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

Supplementary Material

1. Details of Motion Scale and Time Slice

Figure 1 provides details of our experimental setup. To ensure a fair comparison with previous works, Sec. 4.3 of the main paper adopts the same evaluation protocol. DSEC-3DOD provides a fixed-frame-rate sensor at 10 FPS and blind time annotations at 100 FPS. Provided LiDAR and RGB data are fully utilized, and each blind time is evaluated with 10 ground truth annotations.

The evaluation setup of Ev-3DOD [5] allows for assessing detection performance during blind time. However, due to the limited motion in DSEC data and its fixed time intervals, model performance can only be evaluated at restricted points in time. To enable evaluation under asynchronous and diverse temporal conditions, we define motion scale and time slice as key evaluation setup parameters.

Motion scale is a parameter that controls scene motion by adjusting the length of the blind time. This is achieved by skipping consecutive frames of LiDAR and RGB data, thereby modifying the amount of motion. Events are accumulated over the blind time and normalized in the temporal domain. Motion scale control [4, 8, 11, 12, 14, 15], which accumulates data over a longer period to represent large motion, is a well-established method widely used in other works for evaluating performance under large motion conditions. Therefore, following previous work, we also adopted motion scale control to represent dynamic and long-range motion.

Time slice controls the evaluation interval within the blind time. Each blind time is evaluated at multiple points determined by the time slice. This parameter introduces variations in the distribution of event data, making it a challenging factor for assessing the temporal flexibility of event-based methods.

We define the baseline experimental setup with a motion scale of 1 and a time slice of 10. Evaluations were conducted using various experimental parameters within the constraints of the given fixed-frame-rate sensor and available annotations. To ensure a fair evaluation, we assessed detection performance using only the model trained on the baseline setup.

2. Implementation Details

Training Details. Training was conducted on two NVIDIA TITAN RTX GPUs for 60 epochs with a batch size of 2. The AdamW optimizer [10] was employed with a learning rate set to 0.001.

Depth Refinement. To perform actual refinement, we computed a finer probability by considering neighboring probabilities. Thus, the practical implementation of Eq. (6) in the main paper incorporates neighboring pixels as follows:

$$S(u,v) = \langle F_L^{sem}(u,v), F_R^{sem}(u - \frac{fL}{D_{init}^m(u,v)}, v)\rangle, \quad (1)$$

We use the $m = 1, 2, 3, 4, 5$ for neighboring sampling.

Event Grid size and Voxel Size. Following previous works [5], we use the bin size of event voxel grid as 5. 3D geometric voxel and 3D semantic voxel in Sec. 3.3 have range of $[-30.4m, 30.4m]$ in X axis, $[-1.0m, 3.0m]$ in Y axis, and $[2.0m, 56.9m]$ for Z axis. Voxel size is set to $(0.2m, 0.2m, 0.2m)$.

ROI Pooling for Alignment In Sec. 3.5 of the main paper, the ROI P_G estimated by the global detector is divided into a $k \times k$ voxel grid, for $k = 3$. The semantic BEV features are pooled for each grid, and all grid features are aggregated to estimate the local offset.

Anchor Size As mentioned in the main paper, we use anchors with fixed size, height and orientation for each (x, z) coordinate in the 3D voxel space. The fixed anchor sizes are determined by computing the class-wise box statistics from the training set. Anchors for vehicle class and pedestrian class are as follows:

$$\begin{aligned} A_{veh} &= (x, 0.47, z, 1.79, 1.86, 4.28, \{0, \frac{\pi}{2}\}) \\ A_{ped} &= (x, 0.6, z, 1.73, 0.6, 0.8, \{0, \frac{\pi}{2}\}) \end{aligned} \quad (2)$$

3. Effectiveness of Semantic-guided Depth Refinement

Following the KITTI stereo metric [6], we measure a depth estimation accuracy if depth error is below a specified outlier threshold. Table 1 compares performance across different outlier thresholds, highlighting the impact of semantic-guided depth refinement (SDR). The results show that SDR consistently enhances accuracy across all thresholds. Moreover, the depth refinement module improves not only the final detection performance but also the overall depth estimation quality.

4. Visualization on Semantic and Geometric Features

Semantic and geometric event features serve distinct roles, and the model utilizes them collaboratively to enhance both

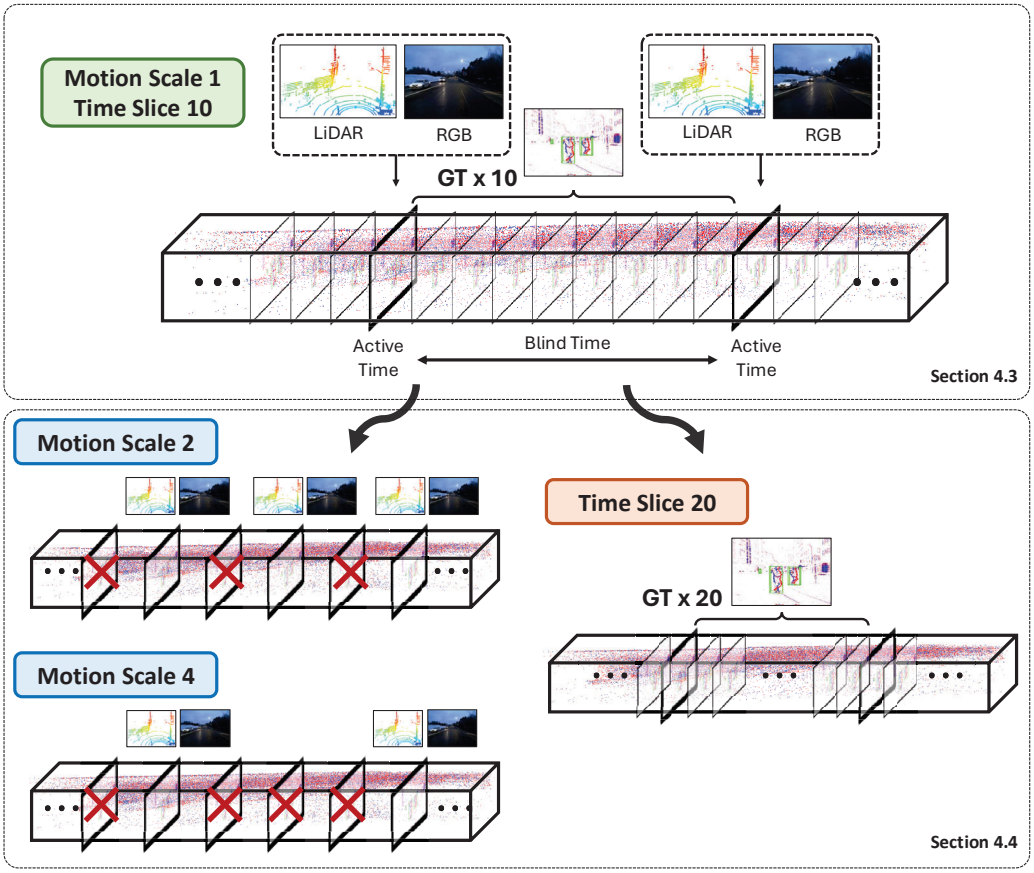

Figure 1. Visualization of the motion scale and time slice used in the experimental setup. The upper part of the figure represents the baseline setup following Ev-3DOD [5], with a motion scale of 1 and a time slice of 10. The lower part illustrates the setup adopted to evaluate the model under large motion and diverse event inputs.

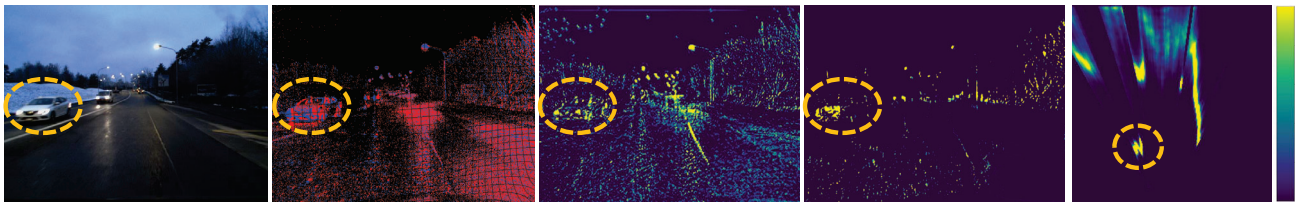

Figure 2. Example of input event, geometric feature, enhanced semantic feature, and semantic BEV feature. Feature values are normalized to $[0, 1]$ for visualization.

Table 1. Effectiveness of semantic-guided depth refinement. SDR: Semantic-Guided Depth Refinement.

Outlier Threshold	$> 1.6m$	$> 0.8m$	$> 0.4m$	$> 0.2m$
w/o SDR	0.236	0.382	0.549	0.715
w/ SDR (Ours)	**0.219**	**0.361**	**0.529**	**0.696**

geometric and semantic information. To provide insights into the characteristics of these features, we present visualizations of the extracted representations. Geometric features highlight complex structures that aid in stereo matching, whereas semantic features exhibit strong attention to target objects. By leveraging such object-centric information, ROI alignment is achieved, enabling fine-grained box regression.

5. Additional Results

Quantitative Results. Table 2 presents the quantitative results of DSEC-3DOD at the moderate difficulty level, evaluated under various motion scales and time slices. Compared to our method, other approaches exhibit more significant performance degradation under large motions and longer

Table 2. Performance evaluation across various motion scales and time slices, presenting results for the moderate 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 [3]	HEDNet [13]	Focals Conv [2]	LoGoNet [9]	Ev-3DOD [5]	DSGN [1]	LIGA [7]	Ours
×2	×10	VEH	4.28 / 12.58	5.24 / 12.06	5.60 / 11.66	4.93 / 12.21	13.52 / 26.56	7.02 / 16.03	5.52 / 11.93	**19.31 / 32.47**
		PED	2.43 / 3.13	1.70 / 2.48	2.14 / 2.81	1.71 / 2.62	4.91 / 8.57	1.27 / 1.88	1.75 / 2.29	**12.56 / 13.99**
	×20	VEH	3.80 / 11.54	4.67 / 11.08	5.56 / 10.79	4.60 / 11.23	14.50 / 27.93	6.75 / 14.36	4.77 / 11.08	**19.62 / 33.03**
		PED	2.31 / 2.48	1.49 / 2.18	1.52 / 2.15	1.44 / 2.24	1.62 / 3.19	1.26 / 1.77	1.42 / 2.12	**12.93 / 14.34**
×4	×10	VEH	2.03 / 5.22	2.85 / 4.78	2.73 / 3.96	1.88 / 4.37	5.59 / 10.50	3.08 / 6.27	1.82 / 4.50	**16.42 / 29.01**
		PED	0.91 / 1.18	0.91 / 0.91	0.91 / 1.06	0.91 / 0.91	2.31 / 3.37	0.41 / 0.49	0.91 / 1.05	**10.61 / 13.86**
	×20	VEH	1.59 / 4.29	2.27 / 4.09	2.73 / 3.55	1.88 / 3.99	5.60 / 10.26	3.06 / 5.62	1.36 / 3.87	**19.31 / 32.47**
		PED	0.91 / 0.91	0.91 / 0.91	0.91 / 0.91	0.45 / 0.91	0.91 / 1.31	0.34 / 0.46	0.45 / 0.69	**12.93 / 14.34**

blind times, as they heavily rely on synchronized sensors (*e.g.*, RGB and LiDAR).

Qualitative Results.

We provide additional qualitative results for the motion scale 2 and time slice 10 setup. The results demonstrate that conventional sensor-based methods suffer from significant detection errors due to large motion. Furthermore, compared to Fig. 5 in the main paper, where the motion scale is set to 10, Ev-3DOD exhibits a substantial performance drop despite utilizing event data, as it remains heavily dependent on LiDAR. In contrast, our fully asynchronous model seamlessly adapts to large motion, ensuring robust detection.

References

[1] 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. 3, 4

[2] 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. 3

[3] 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. 3

[4] 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. 1

[5] 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, 3, 4

[6] 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. 1

[7] 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. 3

[8] Jesse Hagenaars, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. *Advances in Neural Information Processing Systems*, 34:7167–7179, 2021. 1

[9] 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. 3, 4

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022. 1

[12] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 1

[13] 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. 3

[14] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. 1

[15] 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. 1

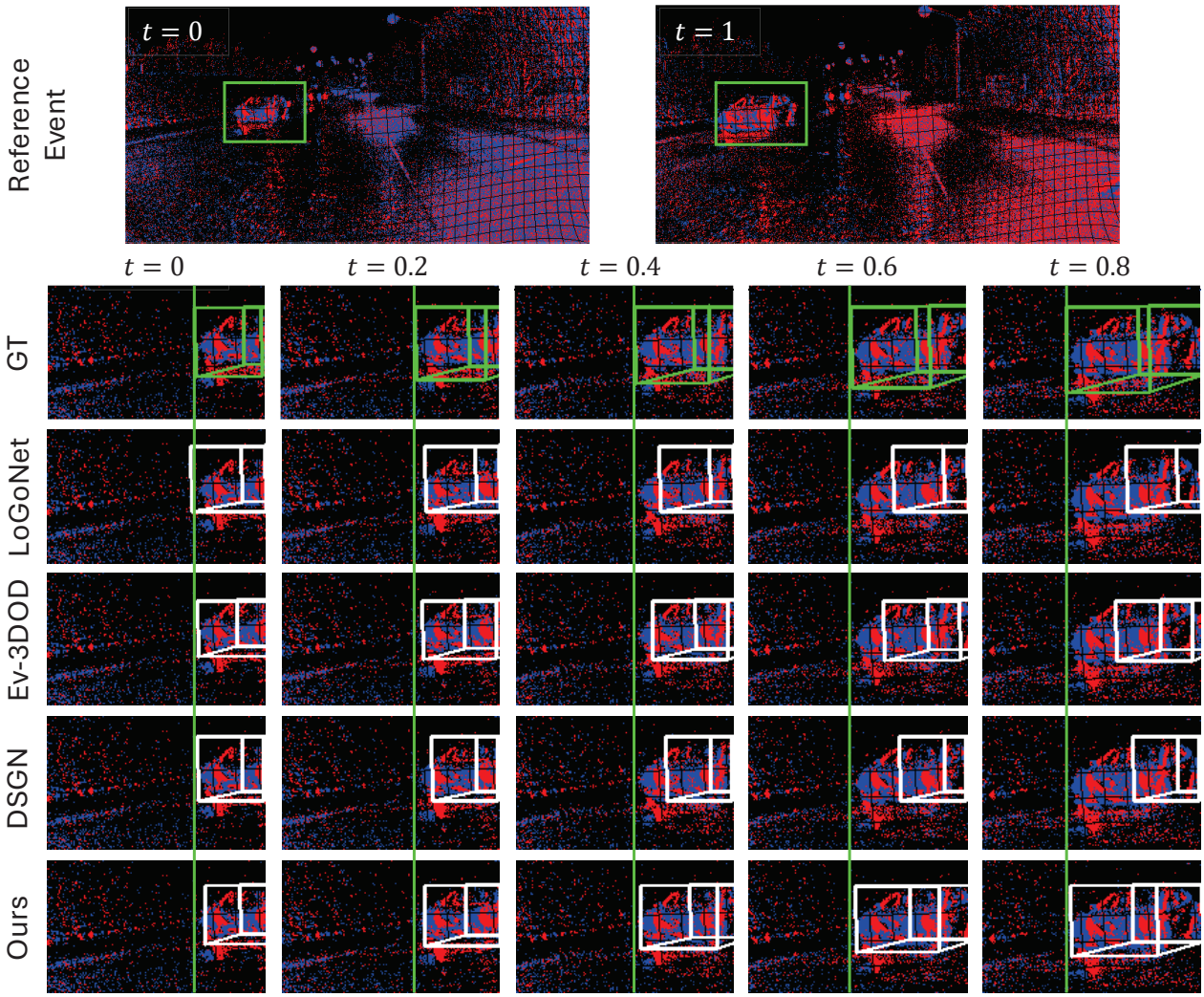

Figure 3. Comparison of 3D detection during blind time. The motion scale and time slice are set to 2 and 10, respectively. Green vertical lines across the image were added to compare the box's relative position. Fixed-frame-rate sensor-based methods (*i.e.*, LoGoNet [9] and DSGN [1]) fail to predict objects during the blind time. Ev-3DOD [5] leverages monocular event data to propagate detection through blind time, but its performance deteriorates under large movements. The proposed method operates in a fully asynchronous manner, consistently producing stable results regardless of the blind time.