% in AUC@ 10° and 18.6% in AUC@ 20° . We attribute the superior performance to our G2PL and CDDL modules. The G2PL module effectively utilizes fine-grained spatio-temporal information encoded in event data to guide the learning of event frame features, capturing diverse keypoint motion patterns across frames. Additionally, the CDDL module enhances keypoint detection by focusing on co-visible regions, ensuring robust keypoint detection and matching, even under extreme viewpoint variations.

Results on VECtor dataset. We integrate state-of-the-art event keypoint detection methods [4, 18, 46] into OKVIS2 [30] and evaluate their performance for SLAM, as detailed in [4]. By setting the two input event streams identical, our method performs single-frame keypoint detection comparable to mainstream approaches. For loop detection, we adopt a coarse-to-fine pipeline: keyframe descriptors are first queried against a DBOW2 vocabulary [15] to retrieve candidates, then the current frame is matched with the top three, and the pair with the highest score identifies the loop. As shown in Table 7, our proposed method outperforms all other methods by a large margin. Specifically, we achieve a 68% reduction in average ATE compared to SuperEvent [4]. This improvement can be attributed to our G2PL module, which effectively leverages fine-grained spatio-temporal cues from event data, enabling the capture of diverse texture change patterns. And our CDDL module can focus keypoint detection on co-visible regions, enabling robust keypoint matching. This significantly improves pose estimation and enhances loop detection, further reducing cumulative errors via loop closure implemented in [30].

Results on TUM-VIE dataset. We adopt the procedure described in [4] to integrate state-of-the-art event keypoint detection methods [4, 18, 46] into OKVIS2 [30] for evaluation on small-scale motion capture sequences. As shown in Table 8, our method consistently outperforms existing approaches across all sequences. Specifically, compared to SuperEvent [4], our approach achieves significant improvements on all sequences, with an average reduction of 0.31 cm in ATE. To further verify the effectiveness of our model in loop detection, we perform an evaluation on long-distance loop closure sequences. As demonstrated in Table 9, our method achieves substantial improvements over existing event keypoint detection methods, with a 78% reduction in ATE compared to SuperEvent [4]. These results clearly highlight the superiority of our approach in event keypoint matching and its potential to enhance SLAM performance. These improvements can be attributed to the well-designed G2PL and CDDL modules. Specifically, the G2PL module can effectively leverage fine-grained spatio-temporal information from the event data, facilitating ro-

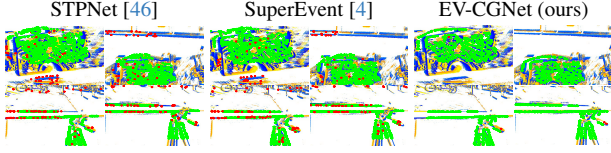


Figure 5. Qualitative comparisons on the HVGA ATIS Corner dataset. Green and red stars denote correct and incorrect matches, respectively (threshold is 3px).

Table 10. Effectiveness of each component on the HVGA ATIS Corner dataset. We report the reprojection error in pixels across event frame pairs at time intervals of $25ms$, $50ms$, and $100ms$.

Models	CDDL	G2PL	$25ms$	$50ms$	$100ms$
[A]	✗	✗	5.36	10.17	16.20
[B]	✗	✓	0.51	0.58	0.73
[C]	✓	✗	0.39	0.46	0.59
[D]	✓	✓	0.27	0.31	0.42

bust keypoint tracking. And the CDDL module can focus keypoint detection on co-visible regions, ensuring accurate keypoint matching, which is crucial for both the front-end and loop closure components in [30].

4.4. Ablation Studies

To assess the contribution of each component in our proposed method, we conduct a series of ablation studies on the HVGA ATIS Corner dataset. In Table 10, model [A] employs a 1×1 convolutional layer to extract event frame features $\mathbf{E}$ and a separate 1×1 convolutional layer for keypoint detection. In contrast, model [B] introduces a dot product operation between the 3D-guided 2D prototypes $\mathbf{P}_D$ and the event frame features $\mathbf{F}_E$ to extract keypoints. Model [C] uses a 1×1 convolutional layer to extract event frame features $\mathbf{E}$, then initializes a set of agents for each of the two event frames. These agents interact with event frame features independently, and the updated agents are subsequently fed into the CDDL module. Finally, model [D] represents the complete architecture of our proposed method.

Effectiveness of the G2PL module. As shown in Table 10, the performance on the HVGA ATIS Corner improves significantly with our G2PL module. Specifically, model [B] reduces reprojection error by 4.85, 9.59, and 15.47 pixels compared to model [A] for event frame pairs with $25ms$, $50ms$, and $100ms$ intervals, respectively. Besides, model [D] also exhibits superior performance compared to model [C]. The main reason is that our G2PL module can effectively leverage fine-grained spatio-temporal cues from event points to enhance event frame features. This proves particularly beneficial in addressing challenging factors, such as diverse motion patterns and significant texture variations.

Impacts about the number of event point prototypes in

Table 11. Impacts of the number of event point prototypes on HVGA ATIS Corner. We report the reprojection error in pixels across event frame pairs at time intervals of $25ms$, $50ms$, and $100ms$.

Models	$25ms$	$50ms$	$100ms$
$N=2$	0.46	0.52	0.66
$N=4$	0.36	0.41	0.54
$N=8$	0.31	0.35	0.47
$N=16$	0.27	0.31	0.42
$N=32$	0.29	0.33	0.44

the G2PL module. To investigate the influences of the number of event point prototypes in the G2PL module, we pick N from the set $\{2, 4, 8, 16, 32\}$ and evaluate the performance on the HVGA ATIS Corner dataset. As shown in Table 11, we find that setting $N = 16$ yields the best performance. There is no performance gain when N continues to increase. This observation suggests that the model with $N = 16$ is sufficient to capture spatio-temporal cues in event points on the HVGA ATIS Corner dataset.

Effectiveness of the CDDL module. As shown in Table 10, the integration of our proposed CDDL module leads to significant performance enhancements on the HVGA ATIS Corner dataset. Specifically, model [C] achieves substantial reductions in reprojection error by 4.97, 9.71, and 15.61 pixels for event frame pairs at time intervals of $25ms$, $50ms$, and $100ms$, respectively, compared to model [A]. Besides, model [D] also demonstrates better performance than model [B]. The primary reason is that our proposed CDDL module can concentrate keypoint detection on co-visible regions, ensuring accurate matches.

5. Conclusion

In this work, we propose a novel event keypoint detection method by integrating a distinctive 3D-guided 2D paradigm and an innovative strategy that focuses keypoint detection on co-visible regions. Our approach consists of two main components, including a G2PL module and a CDDL module. With these well-designed modules, our proposed method can effectively exploit fine-grained spatio-temporal cues in event data and focus keypoint detection on co-visible regions, enabling robust keypoint matching even under some extremely challenging factors. Extensive experimental results on six challenging benchmarks demonstrate the effectiveness of our proposed method.

6. Acknowledgement

This work was partially supported by the National Key Research and Development Program of China (2024YFB3909902) and Key Technology Research Project of TW-3 (TW3004).

References

- [1] Ignacio Alzugaray and Margarita Chli. Asynchronous corner detection and tracking for event cameras in real time. *IEEE Robotics and Automation Letters*, 3(4):3177–3184, 2018. [1](#), [3](#), [5](#)
- [2] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):407–417, 2014. [3](#)
- [3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbrück. A 240×180 130 db $3 \mu\text{s}$ latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. [5](#)
- [4] Yannick Burkhardt, Simon Schaefer, and Stefan Leutenegger. Superevent: Cross-modal learning of event-based keypoint detection for slam. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8918–8928, 2025. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [5] Yang Chen and Lin Zhang. GRE-SLAM: 6-dof pure event-based SLAM with semi-dense depth recovery assisted bundle adjustment. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 90–98, 2025. [1](#)
- [6] Zehao Chen, Zhanfeng Liao, De Ma, Huajin Tang, Qian Zheng, and Gang Pan. Evhdr-nerf: Building high dynamic range radiance fields with single exposure images and events. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, pages 2376–2384, 2025. [1](#)
- [7] Philippe Chiberre, Etienne Perot, Amos Sironi, and Vincent Lepetit. Detecting stable keypoints from events through image gradient prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1387–1394, 2021. [5](#), [6](#)
- [8] Xavier Clady, Sio-Hoi Ieng, and Ryad Benosman. Asynchronous event-based corner detection and matching. *Neural Networks*, 66:91–106, 2015. [1](#), [3](#)
- [9] Hiroyuki Deguchi, Mana Masuda, Takuya Nakabayashi, and Hideo Saito. E2GS: event enhanced gaussian splatting. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1676–1682. IEEE, 2024. [1](#)
- [10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. [2](#), [3](#), [5](#)
- [11] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8092–8101, 2019. [3](#)
- [12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996. [1](#), [3](#)
- [13] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–178. IEEE, 2004. [5](#)
- [14] 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, 2022. [1](#)
- [15] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012. [7](#)
- [16] Ling Gao, Yuxuan Liang, Jiaqi Yang, Shaoxun Wu, Chenyu Wang, Jiaben Chen, and Laurent Kneip. Vector: A versatile event-centric benchmark for multi-sensor SLAM. *IEEE Robotics and Automation Letters*, 7(3):8217–8224, 2022. [5](#)
- [17] Yuan Gao, Jianfeng He, Tianzhu Zhang, Zhe Zhang, and Yongdong Zhang. Dynamic keypoint detection network for image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. [2](#), [3](#)
- [18] Yuan Gao, Yuqing Zhu, Xinjun Li, Yimin Du, and Tianzhu Zhang. Sd2event: Self-supervised learning of dynamic detectors and contextual descriptors for event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3055–3064, 2024. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [19] 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. [5](#)
- [20] Arren Glover, Aiko Dinale, Leandro De Souza Rosa, Simeon Bamford, and Chiara Bartolozzi. Iuvharris: A practical corner detector for event-cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10087–10098, 2022. [1](#)
- [21] Shuang Guo and Guillermo Gallego. Cmax-slam: Event-based rotational-motion bundle adjustment and SLAM system using contrast maximization. *IEEE Transactions on Robotics*, 40:2442–2461, 2024. [1](#)
- [22] Haiqian Han, Jianing Li, Henglu Wei, and Xiangyang Ji. Event-3dgs: Event-based 3d reconstruction using 3d gaussian splatting. In *Advances in Neural Information Processing Systems*, 2024. [1](#)
- [23] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, 1988. [2](#), [3](#)
- [24] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2022. [5](#)
- [25] Sumin Hu, Yeeun Kim, Hyungtae Lim, Alex Junho Lee, and Hyun Myung. ecdt: Event clustering for simultaneous feature detection and tracking. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 3808–3815, 2022. [1](#), [3](#)
- [26] Jian Huang, Chengrui Dong, Xuanhua Chen, and Peidong Liu. Inceventgs: Pose-free gaussian splatting from a single event camera. In *Proceedings of the IEEE Conference*

- on *Computer Vision and Pattern Recognition*, pages 26933–26942, 2025. 1
- [27] Ze Huang, Li Sun, Cheng Zhao, Song Li, and Songzhi Su. Eventpoint: Self-supervised interest point detection and description for event-based camera. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 5396–5405, 2023. 1, 2, 3, 5, 6
- [28] Simon Klenk, Jason Chui, Nikolaus Demmel, and Daniel Cremers. TUM-VIE: The TUM Stereo Visual-Inertial Event Dataset. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 8601–8608, 2021. 5
- [29] Simon Klenk, Marvin Motzet, Lukas Koestler, and Daniel Cremers. Deep event visual odometry. In *Proceedings of the International Conference on 3D Vision*, pages 739–749, 2024. 1
- [30] Stefan Leutenegger. Okvis2: Realtime scalable visual-inertial slam with loop closure. *arXiv preprint arXiv:2202.09199*, 2022. 7, 8
- [31] Ruoxiang Li, Dianxi Shi, Yongjun Zhang, Kaiyue Li, and Ruihao Li. Fa-harris: A fast and asynchronous corner detector for event cameras. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 6223–6229, 2019. 1
- [32] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. Speed invariant time surface for learning to detect corner points with event-based cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10245–10254, 2019. 1, 3, 5, 6
- [33] Elias Mueggler, Chiara Bartolozzi, and Davide Scaramuzza. Fast event-based corner detection. In *Proceedings of the British Machine Vision Conference*, 2017. 1, 3, 6
- [34] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbrück, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. 1, 5
- [35] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. 5
- [36] Tianbo Pan, Zidong Cao, and Lin Wang. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 10695–10702, 2024. 1
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 5
- [38] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2011. 5
- [39] Bharath Ramesh, Hong Yang, Garrick Orchard, Ngoc Anh Le Thi, Shihao Zhang, and Cheng Xiang. Dart: distribution aware retinal transform for event-based cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2767–2780, 2020. 1, 5
- [40] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018. 1
- [41] Paul Rogister, Ryad Benosman, Sio-Hoi Ieng, Patrick Lichtsteiner, and Tobi Delbrück. Asynchronous event-based binocular stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*, 23(2):347–353, 2011. 1
- [42] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Proceedings of the European Conference on Computer Vision*, pages 430–443, 2006. 2, 3
- [43] Valentina Vasco, Arren Glover, and Chiara Bartolozzi. Fast event-based harris corner detection exploiting the advantages of event-driven cameras. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 4144–4149, 2016. 1, 3, 5, 6
- [44] 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. 4
- [45] Tianyi Xiong, Jiayi Wu, Botao He, Cornelia Fermüller, Yiannis Aloimonos, Heng Huang, and Christopher A. Metzler. Event3dgs: Event-based 3d gaussian splatting for high-speed robot egomotion. In *Proceedings of the Conference on Robot Learning*, pages 4100–4118, 2024. 1
- [46] Yuqing Zhu, Yuan Gao, Tianle Ding, Xiang Liu, Wenfei Yang, and Tianzhu Zhang. Spatio-temporal pyramid key-point detection with event cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):9384–9397, 2025. 1, 2, 3, 5, 6, 7, 8