Scaling Dense Event-Stream Pretraining from Visual Foundation Models

Supplementary Material

In this appendix, we supplement the following materials to support the findings and observations in the main body of this paper:

- Section 1 elaborates on detailed implementation specifics to facilitate reproduction;
- Section 2 presents the complete quantitative results of our experiments;
- Section 3 includes extensive qualitative results to indicate clearer visual comparisons;
- Section 4 provides a further analysis of the current limitations and discusses potential improvement methods.

1. Additional Implementation Detail

1.1. Pretraining Datasets

In this work, we assemble an extensive collection of synchronized image-event datasets to pretrain a versatile and reliable event-domain feature encoder. These datasets span diverse sensing conditions, motion patterns, environments, and acquisition pipelines, providing broad coverage for large-scale cross-modal alignment. A summary of the detailed configurations and salient characteristics of these pretrained datasets is shown in Table 1 and Table 2, grouped by real-world and synthetic sources.

1.2. Vision Foundation Models

In this work, we adopt the state-of-the-art visual foundation model DINOv3 [23] as the teacher model to distill fine-grained representations into our event encoder. Before committing to this choice, we conducted a brief comparative analysis of representative VFMs: CLIP [20], DINOv2 [18], SAM [10], SEEM [37], RADIO2.5 [8], OpenSeeD [33], DINOV [12], GLEE [30], and DINOv3, with emphasis on fine-grained representation fidelity (token-level affinities, boundary sharpness, and global-local coherence). Using a controlled toy example (Figure 1), we probed the quality of the learned semantic structure. DINOv3 consistently exhibited the most coherent long-range grouping and the clearest region boundaries, and is therefore selected as our teacher model. Supporting qualitative results are reported in [23].

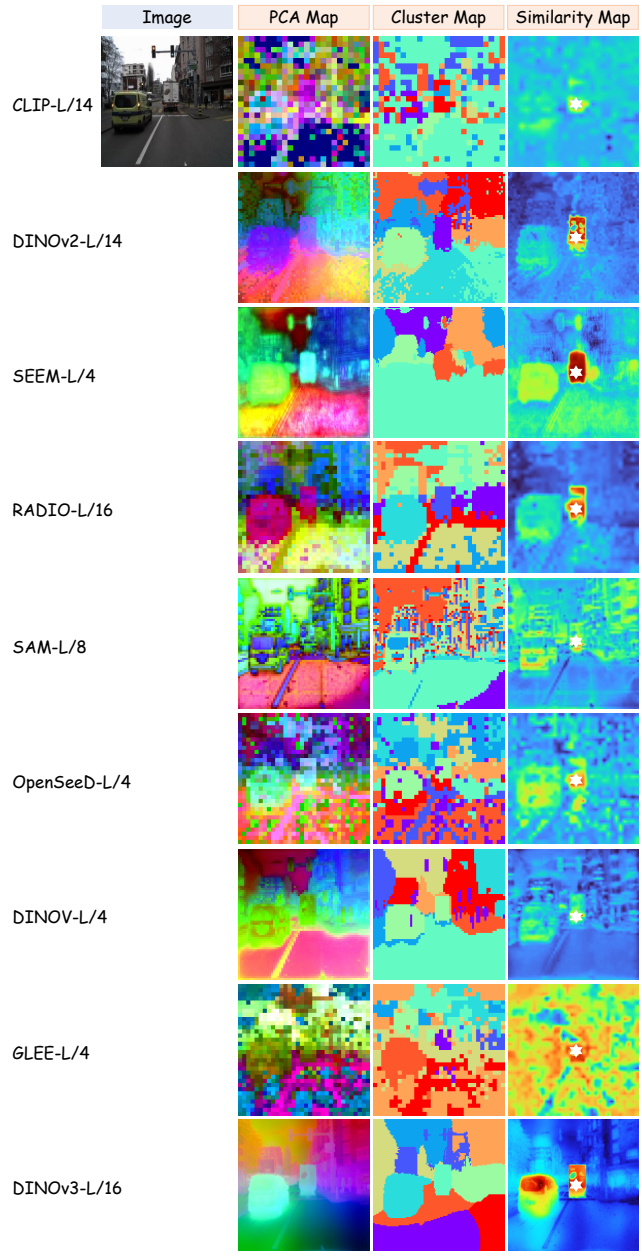


Figure 1. **Comparison of dense image features** under different visual foundation models through a toy example.

1.3. Downstream Datasets

Semantic Segmentation. Following prior works [13, 14, 34], we evaluate event-based semantic segmentation on the DDD17-Seg [1] and DSEC-Semantic [26] datasets.

(i) DDD17-Seg: DDD17-Seg [1] is a semantic segmentation extension of the DDD17 [3] dataset. Alonso and Murillo [1] overlay semantic masks on by leveraging co-registered gray-scale frames with event streams to synthesize approximate labels, which proved effective for training models that segment directly on event data. The dataset provides $15,950$ training and $3,890$ test samples, with semantic maps at 352×200 resolution. Each pixel is annotated with one of six classes: *flat, background, object, vegetation, human, and vehicle*. Download.

(ii) DSEC-Semantic: DSEC-Semantic [26] is a seman-

Table 1. The pretraining dataset configuration and data statistics for the **nine real-world event-image datasets** used in our experiments.

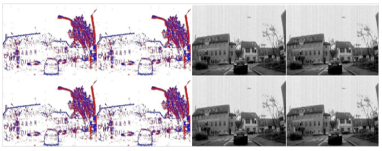
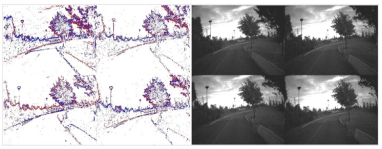
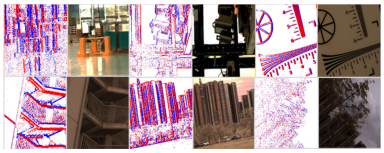
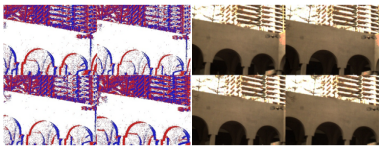
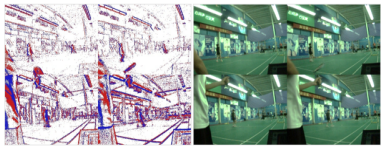
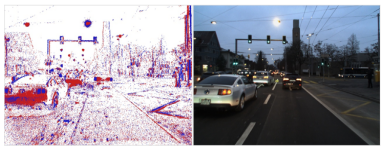
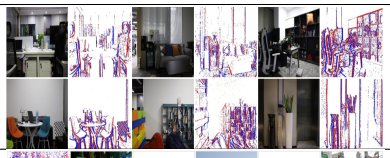
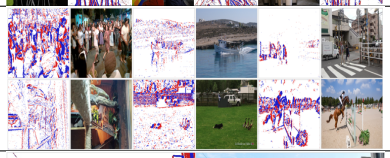
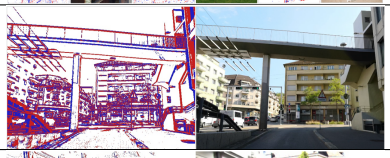
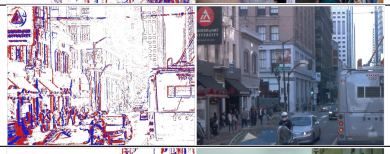
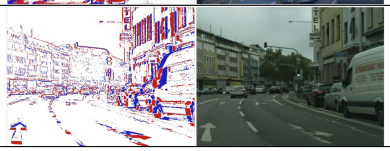
Dataset	Illustration	Resolution	Statistics	Source&Type
DDD17 [3]		346 × 260	5,000 pairs ≈ 20 categories 36 sequences	Real-world DAVIS346B Low-resolution Driving Scene Download
MVSEC [35]		346 × 260	5,000 pairs ≈ 20 categories 9 sequences	Real-world DAVIS346B Low-resolution Driving Scene Download
SEE-600K [16]		346 × 260	5,000 pairs ≈ 20 categories 16 sequences	Real-world DAVIS346C Low-resolution Daliy Scene Download
VisEvent [29]		346 × 260	30,000 pairs ≈ 80 categories 820 sequences	Real-world DAVIS346C Low-resolution Daliy Scene Download
CoeSot [27]		346 × 260	30,000 pairs ≈ 90 categories 1343 sequences	Real-world DAVIS346C Low-resolution Daliy Scene Download
DSEC [6]		640 × 480	20,000 pairs ≈ 40 categories 53 sequences	Real-world Prophesee Gen3.1 High-resolution Driving Scene Download
FEVD [9]		1024 × 768	5,000 pairs ≈ 20 categories 21 sequences	Real-world Prophesee Gen4 High-resolution Daliy Scene Download
M3ED [4]		1280 × 720	20,000 pairs ≈ 40 categories 57 sequences	Real-world Prophesee Gen4 High-resolution Multiple Platforms Download
HighREV [24]		1632 × 1224	10,000 pairs ≈ 20 categories 25 sequences	Real-world High-resolution Multi-modality Daliy Scene Download

Table 2. The pretraining dataset configuration and data statistics for the **seven synthetic event-image datasets** used in our experiments.

Dataset	Illustration	Resolution	Statistics	Source&Type
SDSD [28]		346 × 260	20,000 pairs ≈ 50 categories 150 sequences	VID2E Simulation Low-resolution Daliy Scene Download
DAVIS17 [19]		346 × 260	20,000 pairs ≈ 100 categories 90 sequences	VID2E Simulation Low-resolution Motion Scene Download
DECD [21]		640 × 480	40,000 pairs ≈ 40 categories 120 sequences	VID2E Simulation High-resolution Driving Scene Download
KITTI [7]		1242 × 375	30,000 pairs ≈ 40 categories 60 sequences	VID2E Simulation High-resolution Driving Scene Download
GoPro [17]		1280 × 720	10,000 pairs ≈ 30 categories 35 sequences	VID2E Simulation High-resolution Daliy Scene Download
Waymo [25]		1920 × 1280	50,000 pairs ≈ 40 categories 147 sequences	VID2E Simulation High-resolution Driving Scene Download
Cityscapes [5]		2048 × 1024	200,000 pairs ≈ 40 categories 10000 sequences	VID2E Simulation High-resolution Driving Scene Download

tic segmentation extension of the DSEC [6] dataset. Leveraging DSEC's synchronized, high-resolution RGB images and event streams across diverse driving conditions, Sun et al. [26] applied a pseudo-labeling procedure akin to DDD17-Seg[1] to generate semantic masks for eleven sequences (11/53), yielding the DSEC-Semantic benchmark. The dataset provides $8,082$ training and $2,809$ test samples, with semantic maps at 640×440 resolution. Each pixel is annotated with one of eleven classes: *background, building, fence, person, pole, road, sidewalk, vegetation, car, wall, and traffic-sign*. Download.

Depth Estimation. Following the setup of prior works [2, 15], we evaluate on the MVSEC-Depth [35] and DSEC-Depth datasets [6] for event-based monocular depth estimation.

(i) MVSEC-Depth: MVSEC-Depth is a depth estimation variant of the MVSEC [35] dataset. The dataset provides events at a resolution 346×260 pixels from a stereo event camera consisting of two DAVIS346B sensors. The depth ground-truth is derived from a 16-line LiDAR using Lidar Odometry and Mapping (LOAM), yielding a total of $10,351$ training samples and $21,125$ testing samples. The test set is divided into a $5k$-sample daytime subset and three night-time subsets, each containing $5k$ samples. Download.

(ii) DSEC-Depth: DSEC-Depth is a depth estimation variant of the DSEC [6] dataset. DSEC employs two Prophesee Gen3.1 event cameras in a stereo configuration. The disparity ground-truth is obtained using a 32-beam LiDAR, processed with a Lidar Inertial Odometry algorithm, and further filtered to remove outliers. We convert the disparity ground-truth to depth map based on the stereo setup parameters. The dataset provides $19,181$ training and $7,157$ test samples, with depth maps at 640×480 resolution. Download.

Table 3. **Experimental setup for fine-tuning downstream tasks. lr** denotes learning rate. All configurations are based on the ViT-L encoder. Apart from batch size, which depends on model scale, all other settings remain identical across experiments.

Dataset	Semantic Segmentation		Depth Estimation		Flow Estimation
	DDD17-Seg	DSEC-Semantic	MVSEC-Depth	DSEC-Depth	MVSEC-Flow
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
encoder lr	2×10^{-6}	2×10^{-6}	2×10^{-6}	1×10^{-6}	1×10^{-6}
decoder lr	5×10^{-6}	4×10^{-6}	4×10^{-6}	2×10^{-6}	2×10^{-6}
weight decay	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
batch size	40	12	40	12	24
epochs	20	30	30	30	20
lr scheduler	exponential	exponential	exponential	exponential	exponential
scheduler gamma	0.9	0.9	0.9	0.9	0.9
scheduler epoch	5	5	5	5	5
gradient clipping norm	0.1	0.1	0.1	0.1	0.1

Optical Flow Estimation. Following the setup of prior works [31, 32], we evaluate event-based optical flow estimation on the MVSEC-Flow [36] dataset. **MVSEC-Flow** is an optical flow estimation variant of the MVSEC [35] dataset. MVSEC employs two DAVIS346B event cameras in a stereo configuration. MVSEC-Flow provides per-camera poses and depth maps for each event camera, which were used to generate ground truth optical flow. In this work, we use *outdoor_day2* sequence for training ($26,677$ samples), *indoor_flying1, indoor_flying2, indoor_flying3* sequences for testing ($7,775$ samples). Download.

1.4. Downstream Fine-tuning

- **Experimental Setup.** The details of the fine-tuning procedure are outlined in Table 3.
- **Data Augmentation.** No data augmentation strategies are applied during fine-tuning on downstream tasks.
- **Linear Probing.** The pretrained event feature encoder is frozen with a trainable pixel-wise task head which is trained for 20 epochs, setting the initial learning rate at 5×10^{-4}, with a weight decay of 1×10^{-4}.
- **Few-shot Fine-tuning.** In few-shot fine-tuning, we subsample the training split of each downstream dataset to obtain 1%, 5%, 10%, or 20% annotated scans, generated via fixed-interval sampling over the full training sequences, such as 100, 20, 10, 5.

2. More Quantitative Results

2.1. More Detailed Comparisons

We report the complete results (i.e., the class-wise IoU scores, optical flow/depth metrics) for the **inear probing** and **downstream fine-tuning tasks** outlined in the main paper. Specifically, the detailed performance metrics on the DDD17-Seg, DSEC-Semantic, MVSEC-Depth, DSEC-Depth, and MVSEC-Flow datasets are shown in Table 4, Table 6, Table 7 and Table 5, respectively. These results comprehensively evaluate the model's performance across a variety of dense perception tasks.

Table 4. The per-class segmentation results of our methods on the DDD17-Seg dataset. Scores reported are IoUs in percentage.

Event Model	**mIoU**	flat	background	object	vegetation	human	vehicle	**Acc**
Linear Probing								
ViT-S/16	55.64	79.61	91.18	15.90	57.51	22.02	67.72	91.27
ViT-B/16	57.87	79.92	91.24	15.87	58.04	34.97	67.05	91.31
ViT-L/16	60.30	81.03	91.49	18.83	57.21	44.18	68.95	91.83
Fine-Tuning (1%)								
ViT-S/16	53.87	78.61	90.06	10.03	54.59	25.18	64.63	90.41
ViT-B/16	57.23	79.51	91.03	15.46	57.53	34.87	66.93	91.12
ViT-L/16	59.23	82.34	92.24	18.26	61.68	34.01	69.37	91.68
Fine-Tuning (5%)								
ViT-S/16	54.36	78.96	90.27	10.38	54.92	27.26	64.40	90.62
ViT-B/16	59.54	80.25	91.65	15.24	59.21	44.38	65.80	91.65
ViT-L/16	62.52	81.96	91.97	19.31	61.49	50.77	69.63	92.12
Fine-Tuning (10%)								
ViT-S/16	57.29	79.52	91.16	12.24	59.26	39.77	64.72	91.34
ViT-B/16	61.45	82.37	91.69	20.14	60.69	45.16	68.46	91.72
ViT-L/16	63.71	82.23	92.12	23.75	59.84	51.80	72.95	92.13
Fine-Tuning (20%)								
ViT-S/16	58.37	79.93	91.55	13.02	58.93	41.81	66.27	91.63
ViT-B/16	62.06	82.74	92.05	18.72	61.65	49.79	69.60	92.24
ViT-L/16	64.43	83.10	92.23	23.17	62.62	54.13	71.35	92.44
Fine-Tuning (100%)								
ViT-S/16	59.64	80.68	91.27	17.58	58.88	43.71	65.73	91.39
ViT-B/16	62.81	82.95	92.00	18.79	61.71	51.43	69.98	92.21
ViT-L/16	65.09	83.73	92.34	23.10	62.61	56.43	72.26	92.62

Table 5. The optical flow results of our methods on the MVSEC-Flow dataset.

Event Model	indoor flying1		indoor flying2		indoor flying3	
	EPE ↓	Out↓	EPE ↓	Out↓	EPE ↓	Out↓
ViT-S/16	0.29	0.03	0.38	0.001	0.40	0.001
ViT-B/16	0.28	0.03	0.38	0.001	0.39	0.001
ViT-L/16	0.27	0.03	0.37	0.001	0.39	0.001

Table 6. The per-class segmentation results of our methods on the DSEC-Semantic dataset. Scores reported are IoUs in percentage.

Event Model	**mIoU**	background	building	fence	person	pole	road	sidewalk	vegetation	car	wall	traffic-sign	**Acc**
Linear Probing													
ViT-S/16	55.46	92.81	81.88	17.45	15.67	24.98	93.20	68.12	78.83	77.72	30.79	43.86	90.12
ViT-B/16	58.42	93.46	83.52	23.88	16.69	27.85	93.72	69.27	80.38	80.13	43.06	43.25	91.44
ViT-L/16	61.29	93.91	85.09	27.66	27.37	33.58	93.34	70.94	82.37	82.27	41.82	48.66	91.69
Fine-Tuning (1%)													
ViT-S/16	52.97	92.35	81.36	18.04	7.93	18.77	92.55	60.54	78.50	76.22	22.25	37.06	89.56
ViT-B/16	54.37	93.04	82.55	14.09	16.14	26.06	93.55	65.17	80.79	79.52	12.34	41.40	90.14
ViT-L/16	59.73	92.96	82.55	21.88	20.30	27.87	93.34	69.36	80.68	80.22	40.60	47.24	90.73
Fine-Tuning (5%)													
ViT-S/16	56.55	93.01	82.08	19.92	18.34	19.48	93.15	66.34	79.36	79.22	33.71	46.98	90.78
ViT-B/16	62.87	93.66	84.91	22.14	33.63	31.05	93.91	71.36	81.84	82.27	44.28	49.59	91.52
ViT-L/16	68.03	94.68	86.68	31.01	52.67	39.86	94.77	74.13	84.08	83.89	49.93	58.39	92.83
Fine-Tuning (10%)													
ViT-S/16	58.96	93.56	84.73	21.88	19.25	22.50	93.26	68.90	79.52	78.61	42.29	42.48	91.25
ViT-B/16	63.88	93.59	85.06	23.05	38.58	31.72	94.05	71.48	81.85	82.38	48.44	49.10	91.73
ViT-L/16	68.51	94.52	86.76	26.70	51.15	44.29	94.85	75.24	84.49	84.92	47.71	62.72	92.92
Fine-Tuning (20%)													
ViT-S/16	60.20	93.32	83.99	22.61	27.38	29.98	93.07	69.40	79.62	80.29	33.19	48.69	91.62
ViT-B/16	64.15	93.45	84.93	24.27	40.01	32.64	93.91	71.77	82.13	82.55	50.67	50.30	91.78
ViT-L/16	69.25	94.63	86.77	26.98	51.24	44.27	94.90	75.45	84.74	84.93	47.90	62.95	92.98
Fine-Tuning (100%)													
ViT-S/16	61.12	93.16	83.16	26.28	34.11	30.15	93.12	68.21	80.10	80.04	34.89	49.03	90.76
ViT-B/16	64.93	93.43	85.17	23.69	42.63	34.14	94.08	72.84	82.28	83.13	51.86	50.96	92.00
ViT-L/16	69.65	94.54	86.71	26.88	55.63	45.53	95.13	76.64	83.94	86.13	52.53	62.44	93.10

Table 7. The depth results of our methods on the MVSEC-Depth and DSEC-Depth datasets.

Event Model	Metric	**MVSEC-Depth**						**DSEC-Depth**					
		LP	1%	5%	10%	20%	Full	LP	1%	5%	10%	20%	Full
ViT-S/16	$\delta_1 \uparrow$	0.529	0.526	0.531	0.542	0.560	0.577	0.798	0.795	0.804	0.811	0.816	0.824
	RMSE↓	6.756	6.930	6.712	6.477	6.352	6.145	4.861	4.983	4.751	4.728	4.694	4.564
ViT-B/16	$\delta_1 \uparrow$	0.571	0.561	0.574	0.587	0.591	0.594	0.845	0.839	0.856	0.863	0.867	0.872
	RMSE↓	6.392	6.546	6.339	6.012	5.908	5.891	4.352	4.471	4.264	4.192	4.154	4.032
ViT-L/16	$\delta_1 \uparrow$	0.597	0.592	0.601	0.612	0.619	0.625	0.881	0.856	0.883	0.892	0.893	0.896
	RMSE↓	5.884	5.975	5.855	5.724	5.673	5.554	3.857	3.984	3.841	3.759	3.723	3.694

2.2. More Detailed Ablations

Table 8. Ablative study results of different event aggregation methods.

Event Input	**DDD17-Seg**		**DSEC-Depth**		**MVSEC-Flow**	
	Acc↑	mIoU↑	$\delta_1 \uparrow$	RMSE ↓	EPE↓	Out↓
Color Frame	90.76	56.37	0.784	5.306	1.107	6.720
E2VID	89.25	55.72	0.809	4.928	0.852	3.294
Event Volume	91.39	59.64	0.824	4.564	0.356	0.094

Event Aggregations. In our main study, we aggregate the event stream as a three-dimensional volume (voxel grid) to interface cleanly with vision foundation models. Here, we additionally evaluate alternative renderings, including color-like frames [29] and E2VID reconstructions [21]. For a fair comparison, the event representation is held fixed across pretraining and downstream fine-tuning, and all experiments use a ViT-S encoder. As reported in Table 8, the volumetric encoding delivers the strongest overall performance, indicating that explicit spatio-temporal discretization provides a more effective inductive bias for pretraining than image-like aggregations or reconstructed intensities.

Hyper Parameters. To enable cross-modal distillation, we encode event streams as a multi-channel volume/voxel grid compatible with vision foundation models and introduce an activation mask to suppress spurious event–image alignment during pretraining. We ablate two hyperparameters,

Table 9. Ablative study results of different time bins for event volume aggregation.

Time Bin	DDD17-Seg		DSEC-Depth		MVSEC-Flow	
	Acc↑	mIoU↑	δ_1 ↑	RMSE ↓	EPE↓	Out↓
$B = 1$	91.07	58.43	0.819	4.736	0.365	0.104
$\boldsymbol{B = 3}$	91.39	59.64	0.824	4.564	0.356	0.094
$B = 5$	91.20	59.22	0.822	4.613	0.359	0.095

Table 10. Ablative study results of different density thresholds for activation mask constraint.

Density Threshold	DDD17-Seg		DSEC-Depth		MVSEC-Flow	
	Acc↑	mIoU↑	δ_1 ↑	RMSE ↓	EPE↓	Out↓
$\tau = 32$	91.33	59.51	0.823	4.538	0.362	0.095
$\boldsymbol{\tau = 64}$	91.39	59.64	0.824	4.564	0.356	0.094
$\tau = 128$	91.25	59.32	0.821	4.640	0.367	0.097

Table 11. Ablative study results of different distillation objectives across granularities. CL denotes the contrastive loss.

Alignment Objective	DDD17-Seg		DSEC-Depth		MVSEC-Flow	
	Acc↑	mIoU↑	δ_1 ↑	RMSE ↓	EPE↓	Out↓
patch-level (L1)	90.65	56.06	0.785	4.990	0.367	0.098
superpixel-level (L1)	90.88	56.36	0.790	4.937	0.384	0.106
superpixel-level (CL)	90.92	56.72	0.782	5.031	0.435	0.120
ours	91.39	59.64	0.824	4.564	0.356	0.094

Table 12. Ablative study results of multi-scale distillation.

Alignment Objective	DDD17-Seg		DSEC-Depth		MVSEC-Flow	
	Acc↑	mIoU↑	δ_1 ↑	RMSE ↓	EPE↓	Out↓
multi-scale	91.07	58.83	0.816	4.831	0.377	0.102
single-scale	91.39	59.64	0.824	4.564	0.356	0.094

the number of time bins B for volume aggregation, which controls temporal granularity, and the density threshold τ for the activation mask, which trades coverage for noise suppression. Unless otherwise specified, all comparisons use a ViT-S encoder. Results in Tables 9 and 10 identify the optimal configuration.

Superpixel Alignment. For cross-modal distillation, we formulate a hierarchical objective comprising patch-level supervision (our baseline) and structure-level supervision (our highlight). Here, we further examine a superpixel-level variant. Following OpenESS [11], we partition each image into 100 SAM-derived [10] superpixels and compare two formulations: (i) an L1 regression loss on superpixel-aggregated features, and (ii) a contrastive objective inspired by image-point cloud distillation [22] that enforces intra-superpixel compactness and inter-superpixel separability. Unless otherwise specified, all comparisons use a ViT-S encoder. Results in Table 11 reveal that superpixel-level alignment underperforms, due to semantically ambiguous groupings (e.g., boundary leakage and region fragmentation) that is consistent with our overall analysis.

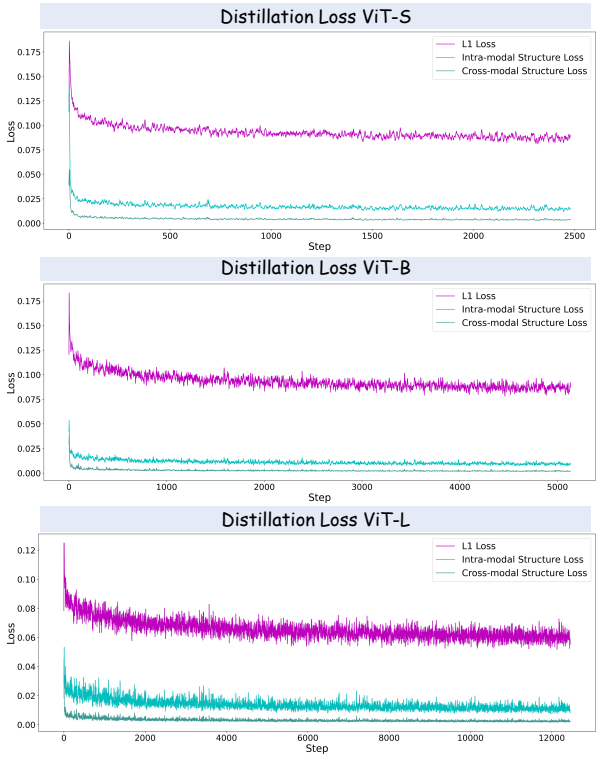


Figure 2. **Cross-modal distillation loss during pretraining** of our event-based ViT-S, ViT-B, and ViT-L feature encoders.

Multi-scale Distillation. For cross-modal distillation, our main study aligns only the terminal features of the encoder. Here, we additionally assess a multi-scale alignment scheme. All comparisons use a ViT-S encoder. Specifically, we align intermediate activations from layers 3, 6, 9, and 12 to their event counterparts with equal loss weights. Results in Table 12 show that multi-scale alignment underperforms, likely because intermediate representations possess weak and unstable semantics and thus exacerbate the event-image modality gap.

2.3. Pretraining Loss

The pretraining losses are depicted in Figure 2.

2.4. Computational Efficiency

The computational efficiency analysis is shown in Table 13.

Table 13. **Computational efficiency** of our downstream task models, setting an input event volume resolution of 480×640.

	Segment Model		Depth Model		Flow Model	
	MParams	GFLOP	MParams	GFLOPs	MParams	GFLOPs
ViT-S	28.23	243.67	18.92	61.23	40.95	349.42
ViT-B	76.74	363.82	74.98	231.93	135.62	962.71
ViT-L	239.09	758.17	257.29	853.16	485.27	3369.48

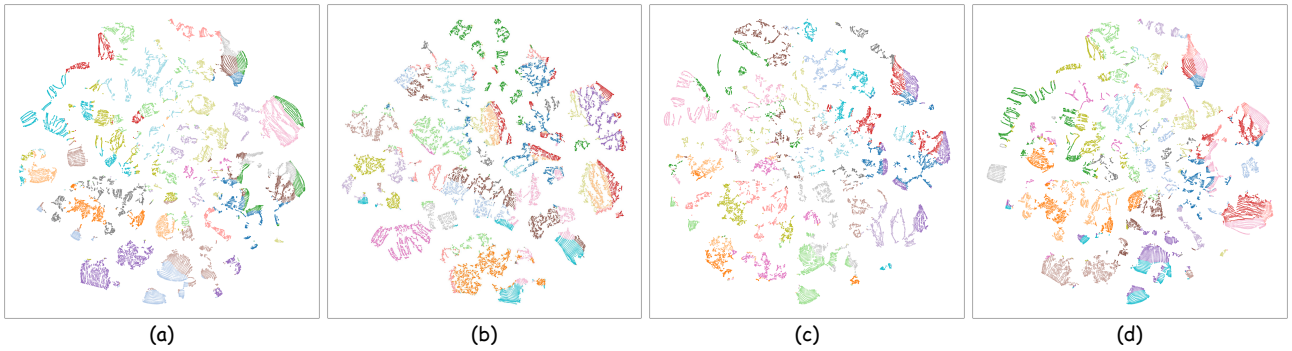


Figure 3. **T-SNE plots of learned event features.** We sample 20,000 event feature vectors from the DSEC dataset [6]. We show features from **(a)** images with pretrained DIVOv3-L; **(b)** event volume with pretrained DIVOv3-L; **(c)** event volume with DIVOv3-L after patch-level distillation; **(d)** event volume with DIVOv3-L using our distillation method.

3. More Qualitative Results

3.1. Representation Visualization

Statistical Analysis. As shown in Figure 3, t-SNE plots of feature vectors from the DSEC dataset [6] highlight the performance of various models and distillation methods. The pretrained DINOv3-L on images shows strong clustering with some overlap, indicating effective feature learning but room for finer distinctions. Event volume with pretrained DINOv3-L shows greater dispersion, highlighting challenges in capturing event-specific features and temporal dynamics. Patch-level distillation improves feature separation, resulting in more compact clusters. Our distillation method achieves the most distinct and well-separated clusters, closely matching the pretrained DINOv3-L while better capturing event-specific features.

Exemplary Analysis. As shown in Figure 4 and Figure 5, exemplary learned event features are visualized through cosine similarity maps, with key points marked by white stars. The RGB reference images and corresponding event data are shown on the left, while the cosine similarity maps (scaled by a factor of 4) highlight the areas where the model focuses. These maps emphasize the spatial locations of distinctive event features, demonstrating how the model captures dynamic, fine-grained details. The alignment of the white stars with key features indicates the model's ability to identify significant event-driven changes. The results highlight the model's effectiveness in learning and refining event features, benefiting from the cross-modal distillation of pretrained image-based models to better capture these event features.

3.2. Downstream Tasks

Representative qualitative results for downstream tasks are provided in Figures 6, 7, and 8.

Semantic Segmentation. As shown in Figure 6, the comparison of event-based semantic segmentation methods on the DSEC-Semantic dataset highlights the effectiveness of cross-modal distillation for dense event pretraining. Our method significantly improves segmentation quality, particularly in fine-grained object boundaries and dynamic features like persons, cars, and traffic signs. The key advantage lies in leveraging pretrained image models through cross-modal distillation, which enhances spatial feature learning in event data. In contrast, methods like ESS-Sup and OpenESS perform well in general segmentation but fail to capture subtle event-driven features, while KWYAF and 6T show some improvement but struggle in dynamic scenes. Our method outperforms them by maintaining high accuracy.

Monocular Depth Estimation. As shown in Figure 7, the comparison of event-based depth estimation methods on the DSEC-Depth dataset demonstrates the benefits of cross-modal distillation for dense event pretraining. Our method produces the most accurate depth maps, especially in dynamic regions with moving objects or occlusions. In contrast, methods like E2Depth and EReformer show noticeable errors, particularly in complex environments. While DepthAnyEvent performs well in static areas, it struggles with depth variations in motion. Our method, leveraging cross-modal pretraining, improves depth accuracy, particularly in foreground-background transitions, by transferring rich spatial knowledge to the event-based depth task.

Optical Flow Estimation. As shown in Figure 8, the comparison of optical flow estimation results on the MVSEC-Flow dataset highlights the effectiveness of our cross-modal distillation approach. Our method produces the highly accurate and consistent flow predictions, thanks to cross-modal distillation from pretrained models, which enhances flow estimation by leveraging fine-grained correlation knowledge. By transferring knowledge from image-based foundation model, our method improves robustness, capturing fine details and rapid motion changes effectively in event-based data.

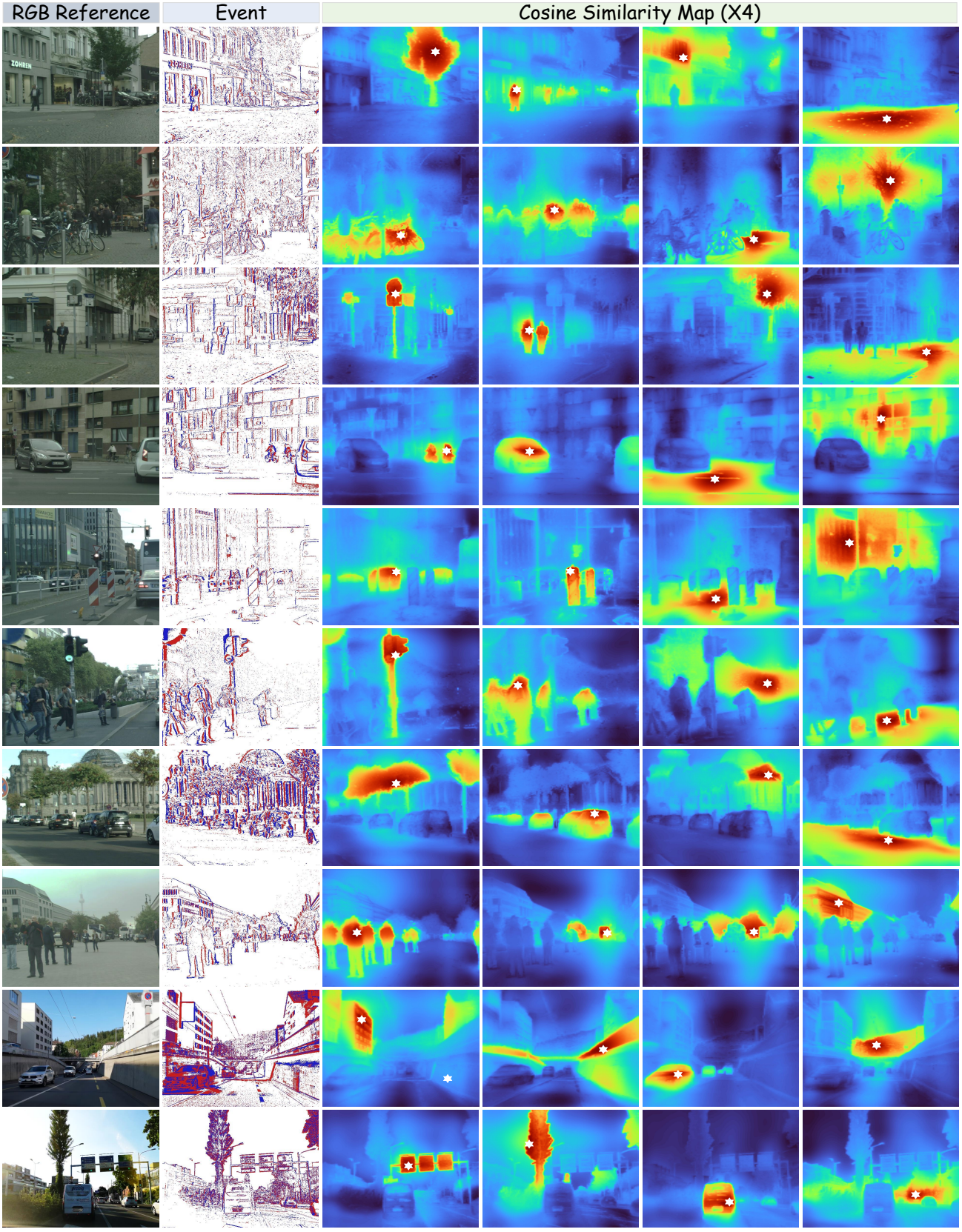


Figure 4. **The learned fine-grained event features (1/2)** of our method are primarily presented through cosine similarity maps, with key points anchored at the distinct white stars. Best viewed in color.

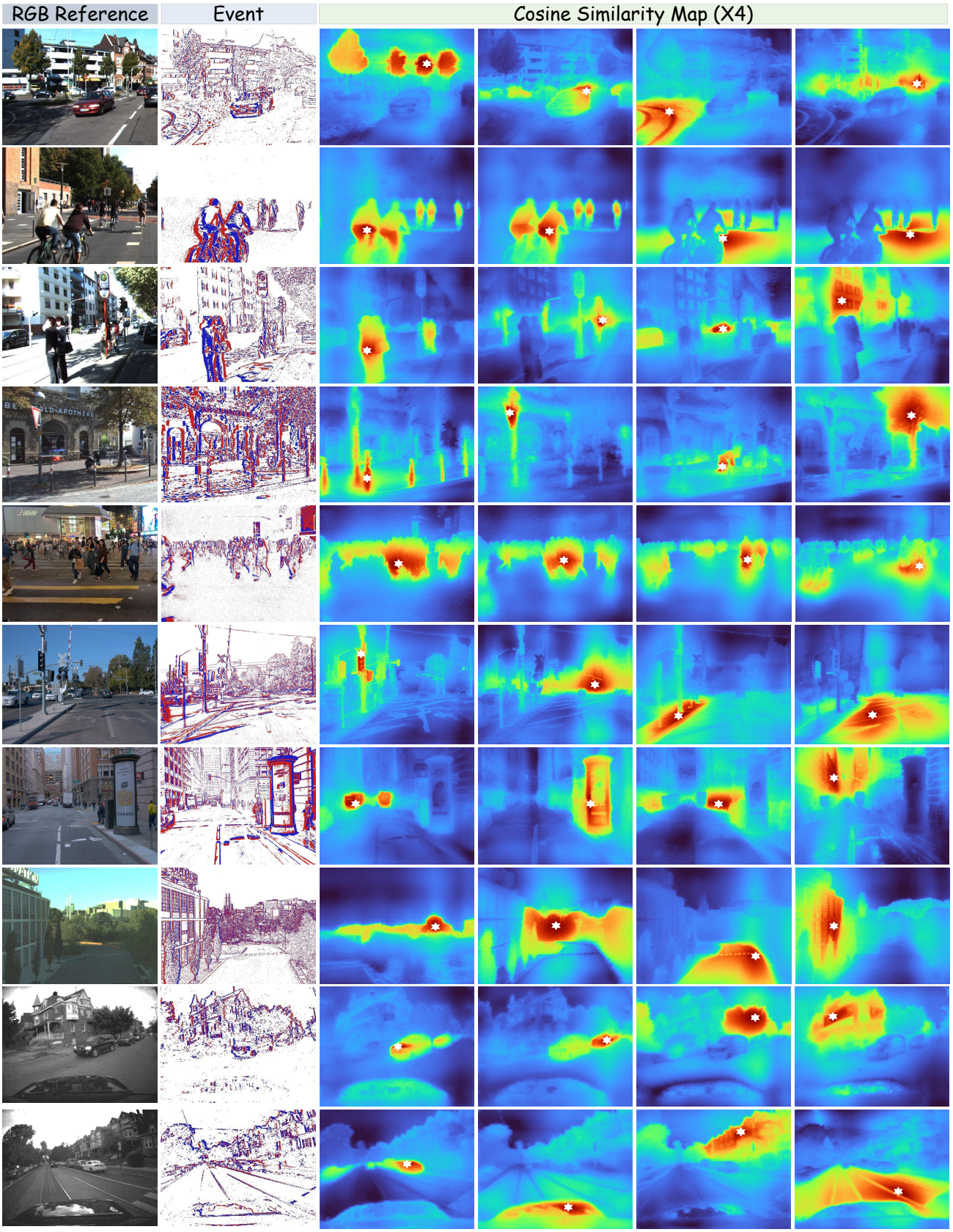


Figure 5. **The learned fine-grained event features (2/2)** of our method are primarily presented through cosine similarity maps, with key points anchored at the distinct white stars. Best viewed in color.

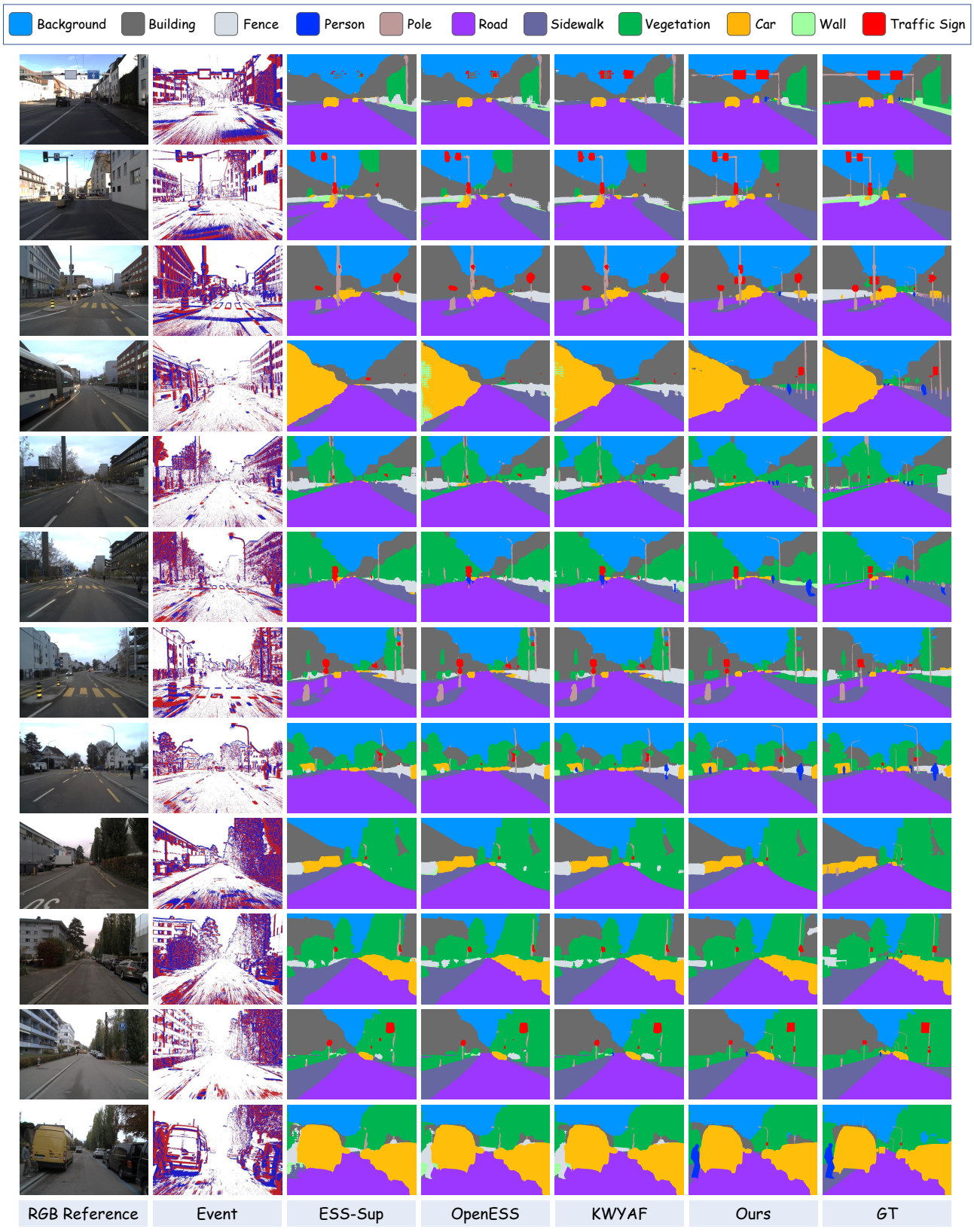


Figure 6. The qualitative comparisons among different **event-based semantic segmention** approaches on the test set of DSEC-Semantic. Best viewed in color.

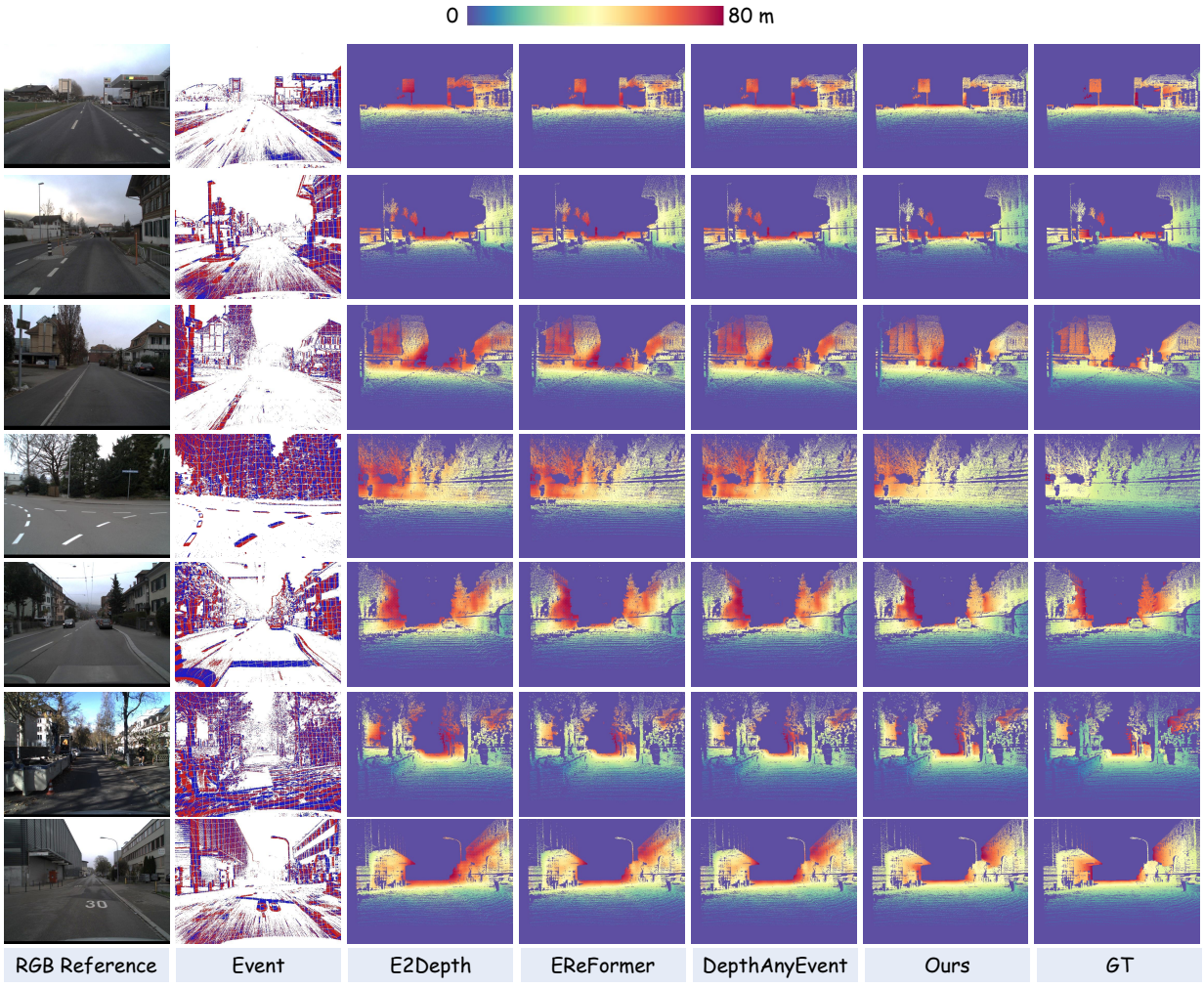


Figure 7. The qualitative comparisons among different **event-based depth estimation** approaches on the test set of DSEC-Depth. Best viewed in color.

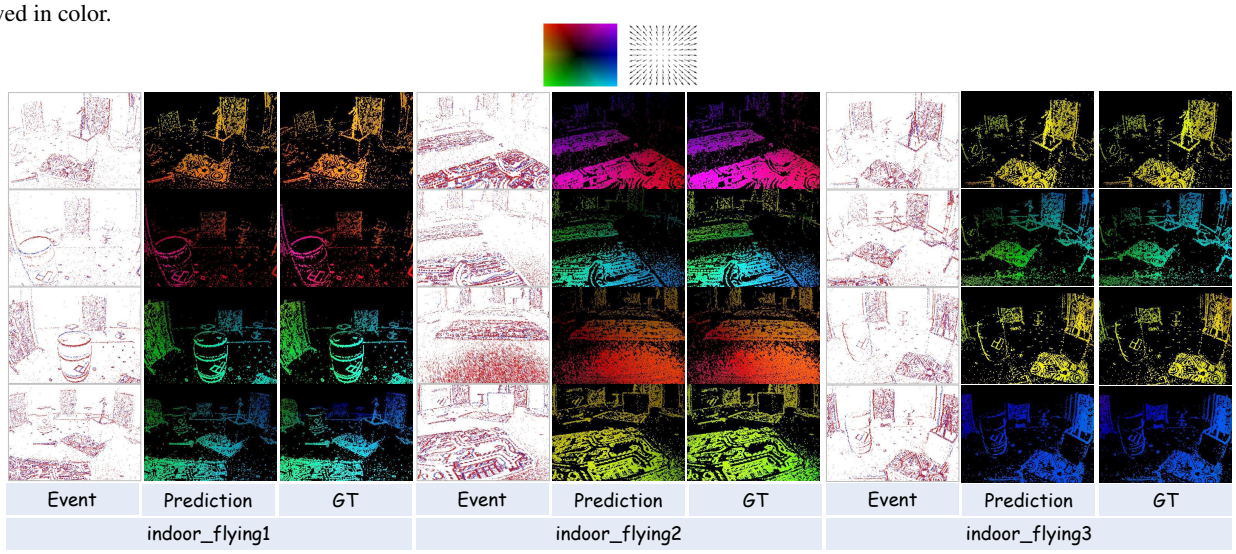


Figure 8. The qualitative results of our **optical flow estimation** approaches on the test set of MVSEC-Flow. Best viewed in color.

4. Limitation and Discussion

While our approach significantly advances event-based pretraining, several limitations remain. First, although our structure-aware distillation improves event representation quality, higher resolutions still face some degradation, particularly with patch- and superpixel-level distillation. This suggests that fine-grained alignment methods could be further refined to handle high-resolution event data more effectively. Second, our method relies on large-scale, synchronized image-event datasets, which may not always be feasible to obtain in certain domains. Future work could explore semi-supervised or unsupervised distillation approaches to reduce reliance on these extensive datasets. Additionally, while our model performs well across standard downstream tasks, its ability to generalize to new or rare event-camera configurations remains limited. Addressing this could involve incorporating domain adaptation or meta-learning strategies to improve robustness in more dynamic or occluded environments. Lastly, the computational efficiency of our method, particularly with large encoder models, presents a challenge. Optimizing for lighter backbones or reducing redundant parameters could enhance the applicability of our approach in resource-constrained real-world scenarios, such as robotics or autonomous vehicles.

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 3

[2] Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19669–19678, 2025. 3

[3] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017. 1, 2

[4] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4015–4022, 2023. 2

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[6] 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. 2, 3, 7

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 3

[8] Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22487–22497, 2025. 1

[9] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024. 2

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 6

[11] Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15686–15698, 2024. 6

[12] Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, et al. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12871, 2024. 1

[13] Ke Li, Gengyu Lyu, Hao Chen, Bochen Xie, Zhen Yang, Youfu Li, and Yongjian Deng. Know where you are from: Event-based segmentation via spatio-temporal propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4806–4814, 2025. 1

[14] Quanmin Liang, Qiang Li, Shuai Liu, Xinzi Cao, Jinyi Lu, Feidiao Yang, Wei Zhang, Kai Huang, and Yonghong Tian. Efficient event camera data pretraining with adaptive prompt fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8656–8667, 2025. 1

[15] Xu Liu, Jianing Li, Jinqiao Shi, Xiaopeng Fan, Yonghong Tian, and Debin Zhao. Event-based monocular depth estimation with recurrent transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7417–7429, 2024. 3

[16] Yunfan Lu, Xiaogang Xu, Hao Lu, Yanlin Qian, Pengteng Li, Huizai Yao, Bin Yang, Junyi Li, Qianyi Cai, Weiyu Guo, et al. See: See everything every time–adaptive brightness adjustment for broad light range images via events. *arXiv preprint arXiv:2502.21120*, 2025. 2

[17] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[19] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 3

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[21] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 3, 5

[22] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 6

[23] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1

[24] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 2

[25] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3

[26] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 1, 3

[27] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Shifeng Chen, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *Pattern Recognition*, page 112718, 2025. 2

[28] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9700–9709, 2021. 3

[29] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3): 1997–2010, 2023. 2, 5

[30] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795, 2024. 1

[31] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10699–10709, 2023. 4

[32] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision*, pages 292–310. Springer, 2024. 4

[33] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 1

[34] Yucheng Zhao, Gengyu Lyu, Ke Li, Zihao Wang, Hao Chen, Zhen Yang, and Yongjian Deng. Eseg: Event-based segmentation boosted by explicit edge-semantic guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10510–10518, 2025. 1

[35] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3 (3):2032–2039, 2018. 2, 3, 4

[36] 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. 4

[37] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 1