# Unleashing the Temporal Potential of Stereo Event Cameras for Continuous-Time 3D Object Detection

Jae-Young Kang*, Hoonhee Cho*, and Kuk-Jin Yoon
KAIST
{mickeykang,gnsgnsgml,kjyoon}@kaist.ac.kr

## Abstract

*3D object detection is essential for autonomous systems, enabling precise localization and dimension estimation. While LiDAR and RGB cameras are widely used, their fixed frame rates create perception gaps in high-speed scenarios. Event cameras, with their asynchronous nature and high temporal resolution, offer a solution by capturing motion continuously. The recent approach, which integrates event cameras with conventional sensors for continuous-time detection, struggles in fast-motion scenarios due to its dependency on synchronized sensors. We propose a novel stereo 3D object detection framework that relies solely on event cameras, eliminating the need for conventional 3D sensors. To compensate for the lack of semantic and geometric information in event data, we introduce a dual filter mechanism that extracts both. Additionally, we enhance regression by aligning bounding boxes with object-centric information. Experiments show that our method outperforms prior approaches in dynamic environments, demonstrating the potential of event cameras for robust, continuous-time 3D perception. The code is available at https://github.com/mickeykang16/Ev-Stereo3D.*

## 1. Introduction

3D object detection is essential for accurately identifying the position and dimensions of surrounding objects, making it a fundamental task in autonomous systems. Multi-modal sensors, such as LiDAR and RGB cameras, play a key role. However, fixed-frame rate sensors have discrete intervals during which data cannot be collected, resulting in inevitable detection delays. This delay in capturing dynamic changes can significantly impact real-time decision-making, emphasizing the importance of high temporal resolution for reliable autonomous driving.

Event cameras [27] are asynchronous sensors that continuously stream data over time, allowing them to break the trade-off between bandwidth and delay. Their high temporal resolution makes them well-suited for capturing fast motion, enabling high-speed perception even in dynamic scenarios. This capability is particularly beneficial for autonomous driving, where rapid response is critical for accident prevention and safe navigation.

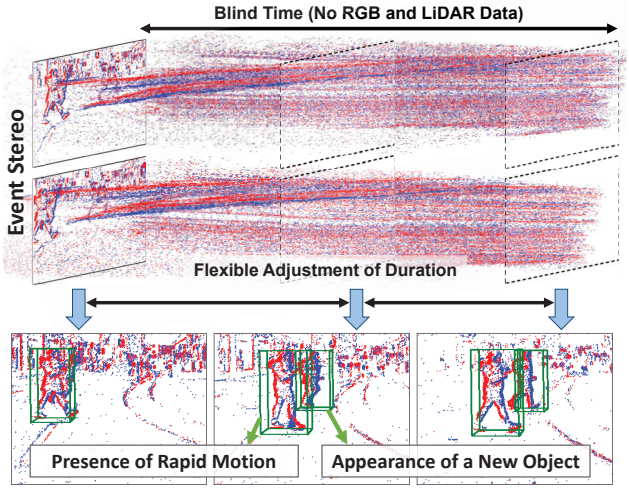


Figure 1. During blind time, our method enables fast and large motion-aware continuous 3D object detection at arbitrary points in time, including for newly appearing objects.

A recent work, Ev-3DOD [22], integrates an RGB camera, LiDAR, and an event camera to overcome the frame-rate limitations of conventional sensors in 3D object detection. By incorporating event data, it enables perception even during periods when LiDAR and RGB cameras are unavailable. Notably, it introduces continuous-time 3D object detection, allowing inference at arbitrary timestamps before the next synchronized sensor data arrives. Continuous-time 3D object detection enables accurate inference beyond the timeframes when synchronized sensors, such as RGB and LiDAR, provide data. Ev-3DOD [22] defined periods with LiDAR and RGB data as active time and intervals without them as blind time, demonstrating that event cameras can facilitate 3D detection during these blind periods.

However, approaches that temporally propagate LiDAR-

*Equal contribution.

and image-based detection with events are highly dependent on synchronized sensors introducing limitations in dynamic environments. For example, in scenarios like drones with rapid motion during adjacent active times, or high-speed driving causing significant movement during blind periods, fixed frame-rate sensor data does not fully reflect the current 3D geometry. As a result, errors are accumulated over blind time, leading to frequent detection failures.

To this end, we propose an event stereo setup that performs precise 3D object detection using only spare and asynchronous event cameras to address the limitations of conventional sensors. As shown in Fig. 1, event stereo can compute 3D information even during blind time, enabling the detection of large-moving and newly appearing objects, which are scenarios that synchronized sensors cannot handle in blind time.

One of the challenges when using only event data for 3D object detection is its sparse nature. Event data lacks both semantic and geometric information compared to images and LiDAR, making accurate classification and regression particularly difficult. To overcome this, we propose a dual semantic-geometric filter module that collaboratively filters and enhances both information, fully leveraging stereo event data. The event semantic feature filters the geometric depth feature while the geometric feature simultaneously enhances the semantic information through the estimated disparity. Additionally, following the global detection head, a dense, object-centric semantic event feature is employed to align the local 3D offsets of the bounding boxes, thereby enhancing regression performance. Our proposed method achieves comparable performance in continuous-time 3D object detection without relying on rich fixed frame-rate sensor data. Specifically, through experiments in which the scene undergoes significant changes due to large motion within the same duration, as well as tests involving various intervals for inference, our approach outperforms existing methods. This result can be attributed to the use of asynchronous sensors, which offer distinct advantages in handling dynamic scenarios while fully utilizing both semantic and geometric information of events. The paper highlights the potential and scalability of event cameras to enhance 3D object detection in various environments.

## 2. Related Works

**Camera-based 3D Object Detection.** Cameras offer significant cost advantages over LiDAR, leading many researchers to focus on developing methods for 3D object detection using image-only inputs [15, 39, 43, 56, 59, 61, 64, 84, 87, 88]. However, image-based 3D detection faces challenges due to the lack of direct depth information. To address this, several works [11, 12, 34, 48, 60] propose using stereo and multi-camera systems to compute 3D geometry through multiple views. Despite these advancements, accurately estimating depth from images remains challenging, causing image-based methods to underperform compared to LiDAR-based approaches.

**LiDAR-based 3D Object Detection.** LiDAR-based 3D object detection can reliably estimate 3D bounding boxes using point clouds captured from LiDAR sensors. Research on representing 3D information from LiDAR has been conducted through various approaches, including point-based methods [25, 38, 80, 82, 95, 98, 104], voxel-based methods [24, 33, 37, 37, 45, 51, 65, 79, 89, 92, 100], and point-voxel fusion networks [36, 66, 81, 94]. To further boost the performance of LiDAR-based 3D detection, multi-modal sensor fusion [9, 40, 41, 49, 50, 58, 83, 90, 99] has emerged, where different sensors, such as cameras and LiDAR, complement each other. Fusion approaches are categorized into early [16, 85, 93], middle [53–55, 57, 62, 74, 76, 96], and late fusion [2, 49, 69]. Multi-modal 3D object detection has garnered significant attention for its ability to improve accuracy and enhance performance compared to unimodal methods. However, there is still room for improvement in terms of safety in autonomous driving. LiDAR and cameras have limited time resolution (*i.e.*, 10-20 Hz) because of their relatively high bandwidths. These limitations are especially noticeable in scenarios like high-speed driving or drone flight. To improve safety, algorithms that can operate at higher speeds are crucial.

**Event-based Stereo.** Event cameras are well-suited for capturing dynamic scenes due to their high temporal resolution, low latency, and ability to operate effectively in challenging lighting conditions. Based on these characteristics, approaches using event cameras in a stereo setup [1, 19–21, 67, 68, 102, 105, 107] have been proposed to better understand the surrounding 3D dynamic scenes. Additionally, research has explored using event cameras in a stereo setup alongside other sensors [3, 10, 17, 18, 63, 103], aiming to leverage the strengths of both event data and conventional sensors to enhance perception capabilities. Building on this research direction, we demonstrate that 3D object detection can be performed using only asynchronous event cameras. Furthermore, we highlight the potential of event cameras for 3D detection in dynamic and rapid motion scenarios.

**Event-based Object Detection.** On the 2D image plane, numerous studies [6, 28, 30, 35, 42, 46, 47, 52, 70, 71, 73, 86, 108] have leveraged the event camera's sensitivity to moving objects for object detection. Various architectures, including graph-based [5, 29, 77], spiking [97], and sparse [72] neural networks, have been explored to achieve streamlined and compelling results tailored to specific objectives. A recent study [22] demonstrated the advantages of high temporal resolution of event cameras in 3D object detection by integrating them with LiDAR and RGB cameras. This study addressed the limitation of synchronized sensors, such as RGB cameras and LiDAR, which operate

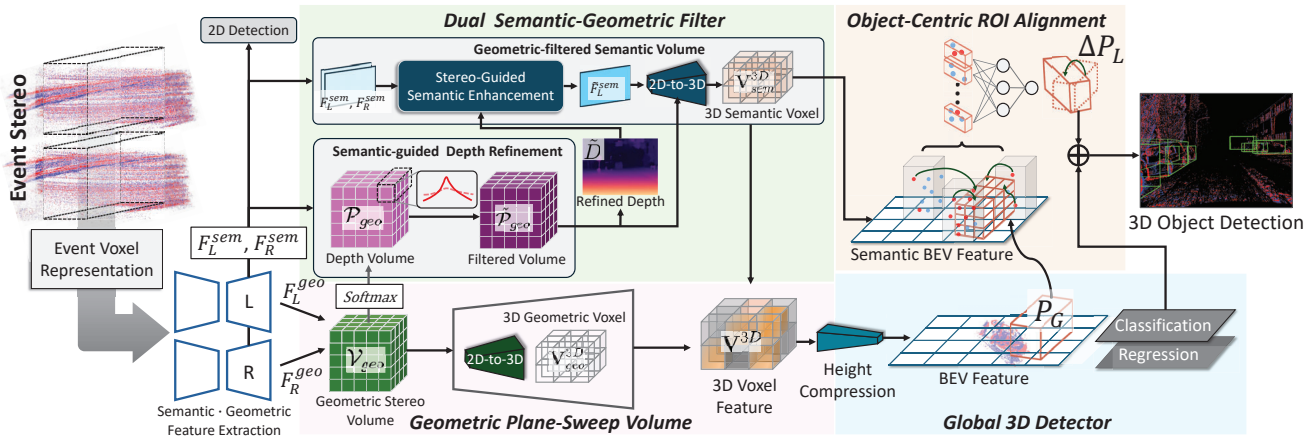


Figure 2. Overview of the proposed method. Our method is categorized into four components: Geometric Plane-Sweep Volume (Sec. 3.2), Dual Semantic-Geometric Filter (Sec. 3.3), Global 3D Detector (Sec. 3.4), and Object-Centric ROI Alignment (Sec. 3.5).

at fixed frame rates and restrict inference to discrete time intervals. By integrating event cameras, the proposed approach overcomes this constraint, enabling continuous 3D detection even in the absence of synchronized sensor input.

Despite these promising results, the approach remains partially dependent on synchronization, as critical information is solely derived from RGB and LiDAR. Consequently, significant errors may occur when rapid motion leads to substantial scene changes during blind time. To address this limitation, we propose the first fully asynchronous system for 3D object detection, utilizing a stereo event camera setup. Thanks to this asynchronous advantage, our method maintains strong performance even when the scene undergoes significant changes during the blind time or when inference is required at a finer continuous-time scale.

# 3. Methods

## 3.1. Preliminaries

**Event Camera.** Traditional frame-based cameras capture images at fixed intervals, whereas event cameras operate asynchronously, recording changes in logarithmic intensity $L(u, v, t)$ only when they exceed a predefined threshold $\mathcal{C}$. This can be expressed as:

$$L(u, v, t) - L(u, v, t - \Delta t) \geq \mathbf{p}\mathcal{C}, \quad \mathbf{p} \in \{-1, +1\}, \quad (1)$$

where $\Delta t$ represents the time difference, and $\mathbf{p}$ denotes event polarity, indicating either negative or positive events. The event stream $\mathcal{E}$ consists of events $e_i$ with spatial coordinates $(u_i, v_i)$, timestamp $t_i$, and polarity $p_i$:

$$\mathcal{E} = \{e_i\}_{i=1}^{N}, \quad e_i = (u_i, v_i, t_i, p_i). \quad (2)$$

**Event Representation.** We adopt a voxel grid [106] representation, where events are discretized along the time axis into $B$ bins. Given $N$ input events, timestamps are normalized to $[0, B-1]$, and the event volume is constructed as:

$$E(u, v, b) = \sum_{i=1}^{N} p_i k_{\mathcal{B}}(u - u_i) k_{\mathcal{B}}(v - v_i) k_{\mathcal{B}}(b - b_i^*), \quad (3)$$

where $b_i^* = (B-1)(t_i - t_1)/(t_N - t_1)$ and $k_{\mathcal{B}}(a)$ is a bilinear sampling kernel. The voxel grid allows efficient feature extraction via 2D convolution across spatial dimensions while conserving temporal information.

We use the event stream from the recent interval $\Delta\tau$, denoted as $\mathcal{E}_{(\tau-\Delta\tau)\to\tau} = \{(u, v, t, p) \mid \tau - \Delta\tau \leq t < \tau\}$, to detect 3D boxes at time $\tau$. To achieve the continuous-time detection framework, $\Delta\tau$ can adapt flexibly to different scenarios. This enables detection at arbitrary time points, making the framework adaptable to various situations. To simplify notation in this section, we represent the voxel grids generated from $\mathcal{E}_{(\tau-\Delta\tau)\to\tau}$ as $E_L$ and $E_R$ for the left and right cameras, respectively.

## 3.2. Geometric Plane-Sweep Volume

We illustrate the overall training pipeline in Fig. 2. We modify the PSMNet [7] to design our feature extractor and incorporated an additional feature head to disentangle the roles of semantics-based detection and geometric scene construction. This allows the extractor to separately extract semantic features $F_{L,R}^{sem}$ and geometric features $F_{L,R}^{geo}$, each with a dimension of $\mathbb{R}^{H \times W \times C}$. The semantic feature activates well on objects through an auxiliary 2D dense head on the 2D image. The geometric feature activates across the entire scene on the stereo images, concentrating on finding correspondences. First, we generate a geometric plane-sweep volume $\mathcal{V}_{geo} \in \mathbb{R}^{H \times W \times D \times C}$ by concatenating left and corresponding right geometric features at each depth level $D$, as done in previous works [11, 34]. The volume is defined as:

$$\mathcal{V}_{geo}(u, v, w) = (F_L^{geo}(u, v) || F_R^{geo}(u - \frac{fL}{d(w)}, v)), \quad (4)$$

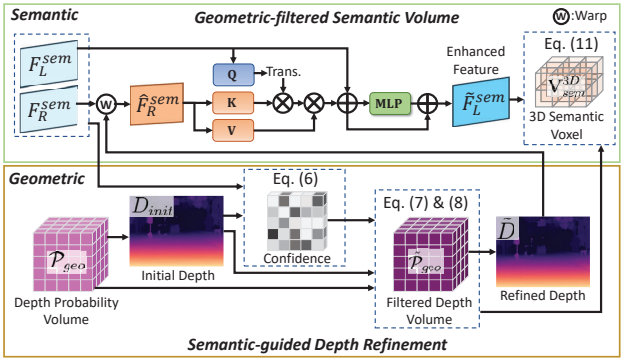


Figure 3. Dual semantic-geometric filter. Semantic and geometric information interact complementarily to enhance both the semantic features and the depth volume.

where $(u, v)$ are pixel coordinates, $||$ denotes the concatenation, $w = 0, 1, \ldots$ is the candidate depth index, and $d(w) = w \cdot v_d + z_{\text{min}}$ is the function to calculate the corresponding depth, where $v_d$ is the depth interval and $z_{\text{min}}$ is the minimum depth of the detection area. $f$ and $L$ represent the camera focal length and the baseline of the stereo camera pair, respectively. From the geometric stereo volume $\mathcal{V}_{\text{geo}}$, we apply 3D convolutions to reduce the channel dimension, followed by a softmax operation along the depth dimension, resulting in a depth probability volume $\mathcal{P}_{\text{geo}} \in \mathbb{R}^{H \times W \times D}$, where $\mathcal{P}_{\text{geo}}(u, v, :)$ represents the depth probability for pixel $(u, v)$ across all depth levels.

To transform the feature volume from stereo space to a 3D space, the 3D detection area is segmented into uniformly sized voxels, resulting in $\mathbf{V}^{3D}_{geo} \in \mathbb{R}^{X \times Y \times Z \times C}$. For each voxel, its center coordinates $(x, y, z)$ are projected back into the stereo space using the intrinsic parameters $K$, resulting in the reprojected pixel coordinates $(u, v)$ and a corresponding depth index $d^{-1}(z) = (z - z_{\text{min}})/v_d$ as:

$$\mathbf{V}^{3D}_{geo}(x, y, z) = \mathcal{V}_{geo}(u, v, d^{-1}(z)). \tag{5}$$

### 3.3. Dual Semantic-Geometric Filter

Along with the geometric volume, we leverage the event camera's robustness in detecting objects on the 2D plane by using semantic features to construct a 3D semantic volume $\mathbf{V}^{3D}_{sem} \in \mathbb{R}^{X \times Y \times Z \times C}$. Creating a 3D volume with only semantic features fails to capture geometric properties. To address this, we propose a dual semantic-geometric filter that combines both semantic and geometric information, as shown in Figure 3.

**Semantic-guided Depth Refinement.** One of the key challenges in 3D detection using stereo cameras is achieving accurate depth inference. Unlike LiDAR, stereo cameras do not provide direct depth measurements from the sensor. To enhance depth accuracy, we utilize semantic information to refine the geometric cost volume. As shown in Fig. 3, we first estimate the initial depth $D_{init}$ by computing the weighted sum over all depth levels using the depth probability volume $\mathcal{P}_{geo}$. Additionally, to refine the depth distribution volume by transferring accuracy through the semantic feature, which effectively captures semantic information about the scene, we compute the similarity for each pixel based on the predicted depth as:

$$S(u, v) = \langle F^{sem}_L(u, v), F^{sem}_R(u - \frac{fL}{D_{init}(u, v)}, v) \rangle, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product for similarity. Similar to previous works [8, 75, 78, 91] that handle confidence in stereo matching, we also compute confidence as follows:

$$C(u, v) = \sum_w \mathcal{P}_{geo}(u, v, w) \cdot \left(d(w) - D^{init}(u, v)\right)^2, \tag{7}$$

We aim to refine the depth distribution by considering both stereo similarity and confidence. Adapting the refinement process from [91], we generate filtered volume, $\tilde{\mathcal{P}}_{geo}$, by incorporating neighboring probabilities as follows:

$$W_m(u, v) = S_m(u, v) \cdot \text{sigmoid}\left(C_m(u, v)\right),$$
$$\tilde{\mathcal{P}}_{geo}(u, v, w) = \sum_{m=1}^{M} \mathcal{P}_{geo}(u, v, w) \cdot \text{softmax}_m(W_m(u, v)), \tag{8}$$

where $m = 1, 2, 3, \ldots, M$ are the sampled neighboring pixels and $\text{softmax}_m$ denotes a softmax across the $m$ dimension. Through this refinement process, depth ambiguity can be resolved, and we estimate the refined depth, $\tilde{D}$, from the refined probability.

**Geometric-filtered Semantic Volume.** Although semantic features are well activated for objects, event data is sparse along the spatial dimension, often leading to ambiguity. To compensate for this lack of information, we enhance the left features using the estimated depth and the semantic features from the right camera. As shown in the bottom of Fig. 3, we warp the right semantic features to the left camera using the estimated depth: $\hat{F}^{sem}_R = \text{Warp}(F^{sem}_R, \tilde{D})$. However, directly using these warped features may introduce occlusion issues and fail to fully address potential misalignments between stereo cameras. To handle these challenges, we apply a transformer-based channel attention mechanism to generate enhanced semantic features. We generate the query, key, and value features as follows: $\mathbf{Q} = W_Q(F^{sem}_L)$, $\mathbf{K} = W_K(\hat{F}^{sem}_R)$, and $\mathbf{V} = W_V(\hat{F}^{sem}_R)$, where $W_{(\cdot)}$ represents a $1 \times 1$ convolution. The results of the transformer are computed in this manner:

$$\mathbb{A}(F^{sem}_L, \hat{F}^{sem}_R) = F^{sem}_L + W_{\mathbb{A}}(\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\alpha}\right) \cdot \mathbf{V}) \tag{9}$$

where $\alpha$ is a learnable parameter. Finally, the geometrically enhanced semantic feature $\tilde{F}^{sem}_L$ is obtained:

$$\tilde{F}^{sem}_L = F^{sem}_L + \text{MLP}(\mathbb{A}) + \mathbb{A}. \tag{10}$$

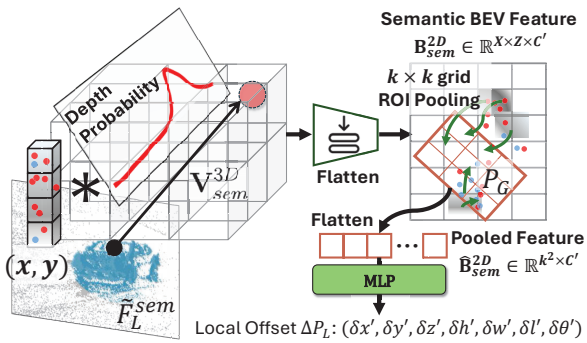


Figure 4. Illustration of the object-centric ROI alignment.

Similar to Eq. (5), voxels can be projected onto the image plane to map 2D semantic features to 3D. However, to ensure the transformation is based on geometric information, we apply a depth probability mask as:

$$\mathbf{V}^{3D}_{sem}(x, y, z) = \tilde{F}^{sem}_L(u, v) \cdot \tilde{\mathcal{P}}_{geo}(u, v, d^{-1}(z)) \quad (11)$$

**Fusion of Semantic and Geometric Voxels.** We generate two types of voxels in the 3D space: one specialized for geometric information, $\mathbf{V}^{3D}_{geo}$, and the other for semantic information, $\mathbf{V}^{3D}_{sem}$. To integrate the two voxels into a unified 3D voxel that effectively combines both aspects, we concatenated and then merged them through 3D convolution, resulting in the fused voxel, $\mathbf{V}^{3D}$.

### 3.4. Global 3D Detector

Inspired by recent stereo 3D detection approaches [11, 12, 34], we designed an anchor-based detector. The bird's-eye view (BEV) feature strikes a balance between accuracy and computational efficiency in 3D detection for autonomous driving. The unified 3D voxel features $\mathbf{V}^{3D} \in \mathbb{R}^{X\times Y\times Z\times C}$ are downsampled along the height dimension to produce a BEV feature of size $(X, Z)$. We utilize fixed-size class anchors, determined based on the training set statistics, and assign a predefined number of anchors with varying sizes and orientations to each BEV feature location $(x, z)$. Similar to previous works [11, 12, 34], the dense head network performs a regression from the anchor $A = (x_a, y_a, z_a, h_a, w_a, l_a, \theta_a)$ to predict the offset $\Delta P_G = (\delta x, \delta y, \delta z, \delta h, \delta w, \delta l, \delta\theta)$. The final prediction is calculated by applying an offset to the anchor to determine the global position using the box decoding function $g_d$,

$$P_G = g_d(A, \Delta P_G) = (x_a + \delta x, y_a + \delta y, z_a + \delta z, h_a e^{\delta h}, w_a e^{\delta w}, l_a e^{\delta l}, \theta_a + \frac{\pi}{2}\tanh(\delta\theta)). \quad (12)$$

### 3.5. Object-Centric ROI Alignment

The geometric-semantic fused 3D voxels, $\mathbf{V}^{3D}$, generated through the semantic-geometric dual filter contain accurate depth and rich semantic information, enabling bounding box prediction in the global scene. However, a single-stage global regression can struggle with ambiguity caused by the motion of objects and the movement of the ego vehicle, especially as motion intensifies, necessitating a method that aligns more accurately with the object. To this end, we propose a method for locally aligning the ROI around the object's bounding box.

**ROI Pooling and Local Regression.** As shown in Fig. 4, we use the semantic voxel, $\mathbf{V}^{3D}_{sem}$, which focuses on the object and provides rich edge information from event-based semantic data, making it advantageous for ROI alignment. To optimize the computation of alignment, and because most stereo-based detections are prone to error along the $X$ and $Z$ dimensions, we project the voxel features onto the BEV, which corresponds to the transformation of $\mathbf{B}^{2D}_{sem} \in \mathbb{R}^{X\times Z\times C'}$ to the BEV of $\mathbf{V}^{3D}_{sem}$, rather than using them directly. Since $\mathbf{B}^{2D}_{sem}$ is splatted based on depth probabilities, it tends to cluster around the high-likelihood hypotheses of the predicted boxes. Therefore, to ensure that not only the exactly overlapping features but also the surrounding information is aggregated during alignment, we use a pooling scheme. Specifically, each bounding box prediction from global 3D detector, $P_G$, is split into an $k \times k$ voxel grid. The semantic BEV feature, $\mathbf{B}^{2D}_{sem}$, is then pooled [23] for each voxel grid. The pooled feature of each bounding box, $\hat{\mathbf{B}}^{2D}_{sem} \in \mathbb{R}^{k^2\times C'}$, encodes object-centric local information, serving as a cue for precisely aligning the box. Each pooled feature passes through an MLP layer to predict the local alignment offset, $\Delta P_L = (\delta x', \delta y', \delta z', \delta h', \delta w', \delta l', \delta\theta')$. The final fine-grained regression, $\tilde{P}_G$, is calculated by box decoding function: $\tilde{P}_G = g_d(P_G, \Delta P_L)$

### 3.6. Objective Function

A multi-task loss was applied to train the network, which includes the geometric plane-sweep volume, the dual semantic-geometric branch, the global 3D detector head, and object-centric ROI alignment module as follows:

$$\mathcal{L} = \mathcal{L}^{init}_{depth} + \mathcal{L}^{refine}_{depth} + \mathcal{L}_{2D} + \mathcal{L}_{cls} + \mathcal{L}^{global}_{reg} + \mathcal{L}^{local}_{reg}. \quad (13)$$

where $\mathcal{L}^{init}_{depth}, \mathcal{L}_{2D}, \mathcal{L}_{cls}$ are adapted from uni-modal depth loss, 2D detection loss, and classification loss in [34]. For refined depth loss, we utilized *smooth* $L_1$ loss [44] as

$$\mathcal{L}^{refine}_{depth} = smooth_{L_1}(D_{gt} - \tilde{D}). \quad (14)$$

The regression loss for the global 3D detector and object-centric ROI alignment are as follows:

$$\mathcal{L}^{global}_{reg} = \mathcal{L}_{reg}(P_G, G), \quad \mathcal{L}^{local}_{reg} = \mathcal{L}_{reg}(\tilde{P}_G, G) \quad (15)$$

where $G$ refers to 3D ground-truth bounding box and $\mathcal{L}_{reg}$ is 3D regression loss adapted from [34].

Table 1. Comparison on the DSEC-3DOD dataset for vehicle (IoU = 0.7) and pedestrian (IoU = 0.5) detection on 3D and BEV. We follow the evaluation protocol of Ev-3DOD [22] for 100 FPS 3D detection in blind time. Mod. is an abbreviation for moderate, and easy and moderate represent difficulty levels. VEH and PED represent vehicle and pedestrian, respectively.

| Modality | Methods | VEH $AP_{3D}$ | | VEH $AP_{BEV}$ | | PED $AP_{3D}$ | | PED $AP_{BEV}$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | Easy | Mod. | Easy | Mod. | Easy | Mod. | Easy | Mod. |
| LiDAR | VoxelNeXt [14] | 12.66 | 11.06 | 31.46 | 28.39 | 10.59 | 6.61 | 12.77 | 7.96 |
| | HEDNet [101] | 14.31 | 12.96 | 29.94 | 27.85 | 10.16 | 8.29 | 11.90 | 6.89 |
| LiDAR+RGB | Focals Conv [13] | 12.71 | 11.69 | 26.14 | 23.83 | 8.92 | 5.50 | 12.54 | 7.83 |
| | LoGoNet [49] | 17.65 | 15.39 | 32.55 | 29.19 | 11.66 | 8.20 | 15.09 | 9.77 |
| LiDAR+RGB+Event | Ev-3DOD [22] | **29.53** | **25.07** | **49.31** | **44.44** | 18.42 | 12.49 | **29.06** | **20.34** |
| RGB-Stereo | DSGN [11] | 16.29 | 13.45 | 31.90 | 27.53 | 4.06 | 2.42 | 6.08 | 3.74 |
| | LIGA [34] | 14.26 | 10.98 | 27.25 | 25.11 | 6.02 | 3.58 | 8.73 | 4.89 |
| Event-Stereo | Ours | 23.47 | 19.62 | 40.13 | 33.03 | **19.86** | **12.93** | 22.91 | 14.34 |

# 4. Experiments

## 4.1. Dataset

**DSEC-3DOD Dataset.** The DSEC-3DOD dataset is the real-world event-based 3D detection dataset introduced in the Ev-3DOD [22]. It is based on the DSEC [31], a stereo event dataset for driving scenarios. DSEC-3DOD contains manually created 10 FPS annotations for vehicle and pedestrian classes and provides blind-time annotations at 100 FPS through interpolation and refinement. We used the widely adopted KITTI metric [32] from frame-based stereo methods, which measures performance in both BEV (bird-eye-view) and 3D for each class.

## 4.2. Experiment Setup

**3D Detection during Blind Time.** Consistent with the training and evaluation protocol of existing event-based 3D object detection [22], we train and evaluate using 100 FPS annotations during blind time, when no LiDAR data is available due to the fixed frame rate of LiDAR (*i.e.*, 10 FPS). To achieve this, we sliced the event stream at $\Delta\tau =$ 10 ms and converted it into a voxel grid.

**Motion Scale and Time Slice.** The strength of event cameras lies in their ability to provide continuous information over time, even in dynamic motion [4, 26]. Therefore, experiments on dynamic and long-range motion during blind time are crucial to demonstrate the advantages of the asynchronous setup. However, the DSEC-3DOD dataset mostly lacks such scenes, as its average ego velocity is below 10 m/s. To address this, we introduce an experimental setup, **motion scale** (MS), which scales time axis for large motion experiments that closely resemble the real world. Accumulation and normalization of events over a longer period can approximate events with larger motion over the same time span. Therefore, increasing the motion scale in experiments can depict dynamic environments by skipping consecutive annotations and accumulating events over a longer period. **Time slice** (TS) refers to the number of divisions applied to the blind time, where detection performance is evaluated at each time slice instance. Since various event input durations are used, this setup reflects the real-world demand for asynchronous detection. The baseline setup is defined with a motion scale of 1 and a time slice of 10, which corresponds to the 10 FPS fixed-frame rate data provided in the DSEC-3DOD dataset and the blind time annotations sliced into 10 parts (*i.e.*, 100 FPS GT). To ensure compatibility with the original ground truth, experiments were conducted with motion scale values of 2 and 4 and time slice values of 10 and 20. The model is trained using the original 100 FPS GT with a motion scale of 1 and a time slice of 10 and is evaluated across various motion scales and time slices. More details about the experimental settings are provided in the supplementary material.

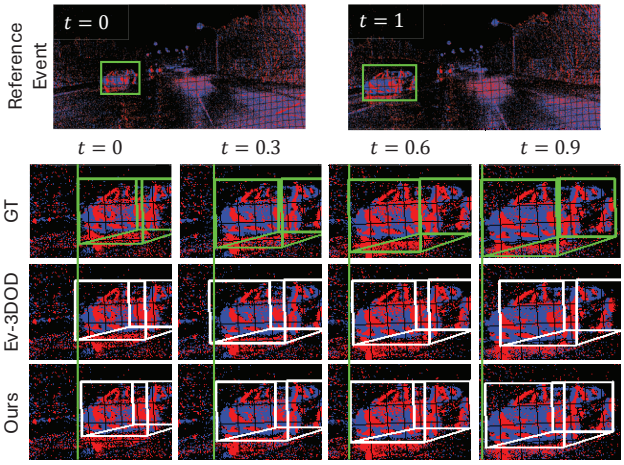


Figure 5. Comparison of 3D detection during blind time. $t = 0$ and $t = 1$ represent active time, while $t = 0.3$, 0.6, and 0.9 indicate normalized blind time intervals.

## 4.3. Results on 3D Detection during Blind Time

Table 1 provides the results of 3D detection performed during blind time. Results from previous works [13, 14, 49,

Table 2. Performance evaluation across various motion scales and time slices, presenting results for the easy difficulty level. Each entry corresponds to 3D / BEV detection results. VEH and PED represent vehicle and pedestrian, respectively.

| Motion Scale | Time Slice | Class | LiDAR | | LIDAR+RGB | | LiDAR+RGB+Event | RGB Stereo | | Event Stereo |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | VoxelNeXt [14] | HEDNet [101] | Focals Conv [13] | LoGoNet [49] | Ev-3DOD [22] | DSGN [11] | LIGA [34] | Ours |
| ×2 | ×10 | VEH | 5.05 / 13.00 | 6.03 / 12.87 | 5.95 / 12.48 | 5.97 / 13.57 | 15.68 / 30.59 | 7.33 / 17.36 | 6.00 / 12.87 | **22.72 / 39.81** |
| | | PED | 4.07 / 4.89 | 3.40 / 4.05 | 3.16 / 4.79 | 2.24 / 3.72 | 7.88 / 12.19 | 1.87 / 3.04 | 2.27 / 4.00 | **19.04 / 22.71** |
| | ×20 | VEH | 4.34 / 11.76 | 5.26 / 11.67 | 5.76 / 11.40 | 5.42 / 12.40 | 15.96 / 31.13 | 7.13 / 16.43 | 5.67 / 12.00 | **23.47 / 40.13** |
| | | PED | 3.83 / 4.31 | 2.60 / 3.49 | 2.66 / 3.86 | 1.91 / 2.90 | 1.47 / 2.63 | 1.84 / 2.99 | 2.08 / 3.01 | **19.86 / 22.91** |
| ×4 | ×10 | VEH | 2.05 / 5.22 | 2.87 / 4.63 | 2.73 / 4.48 | 1.90 / 5.26 | 6.24 / 11.62 | 3.12 / 7.20 | 2.10 / 4.89 | **20.70 / 37.08** |
| | | PED | 2.21 / 2.44 | 1.64 / 1.85 | 1.34 / 1.85 | 1.17 / 1.55 | 3.04 / 4.09 | 0.60 / 0.77 | 1.11 / 1.30 | **16.71 / 21.93** |
| | ×20 | VEH | 1.71 / 4.55 | 2.27 / 4.29 | 2.73 / 3.94 | 1.90 / 4.66 | 6.36 / 11.70 | 3.09 / 6.52 | 1.54 / 4.17 | **22.72 / 39.81** |
| | | PED | 2.14 / 2.17 | 1.10 / 1.42 | 1.02 / 1.44 | 0.53 / 1.28 | 0.39 / 0.43 | 0.59 / 0.75 | 0.82 / 1.12 | **19.04 / 22.71** |

101] were taken from the supplementary material of Ev-3DOD [22], while we trained and evaluated frame-based stereo methods, DSGN [11] and LIGA [34], from scratch. Methods relying solely on synchronized sensors struggle to handle blind time, resulting in poor performance. On the other hand, both Ev-3DOD and our method utilize event cameras to overcome the limitations of fixed frame rates, enabling effective 3D detection during blind time and achieving high performance. While Ev-3DOD benefits from LiDAR-based depth information, generally outperforming our event-only approach, it is noteworthy that our method achieves superior performance in pedestrian $AP_{3D}$. Pedestrian detection requires fine detail, and our high performance on this class is due to specialized modules that effectively leverage semantic features from event data.

Figure 5 illustrates another advantage of the asynchronous setup. While Ev-3DOD shows a decline in qualitative results as it moves further from active time 0, where LiDAR data is available, our method maintains high performance by computing geometric information from events, allowing it to operate independently of active time.

### 4.4. Diverse Motion Scales and Time Slices

Table 2 shows the comparison of ours with different methods across various motion scales and time slices. Methods relying solely on synchronized sensors [11, 13, 14, 34, 49, 101] exhibit a significant performance drop as the motion scale increases during blind time. Our primary comparison method, Ev-3DOD [22], utilizes event data during blind time, allowing it to handle moderate scenarios well, as shown in Table 1. However, as the motion increases, its performance significantly drops. This is because Ev-3DOD relies entirely on LiDAR for 3D geometric information, and when large motion occurs, scene changes make past geometric information less useful for the present. In contrast, our method computes 3D geometric information using asynchronous event data, resulting in more stable performance across different motion scales. Therefore, in scenarios with high-speed driving or significant scene changes due to dynamic motion, our event-based 3D detection proves highly effective during blind time.

Table 3. Ablation study on modules for 3D detection during blind time. DSGF: Dual Semantic-Geometric Filter. SDR: Semantic-Guided Depth Refinement. GSV: Geometric-Filtered Semantic Volume. OCRA: Object-Centric ROI Alignment. The fused voxel source is denoted as G or S, where G represents geometric voxel and S represents semantic voxel.

| DSGF | | 3D Voxel | OCRA | mAP Easy | |
|---|---|---|---|---|---|
| SDR | GSV | Source | | 3D | BEV |
| | | G | | 12.92 | 20.78 |
| | | G+S | | 14.60 | 24.85 |
| | | G | ✓ | 14.56 | 21.73 |
| | | G+S | ✓ | 15.95 | 26.12 |
| ✓ | | G+S | | 15.85 | 27.78 |
| | ✓ | G+S | | 17.08 | 26.53 |
| ✓ | ✓ | G+S | | 19.12 | 29.64 |
| ✓ | ✓ | G+S | ✓ | **21.66** | **31.52** |

## 5. Ablation Study

**Ablation Study on Modules.** As shown in Table 3, we conduct an ablation study on the proposed modules. Starting from a baseline that performs 3D detection using only the geometric plane-sweep volume, we incrementally add the proposed modules to evaluate their impact on performance. We report the easy mAP performance, which is the average value across the two classes, VEH and PED, measured under a motion scale of 1 and a time slice of 10. Compared to the baseline, simply adding semantic information results in sub-optimal performance, yielding only a 1.68 improvement in 3D detection. However, incorporating a dual semantic-geometric filter, which facilitates interaction between semantic and geometric features, leads to a significant performance gain of 6.2 in 3D detection. Additionally, the proposed object-centric ROI alignment consistently improves performance across all settings. When

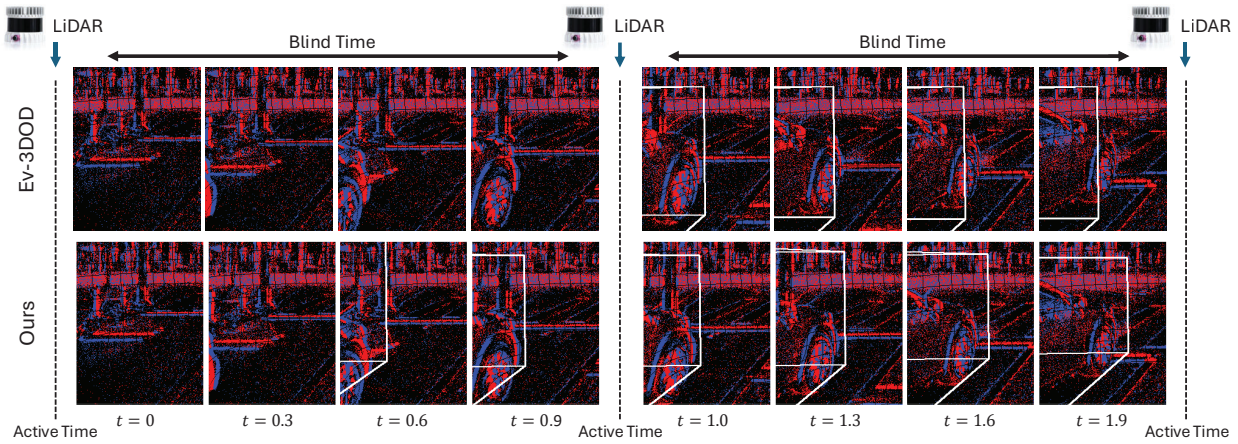


Figure 6. Comparison of 3D detection during blind time. Ev-3DOD detects new objects only at the active timestamps, $t = 0, 1, 2$, using LiDAR and RGB. In contrast, the proposed model can capture the emergence of new objects at arbitrary times.

Table 4. Performance of RGB fusion. Results of easy difficulty level with a time slice of 10.

| Settings | Methods | $AP_{3D}$ | | $AP_{BEV}$ | |
|---|---|---|---|---|---|
| | | VEH | PED | VEH | PED |
| Motion $\times 1$ | Ours (E) | 23.47 | 19.86 | 40.13 | 22.91 |
| | Ours (E+I) | **25.23** | **21.50** | **46.21** | **29.14** |
| Motion $\times 2$ | Ours (E) | **22.72** | **19.04** | **39.81** | **22.71** |
| | Ours (E+I) | 20.14 | 10.50 | 37.93 | 12.76 |

used alongside a dual semantic-geometric filter, it further enhances semantic BEV features, resulting in a more significant performance boost.

**RGB Fusion.** Table 4 presents the results of fusing synchronous RGB information into our method. The stereo RGB features are extracted using a separate feature extractor and concatenated with the event features. This fusion provides improved performance under small motion scales, as the RGB input offers additional semantic guidance. However, under large motion, the performance significantly degrades due to misalignment between the synchronous RGB data and the actual object location. In contrast, our setup that relies solely on event data demonstrates robust performance regardless of scene dynamics.

**Performance Degradation over Blind Time.** Figure 7 shows the performance over time during blind time for our method and other methods [22, 49]. To highlight robustness in challenging scenarios, we use motion scale of 2. $t = 0$ refers to the most recent synchronized sensor input timestamp, while $t = 0.1 \sim 0.9$ represents the blind interval before the next input. The reason for the performance drop of other methods is that they rely on 3D information from synchronous sensors and struggle as the current scene diverges from recent active time ($t = 0$), making it challenging to estimate the present 3D box using past information. Thus, despite using event data, Ev-3DOD struggles to transfer past

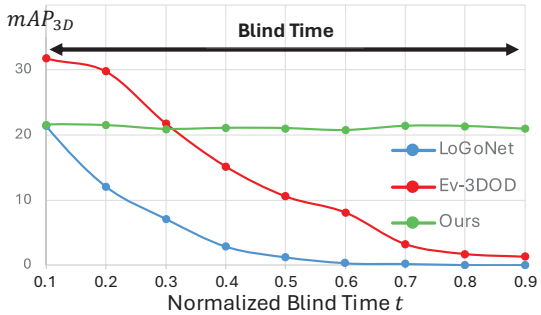


Figure 7. Performance comparison with other methods during blind time under the setting of motion scale 2 and time slice 10.

3D information to the present, causing its performance to resemble LoGoNet as motion increases over time. In contrast, our method asynchronously infers both geometric and semantic information for 3D detection, allowing it to operate independently of active time.

## 6. Conclusion

In this paper, we tackle 3D object detection using event stereo in a fully asynchronous setting for the first time. Unlike the previous continuous-time 3D object detection work that utilizes LiDAR, our approach does not rely on LiDAR, leading to lower performance compared to methods that fuse LiDAR and event data. However, by evaluating various motion scales and time slices, we demonstrate that our method is more adaptable and performs robustly in diverse and extreme conditions. Furthermore, we highlight the potential of event-based 3D perception in dynamic environments where event cameras are particularly beneficial, such as drone flights, high-speed driving, and robots. We hope this work inspires further research in this direction.

## 7. Acknowledgments.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636). This work was also supported by the Technology Innovation Program (1415187329,20024355, Development of autonomous driving connectivity technology based on sensor-infrastructure cooperation) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

## 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*, pages 882–890, 2021. 2

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 2

[3] Luca Bartolomei, Matteo Poggi, Andrea Conti, and Stefano Mattoccia. Lidar-event stereo fusion with hallucinations. In *European Conference on Computer Vision*, pages 125–145. Springer, 2024. 2

[4] Anish Bhattacharya, Marco Cannici, Nishanth Rao, Yuezhan Tao, Vijay Kumar, Nikolai Matni, and Davide Scaramuzza. Monocular event-based vision for obstacle avoidance with a quadrotor. *arXiv preprint arXiv:2411.03303*, 2024. 6

[5] Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 491–501, 2019. 2

[6] Hu Cao, Zehua Zhang, Yan Xia, Xinyi Li, Jiahao Xia, Guang Chen, and Alois Knoll. Embracing events and frames with hierarchical feature refinement network for object detection. In *European Conference on Computer Vision*, pages 161–177. Springer, 2024. 2

[7] 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. 3

[8] Liyan Chen, Weihan Wang, and Philippos Mordohai. Learning the distribution of errors in stereo matching for joint disparity and uncertainty estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17235–17244, 2023. 4

[9] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–181, 2022. 2

[10] 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. 2

[11] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12536–12545, 2020. 2, 3, 5, 6, 7

[12] Yilun Chen, Shijia Huang, Shu Liu, Bei Yu, and Jiaya Jia. Dsgn++: Exploiting visual-spatial relation for stereo-based 3d detectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4416–4429, 2022. 2, 5

[13] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5428–5437, 2022. 6, 7

[14] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 6, 7

[15] Yi-Nan Chen, Hang Dai, and Yong Ding. Pseudo-stereo for monocular 3d object detection in autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 887–897, 2022. 2

[16] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Deformable feature aggregation for dynamic multi-modal 3d object detection. In *European conference on computer vision*, pages 628–644. Springer, 2022. 2

[17] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 454–462, 2022. 2

[18] 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. 2

[19] Hoonhee Cho, Jaeseok Jeong, and Kuk-Jin Yoon. Eomvs: Event-based omnidirectional multi-view stereo. *IEEE Robotics and Automation Letters*, 6(4):6709–6716, 2021. 2

[20] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023.

[21] 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. 2

[22] Hoonhee Cho, Jae-young Kang, Youngho Kim, and Kuk-Jin Yoon. Ev-3dod: Pushing the temporal boundaries of 3d object detection with event cameras. *arXiv preprint arXiv:2502.19630*, 2025. 1, 2, 6, 7, 8

[23] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Pro-*

*ceedings of the AAAI conference on artificial intelligence*, pages 1201–1209, 2021. 5

[24] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8458–8468, 2022. 2

[25] Tuo Feng, Ruijie Quan, Xiaohan Wang, Wenguan Wang, and Yi Yang. Interpretable3d: An ad-hoc interpretable classifier for 3d point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1761–1769, 2024. 2

[26] Benedek Forrai, Takahiro Miki, Daniel Gehrig, Marco Hutter, and Davide Scaramuzza. Event-based agile object catching with a quadrupedal robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12177–12183. IEEE, 2023. 6

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

[28] Daniel Gehrig and Davide Scaramuzza. Pushing the limits of asynchronous graph-based object detection with event cameras. *ArXiv*, abs/2211.12324, 2022. 2

[29] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629:1034 – 1040, 2024. 2

[30] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13884–13893, 2022. 2

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

[32] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 6

[33] Tianrui Guan, Jun Wang, Shiyi Lan, Rohan Chandra, Zuxuan Wu, Larry Davis, and Dinesh Manocha. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 772–782, 2022. 2

[34] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3153–3163, 2021. 2, 3, 5, 6, 7

[35] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22867–22876, 2023. 2

[36] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3d object detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11873–11882, 2020. 2

[37] Chenhang He, Ruihuang Li, Shuai Li, and Lei Zhang. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8417–8427, 2022. 2

[38] Jordan SK Hu, Tianshu Kuai, and Steven L Waslander. Point density-aware voxels for lidar 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8469–8478, 2022. 2

[39] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[40] Junjie Huang, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. In *European Conference on Computer Vision*, pages 439–455. Springer, 2025. 2

[41] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021. 2

[42] Massimiliano Iacono, Stefan Weber, Arren J. Glover, and Chiara Bartolozzi. Towards event-driven object detection with off-the-shelf deep learning. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9, 2018. 2

[43] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 1042–1050, 2023. 2

[44] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, and Peter Henry. End-to-end learning of geometry and context for deep stereo regression. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. 5

[45] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[46] Dianze Li, Jianing Li, and Yonghong Tian. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:14020–14037, 2023. 2

[47] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022. 2

[48] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 2

[49] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023. 2, 6, 7, 8

[50] Xiaotian Li, Baojie Fan, Jiandong Tian, and Huijie Fan. Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21209–21218, 2024. 2

[51] Xiang Li, Junbo Yin, Wei Li, Chengzhong Xu, Ruigang Yang, and Jianbing Shen. Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3215, 2024. 2

[52] Yijin Li, Han Zhou, Bangbang Yang, Yexin Zhang, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Graph-based asynchronous event processing for rapid object recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 914–923, 2021. 2

[53] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022. 2

[54] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1120–1129, 2022.

[55] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17182–17191, 2022. 2

[56] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2

[57] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 2

[58] Zhiwei Lin, Zhe Liu, Zhongyu Xia, Xinhao Wang, Yongtao Wang, Shengxiang Qi, Yang Dong, Nan Dong, Le Zhang, and Ce Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14928–14937, 2024. 2

[59] Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. Multi-view attentive contextualization for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16688–16698, 2024. 2

[60] Yuxuan Liu, Lujia Wang, and Ming Liu. Yolostereo3d: A step back to 2d for efficient stereo 3d detection. In *2021 IEEE international conference on Robotics and automation (ICRA)*, pages 13018–13024. IEEE, 2021. 2

[61] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 2

[62] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 2

[63] Hanyue Lou, Jinxiu Sherry 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, 2025. 2

[64] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 2

[65] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3164–3173, 2021. 2

[66] Zhenwei Miao, Jikai Chen, Hongyu Pan, Ruiwen Zhang, Kaixuan Liu, Peihan Hao, Jun Zhu, Yang Wang, and Xin Zhan. Pvgnet: A bottom-up one-stage 3d object detector with integrated multi-level features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3279–3288, 2021. 2

[67] 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. 2

[68] 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 Patter Recognition*, 2022. 2

[69] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10386–10393. IEEE, 2020. 2

[70] Yan Peng, Yueyi Zhang, Peilin Xiao, Xiaoyan Sun, and Feng Wu. Better and faster: Adaptive event conversion for event-based object detection. In *AAAI Conference on Artificial Intelligence*, 2023. 2

[71] Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-

based vision. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6015–6025, 2023. 2

[72] Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based object detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16794–16804, 2024. 2

[73] Etienne Perot, Pierre de Tournemire, Davide Oscar Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *ArXiv*, abs/2009.13436, 2020. 2

[74] AJ Piergiovanni, Vincent Casser, Michael S Ryoo, and Anelia Angelova. 4d-net for learned multi-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15435–15445, 2021. 2

[75] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5293–5313, 2021. 4

[76] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021. 2

[77] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[78] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4641–4650, 2017. 4

[79] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3d object detection with channel-wise transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2743–2752, 2021. 2

[80] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[81] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 2

[82] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1711–1719, 2020. 2

[83] Ziying Song, Lei Yang, Shaoqing Xu, Lin Liu, Dongyang Xu, Caiyan Jia, Feiyang Jia, and Li Wang. Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection. *arXiv preprint arXiv:2403.11848*, 2024. 2

[84] Jiaming Sun, Linghao Chen, Yiming Xie, Siyu Zhang, Qinhong Jiang, Xiaowei Zhou, and Hujun Bao. Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10548–10557, 2020. 2

[85] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 2

[86] Dongsheng Wang, Xu Jia, Yang Zhang, Xinyu Zhang, Yaoyuan Wang, Ziyang Zhang, D. Wang, and Huchuan Lu. Dual memory aggregation network for event-based object detection with learnable representation. In *AAAI Conference on Artificial Intelligence*, 2023. 2

[87] Wenjie Wang, Yehao Lu, Guangcong Zheng, Shuigen Zhan, Xiaoqing Ye, Zichang Tan, Jingdong Wang, Gaoang Wang, and Xi Li. Bevspread: Spread voxel pooling for bird's-eye-view representation in vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14718–14727, 2024. 2

[88] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[89] Yan Wang, Junbo Yin, Wei Li, Pascal Frossard, Ruigang Yang, and Jianbing Shen. Ssda3d: Semi-supervised domain adaptation for 3d object detection from point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2707–2715, 2023. 2

[90] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17545–17556, 2023. 2

[91] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 4

[92] Jianyun Xu, Zhenwei Miao, Da Zhang, Hongyu Pan, Kaixuan Liu, Peihan Hao, Jun Zhu, Zhengyang Sun, Hongmin Li, and Xin Zhan. Int: Towards infinite-frames 3d detection with an efficient framework. In *European Conference on Computer Vision*, pages 193–209. Springer, 2022. 2

[93] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3047–3054. IEEE, 2021. 2

[94] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1951–1960, 2019. 2

[95] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11040–11048, 2020. 2

[96] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022. 2

[97] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhao-Xu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10201–10210, 2021. 2

[98] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *European conference on computer vision*, pages 17–33. Springer, 2022. 2

[99] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14905–14915, 2024. 2

[100] Gang Zhang, Junnan Chen, Guohuan Gao, Jianmin Li, Si Liu, and Xiaolin Hu. Safdnet: A simple and effective network for fully sparse 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14477–14486, 2024. 2

[101] Gang Zhang, Chen Junnan, Guohuan Gao, Jianmin Li, and Xiaolin Hu. Hednet: A hierarchical encoder-decoder network for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

[102] 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. 2

[103] Fengan Zhao, Qianang Zhou, and Junlin Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In *European Conference on Computer Vision*, pages 190–205. Springer, 2024. 2

[104] Dingfu Zhou, Jin Fang, Xibin Song, Liu Liu, Junbo Yin, Yuchao Dai, Hongdong Li, and Ruigang Yang. Joint 3d instance segmentation and object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1839–1849, 2020. 2

[105] Alex Zihao Zhu, Yibo Chen, and Kostas Daniilidis. Real-time time synchronized event-based stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 433–447, 2018. 2

[106] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3

[107] Dongqing Zou, Feng Shi, Weiheng Liu, Jia Li, Qiang Wang, Paul-KJ Park, Chang-Woo Shi, Yohan J Roh, and Hyunsurk Eric Ryu. Robust dense depth map estimation from sparse dvs stereos. In *British Mach. Vis. Conf.(BMVC)*, 2017. 2

[108] Nikola Zubic, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12800–12810, 2023. 2