OpenESS: Event-based Semantic Scene Understanding with Open Vocabularies

Lingdong Kong[1,2] Youquan Liu[3] Lai Xing Ng[4,5] Benoit R. Cottereau[5,6] Wei Tsang Ooi[1,5]
[1]National University of Singapore [2]CNRS@CREATE [3]Hochschule Bremerhaven
[4]Institute for Infocomm Research, A*STAR [5]IPAL, CNRS IRL 2955, Singapore
[6]CerCo, CNRS UMR 5549, Université Toulouse III
https://github.com/ldkong1205/OpenESS

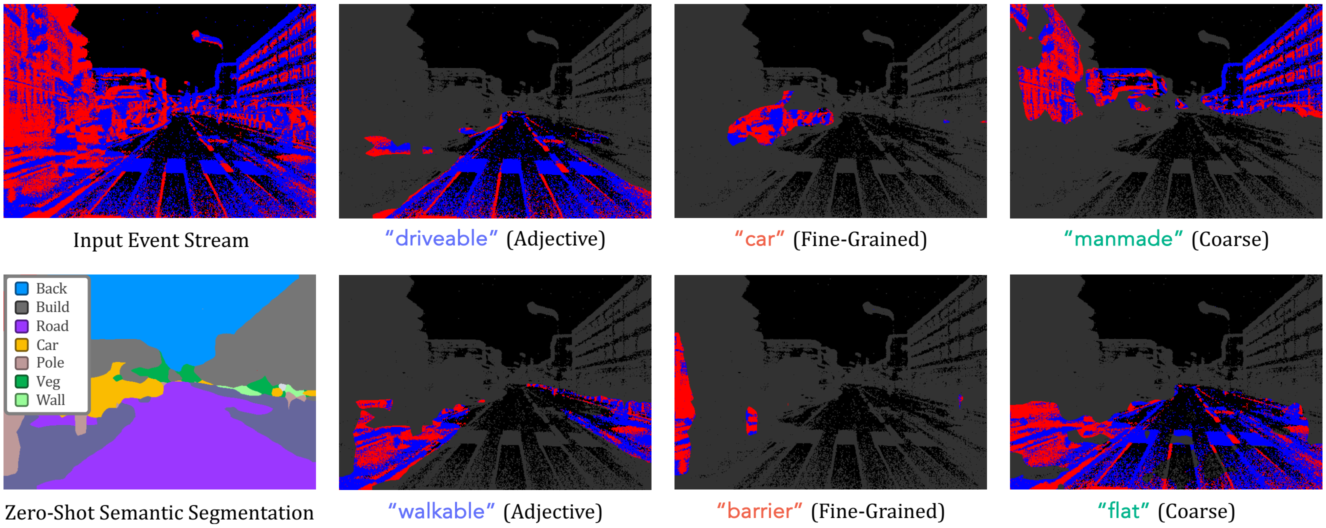


Figure 1. **Open-vocabulary event-based semantic segmentation (OpenESS)**. Our framework is capable of performing zero-shot semantic segmentation of event data streams with open vocabularies. Given raw events and text prompts as inputs, OpenESS outputs semantically coherent open-world predictions across various adjective, fine-grained, and coarse categories. The last three columns show the language-guided attention maps where regions of a high similarity score to the given text prompts are highlighted. Best viewed in colors.

Abstract

*Event-based semantic segmentation (ESS) is a fundamental yet challenging task for event camera sensing. The difficulties in interpreting and annotating event data limit its scalability. While domain adaptation from images to event data can help to mitigate this issue, there exist data representational differences that require additional effort to resolve. In this work, for the first time, we synergize information from image, text, and event-data domains and introduce **OpenESS** to enable scalable ESS in an open-world, annotation-efficient manner. We achieve this goal by transferring the semantically rich CLIP knowledge from image-text pairs to event streams. To pursue better cross-modality adaptation, we propose a frame-to-event contrastive distillation and a text-to-event semantic consistency regularization. Experimental results on popular ESS benchmarks showed our approach outperforms existing methods. Notably, we achieve 53.93% and 43.31% mIoU on DDD17 and DSEC-Semantic without using either event or frame labels.*

1. Introduction

Event cameras, often termed bio-inspired vision sensors, stand distinctively apart from traditional frame-based cameras and are often merited by their low latency, high dynamic range, and low power consumption [28, 44, 76]. The realm of event-based vision perception, though nascent, has rapidly evolved into a focal point of contemporary research [99]. Drawing parallels with frame-based perception and recognition methodologies, a plethora of task-specific applications leveraging event cameras have burgeoned [25].

Event-based semantic segmentation (ESS) emerges as one of the core event perception tasks and has gained increasing attention [2, 6, 38, 79]. ESS inherits the challenges of traditional image segmentation [11, 12, 19, 39, 58], while also contending with the unique properties of event data [2], which opens up a plethora of opportunities for exploration. Although accurate and efficient dense predictions from event cameras are desirable for practical applications, the learning and annotation of the sparse, asynchronous, and high-temporal-resolution event streams pose several challenges [47, 49, 61]. Stemming from the image segmentation community, existing ESS models are trained on *densely annotated* events within a *fixed* and *limited* set of label mapping [2, 79]. Such closed-set learning from expensive annotations inevitably constrains the scalability of ESS systems.

An obvious approach will be to make use of the image domain and transfer knowledge to event data for the same vision tasks. Several recent attempts [30, 61, 79] resort to unsupervised domain adaptation to avoid the need for paired image and event data annotations for training. These methods demonstrate the potential of leveraging frame annotations to train a segmentation model for event data. However, transferring knowledge across frames and events is not straightforward and requires intermediate representations such as voxel grids, frame-like reconstructions, and bio-inspired spikes. Meanwhile, it is also costly to annotate dense frame labels for training, which limits their usage.

A recent trend inclines to the use of multimodal foundation models [13, 50, 67, 69, 94] to train task-specific models in an open-vocabulary and zero-shot manner, removing dependencies on human annotations. This paper continues such a trend. We propose a novel open-vocabulary framework for ESS, aiming at transferring pre-trained knowledge from both image and text domains to learn better representations of event data for the dense scene understanding task. Observing the large domain gap in between heterogeneous inputs, we design two cross-modality representation learning objectives that gradually align the event streams with images and texts. As shown in Fig. 1, given raw events and text prompts as the input, the learned feature representations from our OpenESS framework exhibit promising results for known and unknown class segmentation and can be extended to more open-ended texts such as *"adjectives"*, *"fine-grained"*, and *"coarse-grained"* descriptions.

To sum up, this work poses key contributions as follows:

- We introduce OpenESS, a versatile event-based semantic segmentation framework capable of generating open-world dense event predictions given arbitrary text queries.
- To the best of our knowledge, this work represents the first attempt at distilling large vision-language models to assist event-based semantic scene understanding tasks.
- We propose a frame-to-event (F2E) contrastive distillation and a text-to-event (T2E) consistency regularization to encourage effective cross-modality knowledge transfer.
- Our approach sets up a new state of the art in annotation-free, annotation-efficient, and fully-supervised ESS settings on *DDD17-Seg* and *DSEC-Semantic* benchmarks.

2. Related Work

Event-based Vision. The microsecond-level temporal resolution, high dynamic range (typically 140 dB *vs.* 60 dB of standard cameras), and power consumption efficiency of event cameras have posed a paradigm shift from traditional frame-based imaging [25, 60, 77, 108]. A large variety of event-based recognition, perception, localization, and reconstruction tasks have been established, encompassing object recognition [18, 29, 48, 68], object detection [27, 31, 103, 109], depth estimation [17, 36, 42, 62, 65, 70], optical flow [7, 20, 33, 34, 53, 81, 105], intensity-image reconstruction [23, 24, 73, 98, 107], visual odometry and SLAM [43, 56, 72], stereoscopic panoramic imaging [4, 75], *etc.* In this work, we focus on the recently-emerged task of event-based semantic scene understanding [2, 79]. Such a pursuit is anticipated to tackle sparse, asynchronous, and high-temporal-resolution events for dense predictions, which is crucial for safety-critical in-drone or in-vehicle perceptions.

Event-based Semantic Segmentation. The focus of ESS is on categorizing events into semantic classes for enhancing scene interpretation. Alonso *et al.* [2] contributed the first benchmark based on DDD17 [5]. Subsequent works are tailored to improve the accuracy while mitigating the need for extensive event annotations [30]. EvDistill [84] and DTL [83] utilized aligned frames to enhance event-based learning. EV-Transfer [61] and ESS [79] leveraged domain adaptation to transfer knowledge from existing image datasets to events. Recently, HALSIE [6] and HMNet [38] innovated ESS in cross-domain feature synthesis and memory-based event encoding. Another line of research pursues to use of spiking neural networks for energy-efficient ESS [10, 49, 63, 90]. In this work, different from previous pursuits, we aim to train ESS models in an annotation-free manner by distilling pre-trained vision-language models, hoping to address scalability and annotation challenges.

Open-Vocabulary Learning. Recent advances in vision-language models open up new possibilities for visual perceptions [13, 88, 106]. Such trends encompass image-based zero-shot and open-vocabulary detection [26, 52, 89, 96], as well as semantic [35, 51, 55, 97, 100], instance [45, 87], and panoptic [21, 41, 93] segmentation. As far as we know, only three works studied the adaptation of CLIP for event-based recognition. EventCLIP [92] proposed to convert events to a 2D grid map and use an adapter to align event features with CLIP's knowledge. E-CLIP [102] uses a hierarchical triple contrastive alignment that jointly unifies the event, image, and text feature embedding. Ev-LaFOR [18] designed category-guided attraction and category-agnostic repulsion losses to bridge event with CLIP. Differently, we present the first attempt at adapting CLIP for dense predictions on sparse and asynchronous event streams. Our work is also close to superpixel-driven contrastive learning [46, 74], where pre-processed superpixels are used to

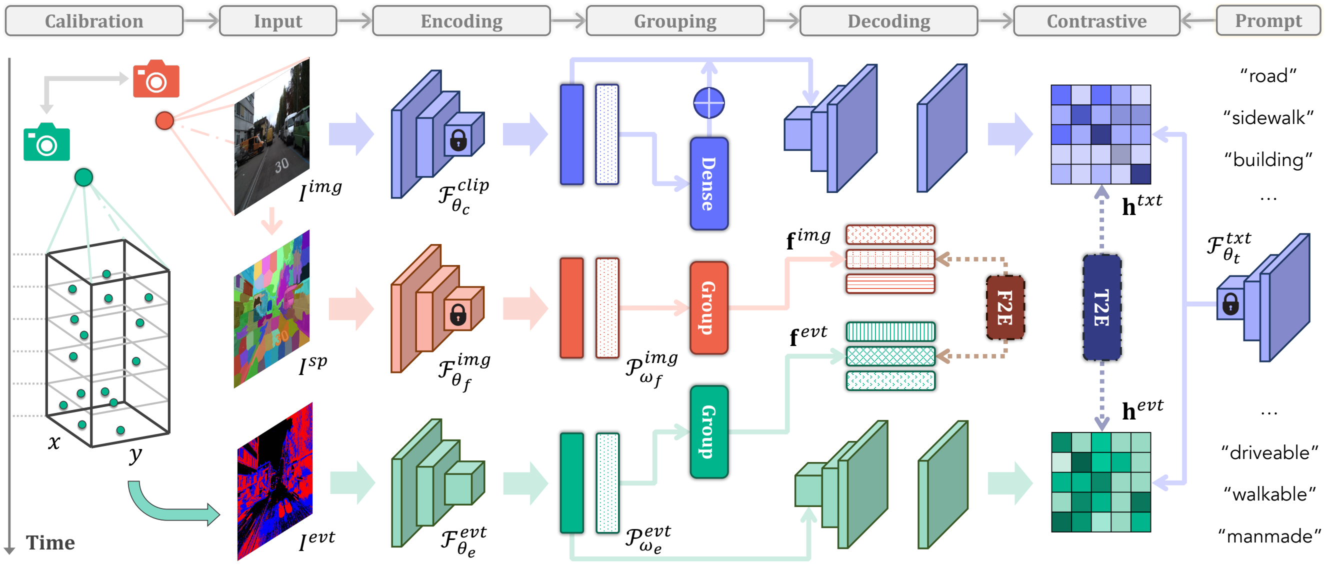


Figure 2. **Architecture overview of the OpenESS framework**. We distill off-the-shelf knowledge from vision-languages models to event representations (*cf*. Sec. 3.1). Given a calibrated event I^{evt} and a frame I^{img}, we extract their features from the event network $\mathcal{F}^{evt}_{\theta_e}$ and the densified CLIP's image encoder $\mathcal{F}^{clip}_{\theta_c}$, which are then combined with the text embedding from CLIP's text encoder $\mathcal{F}^{txt}_{\theta_t}$ for open-world prediction (*cf*. Sec. 3.2). To better serve for cross-modality knowledge transfer, we propose a **frame-to-event (F2E)** contrastive objective (*cf*. Sec. 3.3) via superpixel-driven distillation and a **text-to-event (T2E)** consistency objective (*cf*. Sec. 3.4) via scene-level regularization.

establish contrastive objectives with modalities from other tasks, *e.g*., point cloud understanding [57], remote sensing [37], medical imaging [82], and so on. In this work, we propose OpenESS to explore superpixel-to-event representation learning. Extensive experiments verify that such an approach is promising for annotation-efficient ESS.

3. Methodology

Our study serves as an early attempt at leveraging vision-language foundation models like CLIP [69] to learn meaningful event representations without accessing ground-truth labels. We start with a brief introduction of the CLIP model (*cf*. Sec. 3.1), followed by a detailed elaboration on our proposed open-vocabulary ESS (*cf*. Sec. 3.2). To encourage effective cross-modal event representation learning, we introduce a frame-to-event contrastive distillation (*cf*. Sec. 3.3) and a text-to-event consistency regularization (*cf*. Sec. 3.4). An overview of the OpenESS framework is shown in Fig. 2.

3.1. Revisiting CLIP

CLIP [69] learns to associate images with textual descriptions through a contrastive learning framework. It leverages a dataset of 400 million image-text pairs, training an image encoder (based on a ResNet [39] or Vision Transformer [22]) and a text encoder (using a Transformer architecture [80]) to project images and texts into a shared embedding space. Such a training paradigm enables CLIP to perform zero-shot classification tasks, identifying images based on textual descriptions without specific training on those categories. To achieve annotation-free classification on a custom dataset, one needs to combine class label mappings with hand-crafted text prompts as the input to generate the text embedding. In this work, we aim to leverage the semantically rich CLIP feature space to assist open-vocabulary dense prediction on sparse and asynchronous event streams.

3.2. Open-Vocabulary ESS

Inputs. Given a set of N event data acquired by an event camera, we aim to segment each event $\mathbf{e}_i$ among the temporally ordered event streams ε_i, which are encoded by the pixel coordinates $(\mathbf{x}_i, \mathbf{y}_i)$, microsecond-level timestamp t_i, and the polarity $p_i \in \{-1, +1\}$ which indicates either an increase or decrease of the brightness. Each event camera pixel generates a spike whenever it perceives a change in logarithmic brightness that surpasses a predetermined threshold. Meanwhile, a conventional camera captures gray-scale or color frames $I^{img}_i \in \mathbb{R}^{3\times H\times W}$ which are spatially aligned and temporally synchronized with the events or can be aligned and synchronized to events via sensor calibration, where H and W are the spatial resolutions.

Event Representations. Due to the sparsity, high temporal resolution, and asynchronous nature of event streams, it is common to convert raw events ε_i into more regular representations $I^{evt}_i \in \mathbb{R}^{C\times H\times W}$ as the input to the neural network [25], where C denotes the number of embedding channels which is depended on the event representations themselves. Some popular choices of such embedding include spatiotemporal voxel grids [29, 104, 105], frame-like reconstructions [73], and bio-inspired spikes [49]. We investigate these three methods and show an example of taking voxel grids as the input in Fig. 2. More analyses and comparisons using reconstructions and spikes are in later sections. Specifically, with a predefined number of events, each voxel grid is built from non-overlapping windows as:

$$I_i^{evt} = \sum_{\mathbf{e}_j \in \varepsilon_i} p_j \delta(\mathbf{x}_j - \mathbf{x}) \delta(\mathbf{y}_j - \mathbf{y}) \max\{1 - |t_j^* - t|, 0\}, \tag{1}$$

where δ is the Kronecker delta function; $t_j^* = (B-1)\frac{t_j - t_0}{\Delta T}$ is the normalized event timestamp with B as the number of temporal bins in an event stream; ΔT is the time window and t_0 denotes the time of the first event in the window.

Cross-Modality Encoding. Let $\mathcal{F}_{\theta_e}^{evt} : \mathbb{R}^{C \times H \times W} \mapsto \mathbb{R}^{D_1 \times H_1 \times W_1}$ be an event-based segmentation network with trainable parameters θ_e, which takes as input an event embedding I_i^{evt} and outputs a D_1-dimensional feature of downsampled spatial sizes H_1 and W_1. Meanwhile, we integrate CLIP's image encoder $\mathcal{F}_{\theta_c}^{clip} : \mathbb{R}^{3 \times H \times W} \mapsto \mathbb{R}^{D_2 \times H_2 \times W_2}$ into our framework and keep the parameters θ_c fixed. The output is a D_2-dimensional feature of sizes H_2 and W_2. Our motivation is to transfer general knowledge from $\mathcal{F}_{\theta_c}^{clip}$ to $\mathcal{F}_{\theta_e}^{evt}$, such that the event branch can learn useful representations without using dense event annotations. To enable open-vocabulary ESS predictions, we leverage CLIP's text encoder $\mathcal{F}_{\theta_t}^{txt}$ with pre-trained parameters θ_t. The input of $\mathcal{F}_{\theta_t}^{txt}$ comes from predefined text prompt templates and the output will be a text embedding extracted from CLIP's rich semantic space.

Densifications. CLIP was originally designed for image-based recognition tasks and does not provide per-pixel outputs for dense predictions. Several recent attempts explored the adaptation from global, image-level recognition to local, pixel-level prediction, via either model structure modification [100] or fine-tuning [51, 71, 97]. The former directly reformulates the value-embedding layer in CLIP's image encoder, while the latter uses semantic labels to gradually adapt the pre-trained weights to generate dense predictions. In this work, we implement both solutions to densify CLIP's outputs and compare their performances in our experiments.

Up until now, we have presented a preliminary framework capable of conducting open-vocabulary ESS by leveraging knowledge from the CLIP model. However, due to the large domain gap between the event and image modalities, a naïve adaptation is sub-par in tackling the challenging event-based semantic scene understanding task.

3.3. F2E: Frame-to-Event Contrastive Distillation

Since our objective is to encourage effective cross-modality knowledge transfer for holistic event scene perception, it thus becomes crucial to learn meaningful representations for both *thing* and *stuff* classes, especially their boundary information. However, the sparsity and asynchronous nature of event streams inevitably impede such objectives.

Superpixel-Driven Knowledge Distillation. To pursue a more informative event representation learning at higher granularity, we propose to first leverage calibrated frames to generate coarse, instance-level superpixels and then distill knowledge from a pre-trained image backbone to the event segmentation network. Superpixel groups pixels into conceptually meaningful atomic regions, which can be used as the basis for higher-level perceptions [1, 54, 85]. The semantically coherent frame-to-event correspondences can thus be found using pre-processed or online-generated superpixels. Such correspondences tend to bridge the sparse events to dense frame pixels in a holistic manner without involving extra training or annotation efforts.

Superpixel & Superevent Generation. We resort to the following two ways of generating the superpixels. The first way is to leverage heuristic methods, *e.g.* SLIC [1], to efficiently groups pixels from frame I_i^{img} into a total of M_{slic} segments with good boundary adherence and regularity as $I_i^{sp} = \{\mathcal{I}_i^1, \mathcal{I}_i^2, ..., \mathcal{I}_i^{M_{slic}}\}$, where M_{slic} is a hyperparameter that needs to be adjusted based on the inputs. The generated superpixels satisfy $\mathcal{I}_i^1 \cup \mathcal{I}_i^2 \cup ... \cup \mathcal{I}_i^{M_{slic}} = \{1, 2, ..., H \times W\}$. For the second option, we use the recent Segment Anything Model (SAM) [50] which takes I_i^{img} as the input and outputs M_{sam} class-agnostic masks. For simplicity, we use M to denote the number of superpixels used during knowledge distillation, *i.e.*, $\{I_i^{sp} = \{\mathcal{I}_i^1, ..., \mathcal{I}_i^k\} | k = 1, ..., M\}$ and show more comparisons between SLIC [1] and SAM [50] in later sections. Since I_i^{evt} and I_i^{img} have been aligned and synchronized, we can group events from I_i^{evt} into superevents $\{V_i^{sp} = \{\mathcal{V}_i^1, ..., \mathcal{V}_i^l\} | l = 1, ..., M\}$ by using the known event-pixel correspondences.

Frame-to-Event Contrastive Learning. To encourage better superpixel-level knowledge transfer, we leverage a pre-trained image network $\mathcal{F}_{\theta_f}^{img} : \mathbb{R}^{3 \times H \times W} \mapsto \mathbb{R}^{D_3 \times H_3 \times W_3}$ as the teacher and distill information from it to the event branch $\mathcal{F}_{\theta_e}^{evt}$. The parameters of $\mathcal{F}_{\theta_f}^{img}$, which can come from either CLIP [69] or other pretext task pre-trained backbones such as [8, 15, 64], are kept frozen during the distillation. With $\mathcal{F}_{\theta_e}^{evt}$ and $\mathcal{F}_{\theta_f}^{img}$, we generate the superevent and superpixel features as follows:

$$\mathbf{f}_i^{evt} = \frac{1}{|V_i^{sp}|} \sum_{l \in V_i^{sp}} \mathcal{P}_{\omega_e}^{evt} \left(\mathcal{F}_{\theta_e}^{evt} (I_i^{evt})_l \right), \tag{2}$$

$$\mathbf{f}_i^{img} = \frac{1}{|I_i^{sp}|} \sum_{k \in I_i^{sp}} \mathcal{P}_{\omega_f}^{img} \left(\mathcal{F}_{\theta_f}^{img} (I_i^{img})_k \right), \tag{3}$$

where $\mathcal{P}_{\omega_e}^{evt}$ and $\mathcal{P}_{\omega_f}^{img}$ are projection layers with trainable parameters ω_e and ω_f, respectively, for the event branch and frame branch. In the actual implementation, $\mathcal{P}_{\omega_e}^{evt}$ and

$\mathcal{P}^{img}_{\omega_f}$ consist of linear layers which map the D_1- and D_3-dimensional event and frame features to the same shape. The following contrastive learning objective is applied to the event prediction and the frame prediction:

$$\mathcal{L}_{F2E}(\theta_e, \omega_e, \omega_f) = -\sum_i \log \left[\frac{e^{(\langle \mathbf{f}^{evt}_i, \mathbf{f}^{img}_i \rangle / \tau_1)}}{\sum_{j \neq i} e^{(\langle \mathbf{f}^{evt}_i, \mathbf{f}^{img}_j \rangle / \tau_1)}} \right] , \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product between the superevent and superpixel embedding; $\tau_1 > 0$ is a temperature coefficient that controls the pace of knowledge transfer.

Role in Our Framework. Our F2E contrastive distillation establishes an effective pipeline for transferring superpixel-level knowledge from dense, visual informative frame pixels to sparse, irregular event streams. Since we are targeting the semantic segmentation task, the learned event representations should be able to reason in terms of instances and instance parts at and in between semantic boundaries.

3.4. T2E: Text-to-Event Consistency Regularization

Although the aforementioned frame-to-event knowledge transfer provides a simple yet effective way of transferring off-the-shelf knowledge from frames to events, the optimization objective might encounter unwanted conflicts.

Intra-Class Optimization Conflict. During the model pretraining, the superpixel-driven contrastive loss takes the corresponding superevent and superpixel pair in a batch as the positive pair, while treating all remaining pairs as negative samples. Since heuristic superpixels only provide a coarse grouping of conceptually coherent segments (kindly refer to our Appendix for more detailed analysis), it is thus inevitable to encounter self-conflict during the optimization. That is to say, from hindsight, there is a chance that the superpixels belonging to the same semantic class could be involved in both positive and negative samples.

Text-Guided Semantic Regularization. To mitigate the possible self-conflict in Eq. (4), we propose a text-to-event semantic consistency regularization mechanism that leverages CLIP's text encoder to generate semantically more consistent text-frame pairs $\{I^{img}_i, T_i\}$, where T_i denotes the text embedding extracted from $\mathcal{F}^{txt}_{\theta_t}$. Such a paired relationship can be leveraged via CLIP without additional training. We then construct event-text pairs $\{I^{evt}_i, T_i\}$ by propagating the alignment between events and frames. Specifically, the paired event and text features are extracted as follows:

$$\mathbf{h}^{evt}_i = \mathcal{Q}^{evt}_{\omega_q}\left(\mathcal{F}^{evt}_{\theta_e}\left(I^{evt}_i\right)\right) , \ \mathbf{h}^{txt}_i = \mathcal{F}^{txt}_{\theta_t}\left(T_i\right) , \tag{5}$$

where $\mathcal{Q}^{evt}_{\omega_q}$ is a projection layer with trainable parameters ω_q, which is similar to that of $\mathcal{P}^{evt}_{\omega_e}$. Now assume there are a total of Z classes in the event dataset, the following objective is applied to encourage the consistency regularization:

$$\mathcal{L}_{T2E}(\theta_e, \omega_q) = \tag{6}$$

$$-\sum_{z=1}^{Z} \log \left[\frac{\sum_{T_i \in z, I^{evt}_i} e^{(\langle \mathbf{h}^{evt}_i, \mathbf{h}^{txt}_i \rangle / \tau_2)}}{\sum_{j \neq i, T_i \in z, T_i \notin I^{evt}_i} e^{(\langle \mathbf{h}^{evt}_j, \mathbf{h}^{txt}_i \rangle / \tau_2)}} \right] , \tag{7}$$

where $\tau_2 > 0$ is a temperature coefficient that controls the pace of knowledge transfer. The overall optimization objective of our OpenESS framework is to minimize $\mathcal{L} = \mathcal{L}_{F2E} + \alpha \mathcal{L}_{T2E}$, where α is a weight balancing coefficient.

Role in Our Framework. Our T2E semantic consistency regularization provides a global-level alignment to compensate for the possible self-conflict in the superpixel-driven frame-to-event contrastive learning. As we will show in the following sections, the two objectives work synergistically in improving the performance of open-vocabulary ESS.

Inference-Time Configuration. Our OpenESS framework is designed to pursue segmentation accuracy in annotation-free and annotation-efficient manners, without sacrificing event processing efficiency. As can be seen from Fig. 2, after the cross-modality knowledge transfer, only the event branch will be kept. This guarantees that there will be no extra latency or power consumption added during the inference, which is in line with the practical requirements.

4. Experiments

4.1. Settings

Datasets. We conduct experiments on two popular ESS datasets. ***DDD17-Seg*** [2] is a widely used ESS benchmark consisting of 40 sequences acquired by a DAVIS346B. In total, 15950 training and 3890 testing events of spatial size 352 × 200 are used, along with synchronized gray-scale frames provided by the DAVIS camera. ***DSEC-Semantic*** [79] provides semantic labels for 11 sequences in the DSEC [32] dataset. The training and testing splits contain 8082 and 2809 events of spatial size 640 × 440, accompanied by color frames (with sensor calibration parameters available) recorded at 20Hz. More details are in the Appendix.

Benchmark Setup. In addition to the conventional fully-supervised ESS, we establish two open-vocabulary ESS settings for *annotation-free* and *annotation-efficient* learning, respectively. The former aims to train an ESS model without using any dense event labels, while the latter assumes an annotation budget of 1%, 5%, 10%, or 20% of events in the training set. We treat the first few samples from each sequence as labeled and the remaining ones as unlabeled.

Implementation Details. Our framework is implemented using PyTorch [66]. Based on the use of event representations, we form `frame2voxel`, `frame2recon`, and `frame2spike` settings, where the event branch will adopt E2VID [73], ResNet-50 [39], and SpikingFCN [49], respectively, with an AdamW [59] optimizer with cosine learning rate scheduler. The frame branch uses a pre-trained ResNet-50 [8, 9, 15] and is kept frozen. The number of superpixels

Table 1. **Comparative study** of existing ESS approaches under the annotation-free, fully-supervised, and open-vocabulary ESS settings, respectively, on the *test* sets of the *DDD17-Seg* [5] and *DSEC-Semantic* [79] datasets. All scores are in percentage (%). The **best** score from each learning setting is highlighted in **bold**.

Method	Venue	DDD17		DSEC	
		Acc	mIoU	Acc	mIoU
Annotation-Free ESS					
MaskCLIP [100]	ECCV'22	81.29	31.90	58.96	21.97
FC-CLIP [97]	NeurIPS'23	88.66	51.12	79.20	39.42
OpenESS	**Ours**	**90.51**	**53.93**	**86.18**	**43.31**
Fully-Supervised ESS					
Ev-SegNet [2]	CVPRW'19	89.76	54.81	88.61	51.76
E2VID [73]	TPAMI'19	85.84	48.47	80.06	44.08
Vid2E [30]	CVPR'20	90.19	56.01	-	-
EVDistill [84]	CVPR'21	-	58.02	-	-
DTL [83]	ICCV'21	-	58.80	-	-
PVT-FPN [86]	ICCV'21	94.28	53.89	-	-
SpikingFCN [49]	NCE'22	-	34.20	-	-
EV-Transfer [61]	RA-L'22	51.90	15.52	63.00	24.37
ESS [79]	ECCV'22	88.43	53.09	84.17	45.38
ESS-Sup [79]	ECCV'22	91.08	**61.37**	89.37	53.29
P2T-FPN [91]	TPAMI'23	94.57	54.64	-	-
EvSegformer [47]	TIP'23	**94.72**	54.41	-	-
HMNet-B [38]	CVPR'23	-	-	88.70	51.20
HMNet-L [38]	CVPR'23	-	-	**89.80**	**55.00**
HALSIE [6]	WACV'24	92.50	60.66	89.01	52.43
Open-Vocabulary ESS					
MaskCLIP [100]	ECCV'22	90.50	61.27	89.81	55.01
FC-CLIP [97]	NeurIPS'23	90.68	62.01	89.97	55.67
OpenESS	**Ours**	**91.05**	**63.00**	**90.21**	**57.21**

involved in the calculation of F2E contrastive loss is set to 100 for *DSEC-Semantic* [79] and 25 for *DDD17-Seg* [2]. For evaluation, we extract the feature embedding for each text prompt offline from a frozen CLIP text encoder using pre-defined templates. For linear probing, the pre-trained event network $\mathcal{F}^{evt}_{\theta_e}$ is kept frozen, followed by a trainable point-wise linear classification head. Due to space limits, kindly refer to our Appendix for additional details.

4.2. Comparative Study

Annotation-Free ESS. In Tab. 1, we compare OpenESS with MaskCLIP [100] and FC-CLIP [97] in the absence of event labels. Our approach achieves zero-shot ESS results of 53.93% and 43.31% on *DDD17-Seg* [2] and *DSEC-Semantic* [79], much higher than the two competitors and even comparable to some fully-supervised methods. This validates the effectiveness of conducting ESS in an annotation-free manner for practical usage. Meanwhile, we observe that a fine-tuned CLIP encoder [97] could generate much better semantic predictions than the structure adaptation method [100], as mentioned in Sec. 3.2.

Comparisons to State-of-the-Art Methods. As shown in Tab. 1, the proposed OpenESS sets up several new state-of-the-art results in the two ESS benchmarks. Compared to the previously best-performing methods, OpenESS is 1.63% and 2.21% better in terms of mIoU scores on *DDD17-Seg* [2] and *DSEC-Semantic* [79], respectively. It is worth mentioning that in addition to the performance improvements, our approach can generate open-vocabulary predictions that are beyond the closed sets of predictions of existing methods, which is more in line with the practical usage.

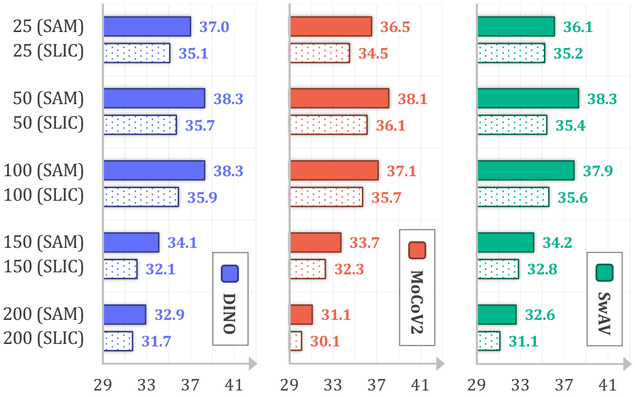


Figure 3. **Ablation study** on the number of superpixels (provided by either SAM [50] or SLIC [1]) involved in calculating the frame-to-event contrastive loss. Models after pre-training are fine-tuned with 1% annotations. All mIoU scores are in percentage (%).

Annotation-Efficient Learning. We establish a comprehensive benchmark for ESS under limited annotation scenarios and show the results in Tab. 3. As can be seen, the proposed OpenESS contributes significant performance improvements over random initialization under linear probing, few-shot fine-tuning, and fully-supervised learning settings. Specifically, using either voxel grid or event reconstruction representation, our approach achieves $>30\%$ relative gains in mIoU on both datasets under liner probing and around 2% higher than prior art in mIoU with full supervisions. We also observe that using voxel grids to represent raw event streams tends to yield overall better ESS performance.

Qualitative Assessment. Fig. 4 provides visual comparisons between OpenESS and other approaches on *DSEC-Semantic* [79]. We find that OpenESS tends to predict more consistent semantic information from sparse and irregular event inputs, especially at instance boundaries. We include more visual examples and failure cases in the Appendix.

Open-World Predictions. One of the core advantages of OpenESS is the ability to predict beyond the fixed label set from the original training sets. As shown in Fig. 1, our approach can take arbitrary text prompts as inputs and generate semantically coherent event predictions without using event labels. This is credited to the alignment between event features and CLIP's knowledge in T2E. Such a flexible way of prediction enables a more holistic event understanding.

Other Representation Learning Approaches. In Tab. 2, we compare OpenESS with recent reconstruction-based [3,

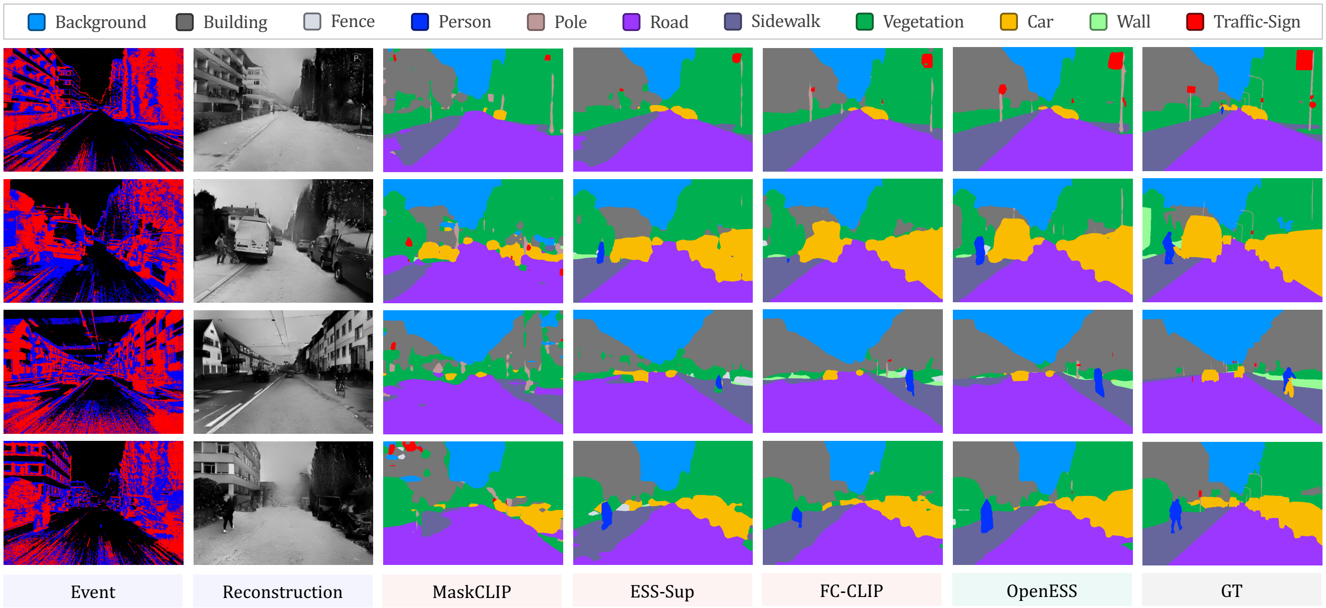


Figure 4. **Qualitative comparisons** of state-of-the-art ESS approaches on the *test* set of *DSEC-Semantic* [79]. Each color corresponds to a distinct semantic category. GT denotes the ground truth semantic maps. Best viewed in colors and zoomed-in for additional details.

Table 2. **Comparative study** of different representation learning methods applied on event data. **OV** denotes whether supporting open-vocabulary predictions. All mIoU scores are in percentage (%). The **best** score from each dataset is highlighted in **bold**.

Method	Venue	Backbone	OV	DDD17	DSEC
Random	-	ViT-S/16	✗	48.76	40.53
MoCoV3 [16]	ICCV'21	ViT-S/16	✗	53.65	49.21
IBoT [101]	ICLR'22	ViT-S/16	✗	49.94	42.53
ECDP [95]	ICCV'23	ViT-S/16	✗	54.66	47.91
Random	-	ViT-B/16	✗	43.89	38.24
BeiT [3]	ICLR'22	ViT-B/16	✗	52.39	46.52
MAE [40]	CVPR'22	ViT-B/16	✗	52.36	47.56
Random	-	ResNet-50	✗	56.96	57.60
SimCLR [14]	ICML'20	ResNet-50	✗	57.22	59.06
ECDP [95]	ICCV'23	ResNet-50	✗	59.15	**59.16**
Random	-	ResNet-50	✗	55.56	52.86
OpenESS	**Ours**	ResNet-50	✓	57.01	55.01
Random	-	E2VID	✗	61.06	54.96
OpenESS	**Ours**	E2VID	✓	**63.00**	57.21

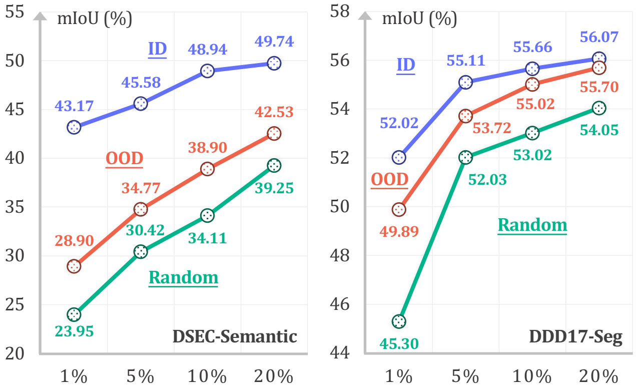


Figure 5. **Cross-dataset representation learning** results of comparing OpenESS pre-training using in-distribution (ID) and out-of-distribution (OOD) data in-between the *DDD17-Seg* [5] and *DSEC-Semantic* [79] datasets. Models after pre-training are fine-tuned with 1%, 5%, 10%, and 20% annotations, respectively.

40, 95, 101] and contrastive learning-based [14, 16] pre-training methods. As can be seen, the proposed OpenESS achieves competitive results over existing approaches. It is worth highlighting again that our framework distinct from prior arts by supporting open-vocabulary learning.

4.3. Ablation Study

Cross-Modality Representation Learning. Tab. 4 provides a comprehensive ablation study on the frame-to-event (F2E) and text-to-event (T2E) learning objectives in OpenESS using three event representations. We observe that both F2E and T2E contribute to an overt improvement over random initialization under linear probing and few-shot fine-tuning settings, which verifies the effectiveness of our proposed approach. Once again, we find that the voxel grids tend to achieve better performance than other representations. The spike-based methods [49], albeit being computationally more efficient, show sub-par performance compared to voxel grids and reconstructions.

Superpixel Generation. We study the utilization of SLIC [1] and SAM [50] in our frame-to-event contrastive distillation and show the results in Fig. 3. Using either frame net-

Table 3. **Comparative study** of different open-vocabulary semantic segmentation methods [97, 100] under the linear probing (LP) and few-shot fine-tuning, and full supervision (Full) settings, respectively, on the *test* sets of the *DDD17-Seg* [5] and *DSEC-Semantic* [79] datasets. All mIoU scores are given in percentage (%). The **best** mIoU scores from each learning configuration are highlighted in **bold**.

Method	**Configuration**	**DSEC-Semantic**						**DDD17-Seg**					
		LP	1%	5%	10%	20%	Full	LP	1%	5%	10%	20%	Full
Random	Voxel Grid	6.70	26.62	31.22	33.67	41.31	54.96	12.30	52.13	54.87	58.66	59.52	61.06
MaskCLIP [100]		33.08	33.89	37.03	38.83	42.40	55.01	31.91	53.91	56.27	59.32	59.97	61.27
FC-CLIP [97]	Voxel Grid	43.00	39.12	43.71	44.09	47.77	55.67	54.07	56.38	58.50	60.05	60.85	62.01
OpenESS (Ours)	frame2voxel	**44.26**	**41.41**	**44.97**	**46.25**	**48.28**	**57.21**	**55.61**	**57.58**	**59.07**	**61.03**	**61.78**	**63.00**
Improve ↑		+33.56	+14.79	+13.75	+12.58	+6.97	+2.25	+43.31	+5.45	+4.20	+2.37	+2.26	+1.94
Random	Reconstruction	6.22	23.95	30.42	34.11	39.25	52.86	13.89	45.30	52.03	53.02	54.05	55.56
MaskCLIP [100]		27.09	30.73	36.33	40.13	43.37	52.97	29.81	49.02	53.65	54.11	54.75	56.12
FC-CLIP [97]	Reconstruction	40.08	38.99	43.34	45.35	47.18	53.05	52.17	51.01	54.09	54.99	55.05	56.34
OpenESS (Ours)	frame2recon	**44.08**	**43.17**	**45.58**	**48.94**	**49.74**	**55.01**	**53.61**	**52.02**	**55.11**	**55.66**	**56.07**	**57.01**
Improve ↑		+37.86	+19.22	+15.16	+14.83	+10.49	+2.15	+39.72	+6.72	+3.08	+2.64	+2.02	+1.45

Table 4. **Ablation study** of OpenESS under linear probing (LP) and few-shot fine-tuning settings from three learning configurations on the *test* set of *DDD17-Seg* [5]. **F2E** denotes the frame-to-event contrastive learning. **T2E** denotes the text-to-event semantic regularization. All mIoU scores are given in percentage (%).

Configuration	**F2E**	**T2E**	**DDD17-Seg**				
			LP	1%	5%	10%	20%
Voxel Grid	Random		12.30	52.13	54.87	58.66	59.52
frame2voxel	✓		52.60	55.41	57.07	59.77	60.21
		✓	54.11	56.77	58.95	60.12	60.99
	✓	✓	55.61	57.58	59.07	61.03	61.78
Reconstruction	Random		13.89	45.30	52.03	53.02	54.05
frame2recon	✓		50.21	50.96	53.67	54.21	54.92
		✓	52.62	51.63	54.27	55.00	55.17
	✓	✓	53.61	52.02	55.11	55.66	56.07
Spike	Random		12.04	10.01	20.02	25.81	26.03
frame2spike	✓		15.07	14.31	21.77	26.89	27.07
		✓	16.11	14.67	22.61	27.97	29.01
	✓	✓	16.27	14.89	23.54	28.51	29.98

works pre-trained by DINO [9], MoCoV2 [15], or SwAV [8], the SAM-generated superpixels consistently exhibit better performance for event representation learning. The number of superpixels involved in calculating tends to affect the effectiveness of contrastive learning. A preliminary search to determine this hyperparameter is required. We empirically find that setting M to 100 for *DSEC-Semantic* [79] and 25 for *DDD17-Seg* [2] will likely yield the best possible segmentation performance in our framework.

Cross-Dataset Knowledge Transfer. Since we are targeting annotation-free representation learning, it is thus intuitive to see the cross-dataset adaptation effect. As shown in Fig. 5, pre-training on OOD datasets also brings appealing improvements over the random initialization baseline. This result highlights the importance of conducting representation learning for an effective transfer to downstream tasks.

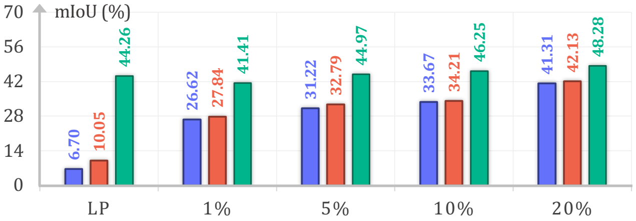


Figure 6. **Single-modality OpenESS representation learning study** on the *DSEC-Semantic* [79] dataset. The results are from models of random initialization (🟦), recon2voxel pre-training (🟥), and frame2voxel pre-training (🟩), respectively, after linear probing (LP) and annotation-efficient fine-tuning.

Framework with Event Camera Only. Lastly, we study the scenario where the frame camera becomes unavailable. We replace the input to the frame branch with event reconstructions [73] and show the results in Fig. 6. Since the limited visual cues from the reconstruction tend to degrade the quality of representation learning, its performance is subpar compared to the frame-based knowledge transfer.

5. Conclusion

In this work, we introduced OpenESS, an open-vocabulary event-based semantic segmentation framework tailored to perform open-vocabulary ESS in an annotation-efficient manner. We proposed to encourage cross-modality representation learning between events and frames using frame-to-event contrastive distillation and text-to-event semantic consistency regularization. Through extensive experiments, we validated the effectiveness of OpenESS in tackling dense event-based predictions. We hope this work could shed light on the future development of more scalable ESS systems.

Acknowledgement. This work is under the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

Appendix

A. Additional Implementation Details

In this section, we provide additional details to assist the implementation and reproduction of the approaches in the proposed OpenESS framework.

A.1. Datasets

In this study, we follow prior works [2, 38, 47, 79] by using the ***DDD17-Seg*** [2] and ***DSEC-Semantic*** [79] datasets for evaluating and validating the baselines, prior methods, and the proposed OpenESS framework. Some specifications related to these two datasets are listed as follows.

- **DDD17-Seg** [2] serves as the first benchmark for ESS. It is a semantic segmentation extension of the DDD17 [5] dataset, which includes hours of driving data, capturing a variety of driving conditions such as different times of day, traffic scenarios, and weather conditions. Alonso and Murillo [2] provide the semantic labels on top of DDD17 to enable event-based semantic segmentation. Specifically, they proposed to use the corresponding gray-scale images along with the event streams to generate an approximated set of semantic labels for training, which was proven effective in training models to segment directly on event-based data. A three-step procedure is applied: *i)* train a semantic segmentation model on the gray-scale images in the *Cityscapes* dataset [19]; *ii)* Use the trained model to label the gray-scale images in DDD17; and *iii)* Conduct a post-processing step on the generated pseudo labels, including class merging and image cropping. The dataset specification is shown in Tab. 5. In total, there are 15950 training and 3890 test samples in the DDD17-Seg dataset. Each pixel is labeled across six semantic classes, including `flat`, `background`, `object`, `vegetation`, `human`, and `vehicle`. For each sample, we convert the event streams into a sequence of 20 voxel grids, each consisting of 32000 events and with a spatial resolution of 352 × 200. For additional details of this dataset, kindly refer to `http://sensors.ini.uzh.ch/news_page/DDD17.html`.
- **DSEC-Semantic** [79] is a semantic segmentation extension of the DSEC (Driving Stereo Event Camera) dataset [32]. DSEC is an extensive dataset designed for advanced driver-assistance systems (ADAS) and autonomous driving research, with a particular focus on event-based vision and stereo vision. Different from DDD17 [5], the DSEC dataset combines data from event-based cameras and traditional RGB cameras. The inclusion of event-based cameras (which capture changes in light intensity) alongside regular cameras provides a rich, complementary data source for perception tasks. The dataset typically features high-resolution images and event data, providing detailed visual information from a wide range of driving conditions, including urban, suburban, and highway environments, various weather conditions, and different times of the day. This diversity is crucial for developing systems that can operate reliably in real-world conditions. Based on such a rich collection, Sun *et al.* [79] adopted a similar pseudo labeling procedure as DDD17-Seg [2] and generated the semantic labels for eleven sequences in DSEC, dubbed as DSEC-Semantic. The dataset specification is shown in Tab. 6. In total, there are 8082 training and 2809 test samples in the DSEC-Semantic dataset. Each pixel is labeled across eleven semantic classes, including `background`, `building`, `fence`, `person`, `pole`, `road`, `sidewalk`, `vegetation`, `car`, `wall`, and `traffic-sign`. For each sample, we convert the event streams into a sequence of 20 voxel grids, each consisting of 100000 events and with a spatial resolution of 640 × 440. For additional details of this dataset, kindly refer to `https://dsec.ifi.uzh.ch/dsec-semantic`.

A.2. Text Prompts

To enable the conventional evaluation of our proposed open-vocabulary approach on an event-based semantic segmentation dataset, we need to use the pre-defined class names as text prompts to generate the text embedding. Specifically, we follow the standard templates [69] when generating the embedding. The dataset-specific text prompts defined in our framework are listed as follows.

- **DDD17-Seg.** There is a total of six semantic classes in

Table 5. The specifications of the *DDD17-Seg* dataset [2].

-	Training					Test
Seq	dir0	dir3	dir4	dir6	dir7	dir1
# Frames	11785	20051	41071	28411	58650	71680
# Events	5550	1320	6945	1140	995	3890
Resolution	352 × 200					352 × 200
# Classes	6 Classes					6 Classes

Table 6. The specifications of the *DSEC-Semantic* dataset [79].

-	Training								Test		
Seq	00_a	01_a	02_a	04_a	05_a	06_a	07_a	08_a	13_a	14_c	15_a
# Frames	939	681	235	701	1753	1523	1463	787	379	1191	1239
# Events	933	675	229	695	1747	1517	1457	781	373	1185	1233
Resolution	640 × 440								640 × 440		
# Classes	11 Classes								11 Classes		

the DDD17-Seg dataset [2], with static and dynamic components of driving scenes. Our defined text prompts of this dataset are summarized in Tab. 7. For each semantic class, we generate for each text prompt the text embedding using the CLIP text encoder and then average the text embedding of all text prompts as the final embedding of this class.

- **DSEC-Semantic.** There is a total of eleven semantic classes in the DSEC-Semantic dataset [79], ranging from static and dynamic components of driving scenes. Our defined text prompts of this dataset are summarized in Tab. 8. For each semantic class, we generate for each text prompt the text embedding using the CLIP text encoder and then average the text embedding of all text prompts as the final embedding of this class.

A.3. Superpixels

In image processing and computer vision, superpixels can be defined as a scheme that groups pixels in an image into perceptually meaningful atomic regions, which are used to replace the rigid structure of the pixel grid [1]. Superpixels provide a more natural representation of the image structure, often leading to more efficient and effective image processing. Here are some of their key aspects:

- **Grouping Pixels.** Superpixels are often formed by clustering pixels based on certain criteria like color similarity, brightness, texture, and other low-level patterns [1], or more recently, semantics [50]. This results in contiguous regions in the image that are more meaningful than individual pixels for many applications [13, 57, 67, 94].
- **Reducing Complexity.** By aggregating pixels into superpixels, the complexity of image data is significantly reduced [78]. This reduction helps in speeding up subsequent image processing tasks, as algorithms have fewer elements (superpixels) to process compared to the potentially millions of pixels in an image.
- **Preserving Edges.** One of the primary goals of superpixel segmentation is to preserve important image edges. Superpixels often adhere closely to the boundaries of objects in the image, making them useful for tasks that rely on accurate edge information, like object recognition and scene understanding.

In this work, we propose to first leverage calibrated frames to generate coarse, instance-level superpixels and then distill knowledge from a pre-trained image backbone to the event segmentation network. Specifically, we resort to the following two ways to generate the superpixels.

- **SLIC.** The first way is to leverage the heuristic Simple Linear Iterative Clustering (SLIC) approach [1] to efficiently group pixels from frame I_i^{img} into a total of M_{slic} segments with good boundary adherence and regularity. The superpixels are defined as $I_i^{sp} = \{\mathcal{I}_i^1, \mathcal{I}_i^2, ..., \mathcal{I}_i^{M_{slic}}\}$, where M_{slic} is a hyperparameter that needs to be adjusted based on the inputs. The generated superpixels satisfy $\mathcal{I}_i^1 \cup \mathcal{I}_i^2 \cup ... \cup \mathcal{I}_i^{M_{slic}} = \{1, 2, ..., H \times W\}$. Several examples of the SLIC-generated superpixels are shown in the second row of Fig. 7, where each of the color-coded patches represents one distinct and semantically coherent superpixel.
- **SAM.** For the second option, we use the recent Seg-

Table 7. The text prompts defined on the *DDD17-Seg* dataset [2] (6 classes) used for generating the CLIP text embedding.

DDD17 (6 classes)		
#	**class**	**text prompt**
0	`flat`	‘road’, ‘driveable’, ‘street’, ‘lane marking’, ‘bicycle lane’, ‘roundabout lane’, ‘parking lane’, ‘terrain’, ‘grass’, ‘soil’, ‘sand’, ‘lawn’, ‘meadow’, ‘turf’
1	`background`	‘sky’, ‘building’
2	`object`	‘pole’, ‘traffic sign pole’, ‘traffic light pole’, ‘traffic light box’, ‘traffic-sign’, ‘parking-sign’, ‘direction-sign’
3	`vegetation`	‘vegetation’, ‘vertical vegetation’, ‘tree’, ‘tree trunk’, ‘hedge’, ‘woods’, ‘terrain’, ‘grass’, ‘soil’, ‘sand’, ‘lawn’, ‘meadow’, ‘turf’
4	`human`	‘person’, ‘pedestrian’, ‘walking people’, ‘standing people’, ‘sitting people’, ‘toddler’
5	`vehicle`	‘car’, ‘jeep’, ‘SUV’, ‘van’, ‘caravan’, ‘truck’, ‘box truck’, ‘pickup truck’, ‘trailer’, ‘bus’, ‘public bus’, ‘train’, ‘vehicle-on-rail’, ‘tram’, ‘motorbike’, ‘moped’, ‘scooter’, ‘bicycle’

Table 8. The ext prompts defined on the *DSEC-Semantic* dataset [79] (11 classes) used for generating the CLIP text embedding.

DSEC-Semantic (11 classes)		
#	**class**	**text prompt**
0	`background`	‘sky’
1	`building`	‘building’, ‘skyscraper’, ‘house’, ‘bus stop building’, ‘garage’, ‘carport’, ‘scaffolding’
2	`fence`	‘fence’, ‘fence with hole’
3	`person`	‘person’, ‘pedestrian’, ‘walking people’, ‘standing people’, ‘sitting people’, ‘toddler’
4	`pole`	‘pole’, ‘electric pole’, ‘traffic sign pole’, ‘traffic light pole’
5	`road`	‘road’, ‘driveable’, ‘street’, ‘lane marking’, ‘bicycle lane’, ‘roundabout lane’, ‘parking lane’
6	`sidewalk`	‘sidewalk’, ‘delimiting curb’, ‘traffic island’, ‘walkable’, ‘pedestrian zone’
7	`vegetation`	‘vegetation’, ‘vertical vegetation’, ‘tree’, ‘tree trunk’, ‘hedge’, ‘woods’, ‘terrain’, ‘grass’, ‘soil’, ‘sand’, ‘lawn’, ‘meadow’, ‘turf’
8	`car`	‘car’, ‘jeep’, ‘SUV’, ‘van’, ‘caravan’, ‘truck’, ‘box truck’, ‘pickup truck’, ‘trailer’, ‘bus’, ‘public bus’, ‘train’, ‘vehicle-on-rail’, ‘tram’, ‘motorbike’, ‘moped’, ‘scooter’, ‘bicycle’
9	`wall`	‘wall’, ‘standing wall’
10	`traffic-sign`	‘traffic-sign’, ‘parking-sign’, ‘direction-sign’, ‘traffic-sign without pole’, ‘traffic light box’

ment Anything Model (SAM) [50] which takes I_i^{img} as the input and outputs M_{sam} class-agnostic masks. For simplicity, we use M to denote the number of superpixels used during knowledge distillation, *i.e.*, $\{I_i^{sp} = \{\mathcal{I}_i^1, ..., \mathcal{I}_i^k\} | k = 1, ..., M\}$. Several examples of the SAM-generated superpixels are shown in the third row of Fig. 7, where each of the color-coded patches represents one distinct and semantically coherent superpixel.

We calculate the SLIC and SAM superpixel distributions on the training set of the DSEC-Semantic dataset [79] and show the corresponding statistics in Fig. 8. As can be observed, the SLIC-generated superpixels often contain

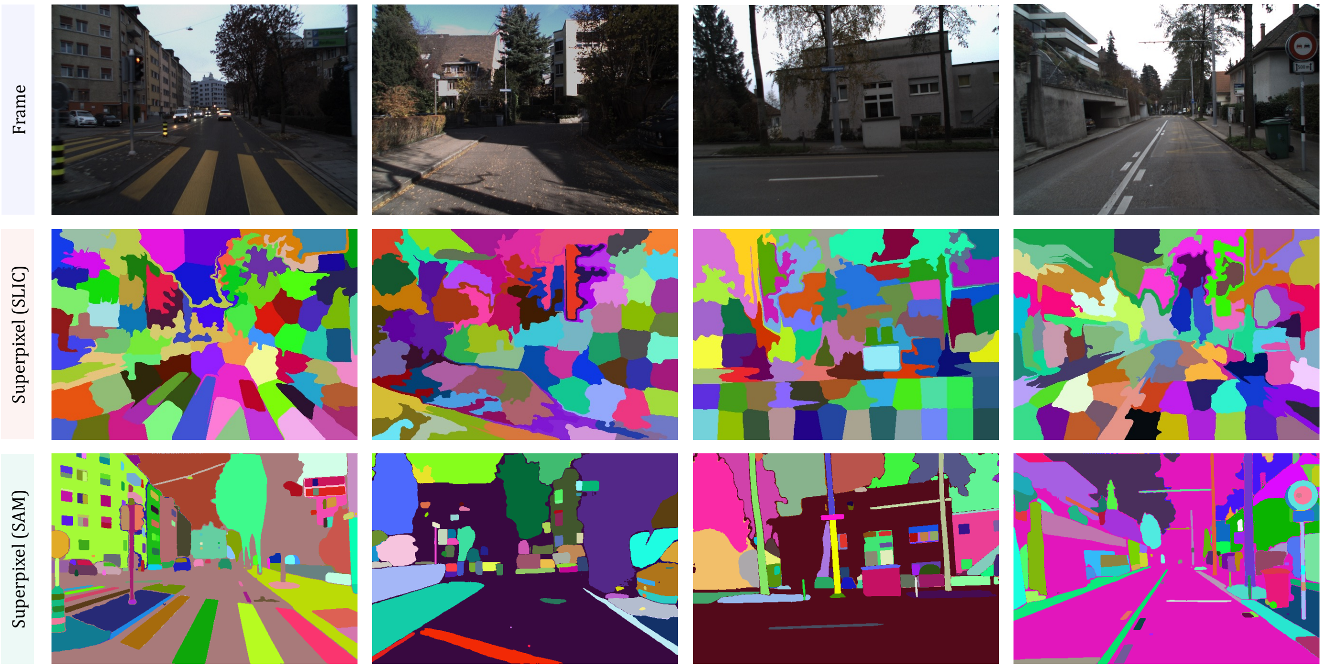
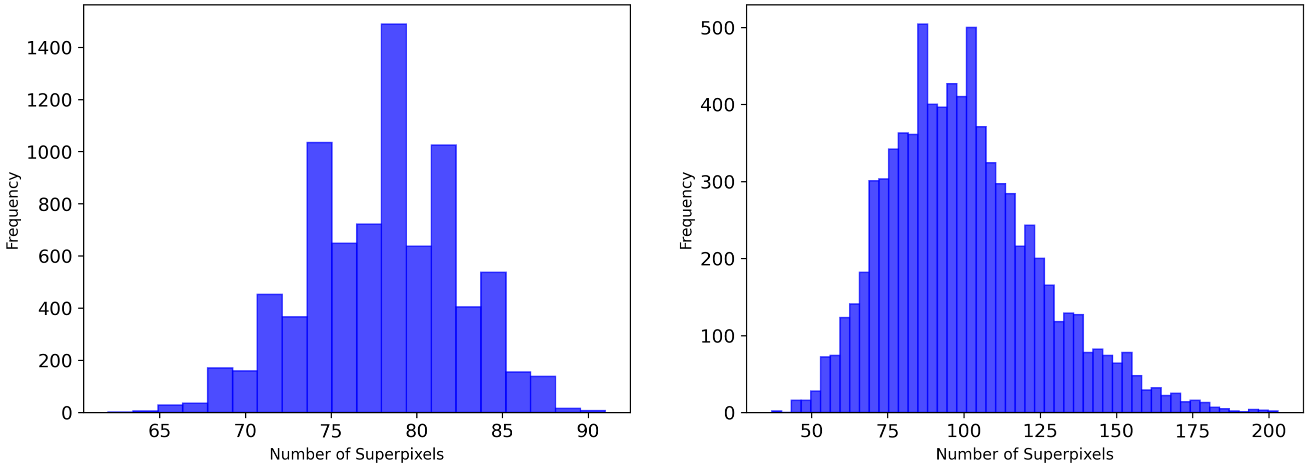


Figure 7. **Examples of superpixels** generated by SLIC [1] (the 2nd row) and SAM [50] (the 3rd row). The parameter M_{slic} in the SLIC algorithm is set to 100. Each colored patch represents one distinct and semantically coherent superpixel. Best viewed in colors.



(a) Histogram of SLIC-Generated Superpixels

(b) Histogram of SAM-Generated Superpixels

Figure 8. **The statistical distributions** of superpixels generated by SLIC [1] (subfigure a) and SAM [50] (subfigure b).

more low-level visual cues, such as color similarity, brightness, and texture. On the contrary, superpixels generated by SAM exhibit clear semantic coherence and often depict the boundaries of objects and backgrounds. As verified in the main body of this paper, the semantically richer SAM superpixels bring higher performance gains in our Frame-to-Event Contrastive Learning framework.

Meanwhile, we provide more fine-grained examples of the SLIC algorithm using different M_{slic}, *i.e.*, 25, 50, 100, 150, and 200. The results are shown in Fig. 9. Specifically, the number of superpixels M_{slic} should reflect the complexity and detail of the image. For images with high detail or complexity (like those with many objects or textures), a larger M_{slic} can capture more of this detail. Conversely, for simpler images, fewer superpixels might be sufficient. Usually, more superpixels mean smaller superpixels. Smaller superpixels can adhere more closely to object boundaries and capture finer details, but they might also capture more noise. Fewer superpixels result in larger, more homogeneous regions but may lead to a loss of detail, especially

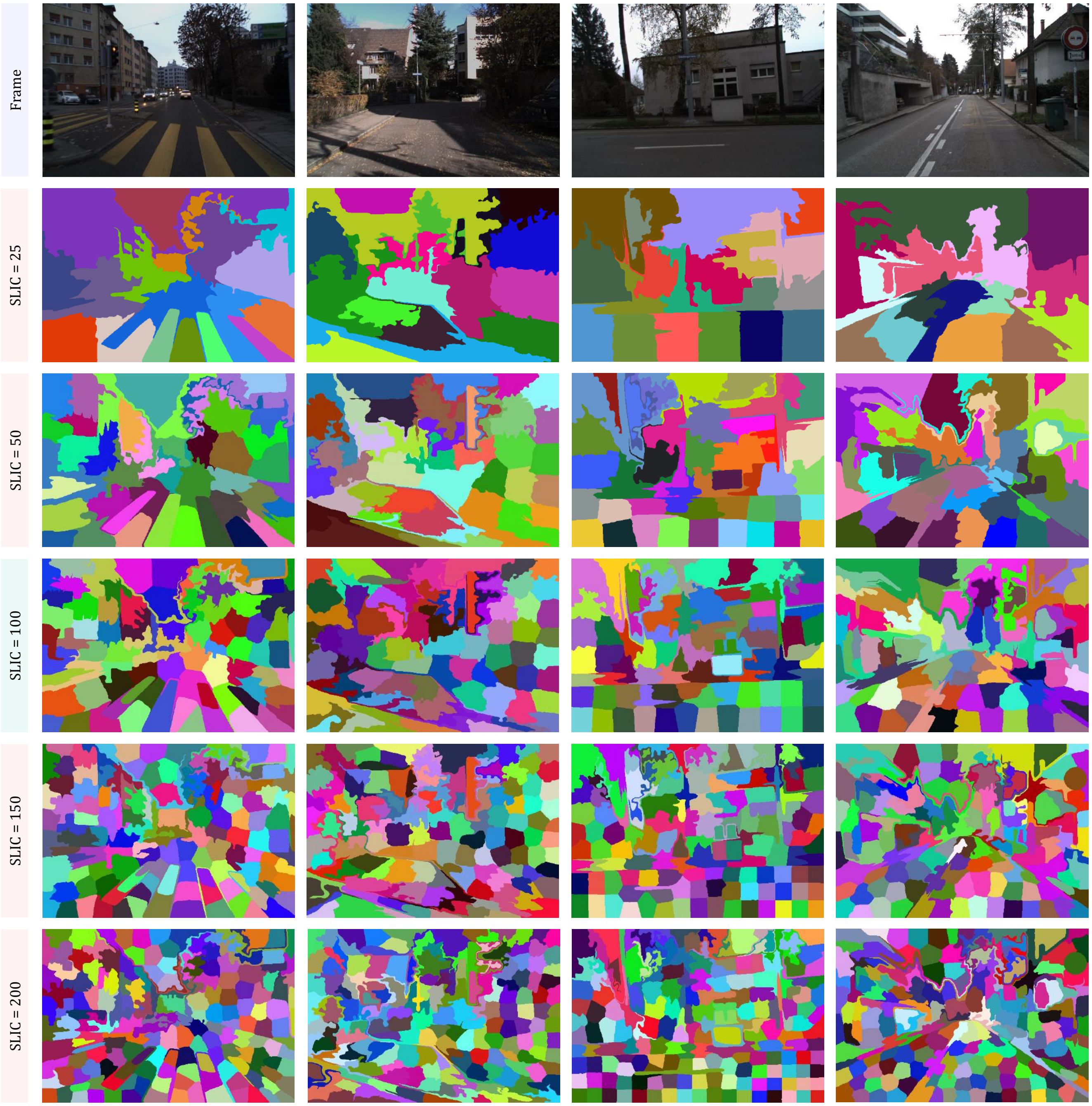


Figure 9. **Examples of superpixels** generated by SLIC [1] with different numbers of superpixels M_{slic} (25, 50, 100, 150, and 200). Each colored patch represents one distinct and semantically coherent superpixel. Best viewed in colors.

at the edges of objects. The choice also depends on the specific application. For instance, in object detection or segmentation tasks where boundary adherence is crucial, a higher number of superpixels might be preferable. In contrast, for tasks like image compression or abstraction, fewer superpixels might be more appropriate. Often, the optimal number of superpixels is determined empirically. This involves experimenting with different values and evaluating the results based on the specific criteria of the task or application. In our event-based semantic segmentation task, we choose $M_{slic} = 100$ for our Frame-to-Event Contrastive Learning on the DSEC-Semantic dataset [79], and $M_{slic} = 25$ on the DDD17-Seg dataset [2].

Since I_i^{evt} and I_i^{img} have been aligned and synchronized, we can group events from I_i^{evt} into superevents $\{V_i^{sp} = \{\mathcal{V}_i^1, ..., \mathcal{V}_i^l\} | l = 1, ..., M\}$ by using the known event-pixel correspondences.

A.4. Backbones

As mentioned in the main body of this paper, we establish three open-vocabulary event-based semantic segmentation settings based on the use of three different event representations, *i.e.*, `frame2voxel`, `frame2recon`, and `frame2spike`. It is worth noting that these three event representations tend to have their own advantages.

We supplement additional implementation details regarding the used event representations as follows.

- **Frame2Voxel.** For the use of *voxel grids* as the event embedding, we follow Sun *et al.* [79] by converting the raw events ε_i into the regular voxel grids $I_i^{vox} \in \mathbb{R}^{C \times H \times W}$ as the input to the event-based semantic segmentation network. This representation is intuitive and aligns well with conventional event camera data processing techniques. It is suitable for convolutional neural networks as it maintains spatial and temporal relationships. Specifically, with a predefined number of events, each voxel grid is built from non-overlapping windows as follows:
$$I_i^{vox} = \sum_{\mathbf{e}_j \in \varepsilon_i} p_j \delta(\mathbf{x}_j - \mathbf{x}) \delta(\mathbf{y}_j - \mathbf{y}) \max\{1 - |t_j^* - t|, 0\}, \tag{8}$$
where δ is the Kronecker delta function; $t_j^* = (B - 1)\frac{t_j - t_0}{\Delta T}$ is the normalized event timestamp with B as the number of temporal bins in an event stream; ΔT is the time window and t_0 denotes the time of the first event in the window. It is worth noting that *voxel grids* can be memory-intensive, especially for high-resolution sensors or long-time windows. They might also introduce quantization errors due to the discretization of space and time. For additional details on the use of *voxel grids*, kindly refer to https://github.com/uzh-rpg/ess.
- **Frame2Recon.** For the use of *event reconstructions* as the event embedding, we follow Sun *et al.* [79] and Rebecq *et al.* [73] by converting the raw events ε_i into the regular frame-like event reconstructions $I_i^{rec} \in \mathbb{R}^{H \times W}$ as the input to the event-based semantic segmentation network. This can be done by accumulating events over short time intervals or by using algorithms to interpolate or simulate frames. This approach is compatible with standard image processing techniques and algorithms developed for frame-based vision. It is more familiar to practitioners used to working with conventional cameras. In this work, we adopt the E2VID model [73] to generate the *event reconstructions*. This process can be described as follows:
$$\mathbf{z}_k^{rec} = E_{\text{e2vid}}(I_k^{vox}, \mathbf{z}_{k-1}^{rec}), \quad k = 1, ..., N, \tag{9}$$
$$I_i^{rec} = D_{\text{e2vid}}(\mathbf{z}^{rec}), \tag{10}$$
where I_k^{vox} denotes the *voxel grids* as defined in Eq. (8); E_{e2vid} and D_{e2vid} are the encoder of decoder of the E2VID model [73], respectively. It is worth noting that *event reconstructions* can lose the fine temporal resolution that event cameras provide. They might also introduce artifacts or noise, especially in scenes with fast-moving objects or low event rates. For additional details on the use of *event reconstructions*, kindly refer to https://github.com/uzh-rpg/rpg_e2vid.
- **Frame2Spike.** For the use of *spikes* as the event embedding, we follow Kim *et al.* [49] by converting the raw events ε_i into spikes $I_i^{spk} \in \mathbb{R}^{H \times W}$ as the input to the event-based semantic segmentation network. The spike representation keeps the data in its raw form – as individual spikes or events. This representation preserves the high temporal resolution of the event data and is highly efficient in terms of memory and computation, especially for sparse scenes. The rate coding is used as the spike encoding scheme due to its reliable performance across various tasks. Each pixel value with a random number ranging between $[s_{min}, s_{max}]$ at every time step is recorded, where s_{min} and s_{max} are the minimum and maximum possible pixel intensities, respectively. If the random number is greater than the pixel intensity, the Poisson spike generator outputs a spike with amplitude 1. Otherwise, the Poisson spike generator does not yield any spikes. The spikes in a certain time window are accumulated to generate a frame, where such frames will serve as the input to the event-based semantic segmentation network. It is worth noting that processing raw spike data requires specialized algorithms, often inspired by neuromorphic computing. It might not be suitable for traditional image processing techniques and can be challenging to interpret and visualize. For additional details on the use of *spikes*, kindly refer to https://github.com/Intelligent-Computing-Lab-Yale/SNN-Segmentation.

To sum up, each event representation has its unique characteristics and is suitable for different applications or processing techniques. Our proposed OpenESS framework is capable of leveraging each of the above event representations for efficient and accurate event-based semantic segmentation in an annotation-free and open-vocabulary manner. Such a versatile and flexible way of learning verifies the broader application potential of our proposed framework.

A.5. Evaluation Configuration

Following the convention, we use the Intersection-over-Union (`IoU`) metric to measure the semantic segmentation performance for each semantic class. The IoU score can be calculated via the following equation:

$$\texttt{IoU} = \frac{TP}{TP + FP + FN}, \tag{11}$$

Table 9. **The per-class segmentation results** of annotation-free event-based semantic segmentation approaches on the test set of *DSEC-Semantic* [79]. Scores reported are IoUs in percentage (%). For each semantic class, the best score in each column is highlighted in **bold**.

Method	**mIoU**	background	building	fence	person	pole	road	sidewalk	vegetation	car	wall	traffic-sign	**Acc**
Annotation-Free ESS													
MaskCLIP [100]	21.97	26.45	52.59	0.20	0.04	**4.19**	65.76	2.96	48.02	40.67	0.67	0.08	58.96
FC-CLIP [97]	39.42	87.49	69.68	**14.39**	**17.53**	0.29	71.76	34.56	71.30	63.19	2.98	0.50	79.20
OpenESS (Ours)	**43.31**	**92.53**	**74.22**	11.96	0.00	0.41	**87.32**	**55.09**	**74.23**	**64.25**	**7.98**	**8.47**	**86.18**

Table 10. **The per-class segmentation results** of annotation-efficient event-based semantic segmentation approaches on the test set of *DSEC-Semantic* [79]. All approaches adopted the `frame2voxel` representation. Scores reported are IoUs in percentage (%). For each semantic class under each experimental setting, the best score in each column is highlighted in **bold**.

Method	**mIoU**	background	building	fence	person	pole	road	sidewalk	vegetation	car	wall	traffic-sign	**Acc**
Linear Probing													
Random	6.70	7.85	3.37	0.00	0.00	0.00	38.60	0.00	23.83	0.01	0.00	0.00	37.94
MaskCLIP [100]	33.08	75.04	65.06	4.63	0.00	**6.47**	77.06	17.07	55.89	52.17	0.69	**9.78**	76.39
FC-CLIP [97]	43.00	92.53	72.59	**12.43**	0.02	0.00	88.14	52.84	71.92	64.02	**10.54**	7.95	86.00
OpenESS (Ours)	**44.26**	**93.64**	**75.40**	11.82	**1.16**	0.75	**90.29**	**57.96**	**73.15**	**65.36**	9.69	7.67	**87.55**
Fine-Tuning (1%)													
Random	26.62	81.63	33.13	1.77	0.97	7.58	76.81	17.45	51.05	18.64	0.37	3.40	70.04
MaskCLIP [100]	33.89	87.56	53.24	2.34	0.60	8.92	81.71	25.76	59.37	42.56	2.52	8.24	77.79
FC-CLIP [97]	39.12	91.64	59.78	**8.93**	0.00	7.84	**87.58**	**46.58**	66.87	51.30	**4.74**	5.10	82.12
OpenESS (Ours)	**41.41**	**93.01**	**74.01**	3.21	**10.78**	**14.58**	84.50	34.78	**69.82**	**55.12**	4.47	**11.21**	**84.41**
Fine-Tuning (5%)													
Random	31.22	77.13	50.32	**12.36**	1.26	0.00	86.03	41.22	21.48	50.67	2.96	0.04	71.38
MaskCLIP [100]	37.03	91.09	60.52	4.35	11.90	**11.73**	81.24	23.56	61.77	45.93	2.75	12.45	79.58
FC-CLIP [97]	43.71	92.91	**71.21**	10.84	0.00	5.60	**90.11**	57.54	**71.30**	**61.04**	**11.41**	8.81	**86.38**
OpenESS (Ours)	**44.97**	**93.58**	70.18	8.44	**18.22**	11.01	89.72	**57.76**	67.44	56.06	9.59	**12.70**	85.46
Fine-Tuning (10%)													
Random	33.67	85.79	49.85	6.78	8.00	**15.51**	80.78	25.72	58.18	29.97	0.82	8.93	76.69
MaskCLIP [100]	38.83	92.34	69.96	3.64	5.85	12.98	82.23	23.61	66.39	53.23	3.47	13.46	82.36
FC-CLIP [97]	44.09	93.62	72.86	**10.88**	0.00	8.23	**89.81**	**57.05**	**71.95**	60.64	9.58	10.42	86.66
OpenESS (Ours)	**46.25**	**93.92**	**73.34**	8.13	**18.61**	15.41	89.03	52.56	71.76	**61.71**	**9.99**	**14.26**	**86.72**
Fine-Tuning (20%)													
Random	41.31	91.08	67.90	4.68	17.90	**17.41**	85.11	43.24	66.62	43.95	5.03	11.55	82.99
MaskCLIP [100]	42.40	93.19	72.49	5.52	18.21	16.17	84.29	35.04	69.44	54.47	2.43	15.15	84.09
FC-CLIP [97]	47.77	91.05	70.90	7.04	**21.10**	14.84	**91.13**	**64.28**	71.62	61.73	**13.25**	**18.55**	86.95
OpenESS (Ours)	**48.28**	**94.21**	**74.66**	**10.49**	20.46	16.27	90.15	57.66	**73.71**	**63.95**	11.20	18.29	**87.57**

where TP (True Positive) denotes pixels correctly classified as belonging to the class; FP (False Positive) denotes pixels incorrectly classified as belonging to the class; and FN (False Negative) denotes pixels that belong to the class but are incorrectly classified as something else.

The `IoU` metric measures the overlap between the predicted segmentation and the ground truth for a specific class. It returns a value between 0 (no overlap) and 1 (perfect overlap). It is a way to summarize the `mIoU` values for each class into a single metric that captures the overall performance of the model across all classes, *i.e.*, mean IoU (`mIoU`). The `mIoU` of a given prediction is calculated as:

$$\texttt{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\texttt{IoU}_i~, \tag{12}$$

where C is the number of classes and $\texttt{IoU}_i$ denotes the score of class i. mIoU provides a balanced measure since each class contributes equally to the final score, regardless of its size or frequency in the dataset. A higher mIoU indicates better semantic segmentation performance. A score of 1 would indicate perfect segmentation for all classes, while a score of 0 would imply an absence of correct predictions. In this work, all the compared approaches adopt the same mIoU calculation as in the ESS benchmarks [2, 79]. Additionally, we also report the semantic segmentation accuracy (Acc) for the baselines and the proposed framework.

B. Additional Experimental Results

In this section, we provide the class-wise IoU scores for the experiments conducted in the main body of this paper.

B.1. Annotation-Free ESS

The per-class zero-shot event-based semantic segmentation results are shown in Tab. 9. For almost every semantic class, we observe that the proposed OpenESS achieves much higher IoU scores than MaskCLIP [100] and FC-CLIP [97]. This validates the effectiveness of OpenESS for conducting efficient and accurate event-based semantic segmentation without using either the event or frame labels.

B.2. Annotation-Efficient ESS

The per-class linear probing event-based semantic segmentation results are shown in the first block of Tab. 10 and Tab. 11. Specifically, compared to the random initialization baseline, a self-supervised pre-trained network always provides better features. The quality of representation learning often determines the linear probing performance. The network pre-trained using our frame-to-event contrastive distillation and text-to-event consistency regularization tends to achieve higher event-based semantic segmentation results than MaskCLIP [100] and FC-CLIP [97]. Notably, such improvements are holistic across almost all eleven semantic classes in the dataset. These results validate the effectiveness of the proposed OpenESS framework in tackling the challenging event-based semantic segmentation task.

The per-class annotation-efficient event-based semantic segmentation results of the frame2vodel and frame2recon settings under 1%, 5%, 10%, and 20% annotation budgets are shown in Tab. 10 and Tab. 11, respectively. Similar to the findings and conclusions drawn above, we observe clear superiority of the proposed OpenESS framework over the random initialization, MaskCLIP [100], and FC-CLIP [97] approaches. Such consistent performance improvements validate again the effectiveness and superiority of the proposed frame-to-event contrastive distillation and text-to-event consistency regularization. We hope our framework can lay a solid foundation for future works in the established annotation-efficient event-based semantic segmentation.

C. Qualitative Assessment

In this section, we provide sufficient qualitative examples to further attest to the effectiveness and superiority of the proposed framework.

C.1. Open-Vocabulary Examples

The key advantage of our proposed OpenESS framework is its capability to leverage open-world vocabularies from the CLIP text embedding space. Unlike prior event-based semantic segmentation, which relies on pre-defined and fixed categories, our open-vocabulary segmentation aims to understand and categorize image regions into a broader, potentially unlimited range of categories. We provide more open-vocabulary examples in Fig. 10. As can be observed, given proper text prompts like *"road"*, *"sidewalk"*, and *"building"*, our proposed OpenESS framework is capable of generating semantically meaningful attention maps for depicting the corresponding regions. Such a flexible framework can be further adapted to new or unseen categories without the need for extensive retraining, which is particularly beneficial in dynamic environments where new objects or classes might frequently appear. Additionally, the open-vocabulary segmentation pipeline allows users to work with a more extensive range of objects and concepts, enhancing the user experience and interaction capabilities.

C.2. Visual Comparisons

In this section, we provide more qualitative comparisons of our proposed OpenESS framework over prior works [79, 100] on the DSEC-Semantic dataset. Specifically, the visual comparisons are shown in Fig. 11 and Fig. 12. As can be observed, OpenESS shows superior event-based semantic segmentation performance over prior works across a wide range of event scenes under different lighting and weather conditions. Such consistent segmentation performance improvements provide a solid foundation to validate the effectiveness and superiority of the proposed frame-to-event contrastive distillation and text-to-event consistency regularization. For additional qualitative comparisons, kindly refer to Appendix C.4.

C.3. Failure Cases

As can be observed from Fig. 10, Fig. 11, and Fig. 12, the existing event-based semantic segmentation approaches still have room for further improvements. Similar to the conventional semantic segmentation task, it is often hard to accurately segment the boundaries between the semantic objects and backgrounds. In the context of event-based semantic segmentation, such a problem tends to be particularly overt. Unlike traditional cameras that capture dense,

Table 11. **The per-class segmentation results** of annotation-efficient event-based semantic segmentation approaches on the test set of *DSEC-Semantic* [79]. All approaches adopted the `frame2recon` representation. Scores reported are IoUs in percentage (%). For each semantic class under each experimental setting, the best score in each column is highlighted in **bold**.

Method	**mIoU**	background	building	fence	person	pole	road	sidewalk	vegetation	car	wall	traffic-sign	**Acc**
Linear Probing													
Random	6.22	7.55	5.48	0.00	0.00	0.00	39.79	0.00	15.64	0.01	0.00	0.00	36.60
MaskCLIP [100]	27.09	59.82	62.14	1.60	0.00	4.54	69.71	5.34	47.85	38.51	0.40	8.12	70.59
FC-CLIP [97]	40.08	**89.22**	**69.08**	**14.62**	**26.90**	0.00	83.14	21.79	**69.56**	57.78	**7.86**	0.92	82.70
OpenESS (Ours)	**44.08**	88.56	61.43	6.05	21.54	**12.36**	**91.43**	**63.04**	64.01	**60.52**	6.18	**9.76**	**84.48**
Fine-Tuning (1%)													
Random	23.95	76.37	29.59	1.73	0.00	5.75	78.12	9.73	48.96	11.56	0.28	1.38	69.20
MaskCLIP [100]	30.73	79.25	47.26	0.13	1.17	5.04	78.78	19.72	56.13	43.74	1.13	5.70	74.25
FC-CLIP [97]	38.99	87.75	61.48	3.47	4.60	8.06	88.96	55.12	64.41	47.16	**3.61**	4.23	82.90
OpenESS (Ours)	**43.17**	**87.85**	**66.15**	**8.82**	**21.52**	**12.41**	**89.36**	**55.35**	**72.45**	**48.76**	3.40	**8.81**	**84.56**
Fine-Tuning (5%)													
Random	30.42	80.25	38.43	5.50	13.45	9.08	83.45	30.88	51.75	19.53	0.16	2.19	73.65
MaskCLIP [100]	36.33	85.80	60.43	2.60	8.70	7.47	83.10	34.04	64.80	39.60	3.07	**10.00**	80.37
FC-CLIP [97]	43.34	88.28	64.90	6.94	**20.96**	**9.58**	**91.18**	**62.35**	68.09	52.39	4.93	7.16	84.93
OpenESS (Ours)	**45.58**	**89.11**	**70.83**	**10.92**	20.21	1.99	91.04	60.76	**72.07**	**67.91**	**12.90**	3.69	**86.93**
Fine-Tuning (10%)													
Random	34.11	81.85	46.28	4.87	11.30	10.20	85.32	43.16	55.34	32.72	1.28	2.90	77.48
MaskCLIP [100]	40.13	87.31	62.54	4.93	5.09	**12.86**	88.30	50.60	64.74	55.21	0.32	9.51	83.52
FC-CLIP [97]	45.35	89.71	69.00	6.64	22.37	8.33	91.20	64.09	69.34	61.73	7.23	9.19	86.29
OpenESS (Ours)	**48.94**	**90.63**	**71.68**	**12.41**	**29.32**	9.42	**92.53**	**66.19**	**73.76**	**69.03**	**10.71**	**12.71**	**87.84**
Fine-Tuning (20%)													
Random	39.25	87.14	61.80	6.77	3.51	13.19	88.53	56.12	61.95	44.65	1.29	6.84	82.51
MaskCLIP [100]	43.37	89.83	69.80	7.07	8.93	10.67	88.88	52.65	70.71	60.03	3.10	15.39	85.69
FC-CLIP [97]	47.18	91.20	71.39	**11.53**	24.92	9.60	91.58	63.88	71.52	63.44	7.55	12.36	87.07
OpenESS (Ours)	**49.74**	**91.28**	**73.43**	10.69	**27.18**	**13.85**	**92.84**	**67.59**	**74.20**	**69.22**	**10.62**	**16.21**	**88.26**

synchronous frames, event cameras generate sparse, asynchronous events, which brings extra difficulties for accurate boundary segmentation. Meanwhile, the current framework finds it hard to accurately predict the minor classes, such as *fence*, *pole*, *wall*, and *traffic-sign*. We believe these are potential directions that future works can explore to further improve the event-based semantic segmentation performance on top of existing frameworks.

C.4. Video Demos

In addition to the qualitative examples shown in the main body and this supplementary file, we also provide several video clips to further validate the effectiveness and superiority of the proposed approach. Specifically, we provide three video demos in the attachment, named `demo1.mp4`, `demo2.mp4`, and `demo3.mp4`. The first two video demos show open-vocabulary event-based semantic segmentation examples using the class names and open-world vocabularies as the input text prompts, respectively. The third video demo contains qualitative comparisons of the semantic segmentation predictions among our proposed OpenESS and prior works. All the provided video sequences validate again the unique advantage of the proposed open-vocabulary event-based semantic segmentation framework. Kindly refer to our GitHub repository[1] for additional details on accessing these video demos.

D. Broader Impact

In this section, we elaborate on the positive societal influence and potential limitations of the proposed open-vocabulary event-based semantic segmentation framework.

D.1. Positive Societal influence

Event-based cameras can capture extremely fast motions that traditional cameras might miss, making them ideal for dynamic environments. In robotics, this leads to better object detection and scene understanding, enhancing the capabilities of robots in the manufacturing, healthcare, and

[1] https://github.com/ldkong1205/OpenESS

service industries. In autonomous driving, event-based semantic segmentation provides high temporal resolution and low latency, which is crucial for detecting sudden changes in the environment. This can lead to faster and more accurate responses, potentially reducing accidents and enhancing road safety. Our proposed OpenESS is designed to reduce the annotation budget and training burden of existing event-based semantic segmentation approaches. We believe such an efficient way of learning helps increase the scalability of event-based semantic segmentation systems and in turn contributes positively to impact society by enhancing safety, efficiency, and performance in various aspects.

D.2. Potential Limitation

Although our proposed framework is capable of conducting annotation-free and open-vocabulary event-based semantic segmentation and achieves promising performance, there tend to exist several potential limitations. Firstly, our current framework requires the existence of synchronized event and RGB cameras, which might not be maintained by some older event camera systems. Secondly, we directly adopt the standard text prompt templates to generate the text embedding, where a more sophisticated design could further improve the open-vocabulary learning ability of the existing framework. Thirdly, there might still be some self-conflict problems in our frame-to-event contrastive distillation and text-to-event consistency regularization. The design of a better representation learning paradigm on the event-based data could further resolve these issues. We believe these are promising directions that future works can explore to further improve the current framework.

E. Public Resources Used

In this section, we acknowledge the use of public resources, during the course of this work.

E.1. Public Datasets Used

We acknowledge the use of the following public datasets, during the course of this work:

• DSEC[2]	CC BY-SA 4.0
• DSEC-Semantic[3]	CC BY-SA 4.0
• DDD17[4]	CC BY-SA 4.0
• DDD17-Seg[5]	Unknown
• E2VID-Driving[6]	GNU General Public License v3.0

E.2. Public Implementations Used

We acknowledge the use of the following public implementations, during the course of this work:

• ESS[7]	GNU General Public License v3.0
• E2VID[8]	GNU General Public License v3.0
• HMNet[9]	BSD 3-Clause License
• EV-SegNet[10]	Unknown
• SNN-Segmentation[11]	Unknown
• CLIP[12]	MIT License
• MaskCLIP[13]	Apache License 2.0
• FC-CLIP[14]	Apache License 2.0
• SLIC-Superpixels[15]	Unknown
• Segment-Anything[16]	Apache License 2.0

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

[3] https://dsec.ifi.uzh.ch/dsec-semantic

[4] http://sensors.ini.uzh.ch/news_page/DDD17.html

[5] https://github.com/Shathe/Ev-SegNet

[6] https://rpg.ifi.uzh.ch/E2VID.html

[7] https://github.com/uzh-rpg/ess

[8] https://github.com/uzh-rpg/rpg_e2vid

[9] https://github.com/hamarh/HMNet_pth

[10] https://github.com/Shathe/Ev-SegNet

[11] https://github.com/Intelligent-Computing-Lab-Yale/SNN-Segmentation

[12] https://github.com/openai/CLIP

[13] https://github.com/chongzhou96/MaskCLIP

[14] https://github.com/bytedance/fc-clip

[15] https://github.com/PSMM/SLIC-Superpixels

[16] https://github.com/facebookresearch/segment-anything

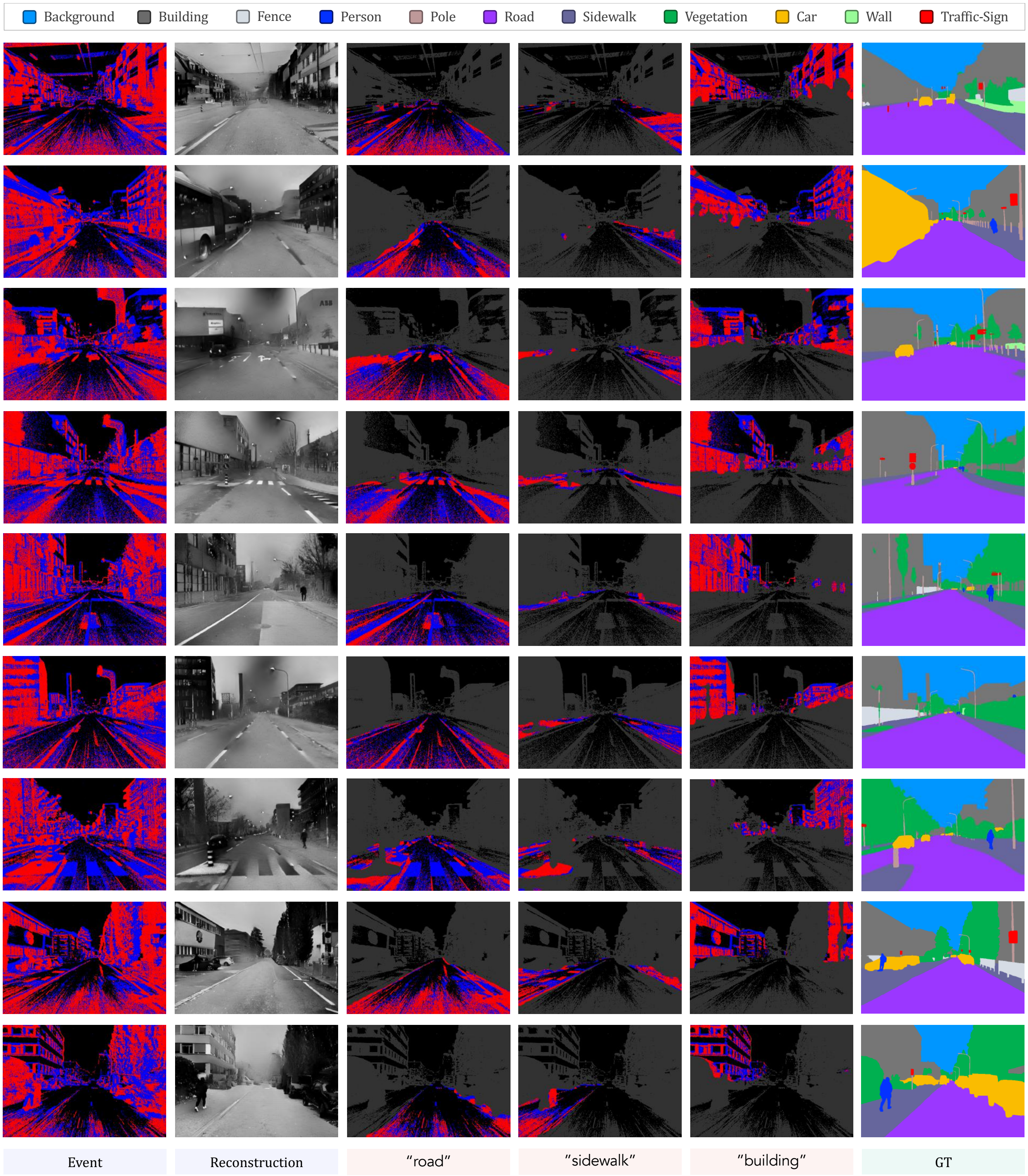


Figure 10. **Qualitative examples** of the language-guided attention maps generated by the proposed OpenESS framework. For each sample, the regions with a high similarity score to the text prompts are highlighted. Best viewed in colors and zoomed-in for additional details.

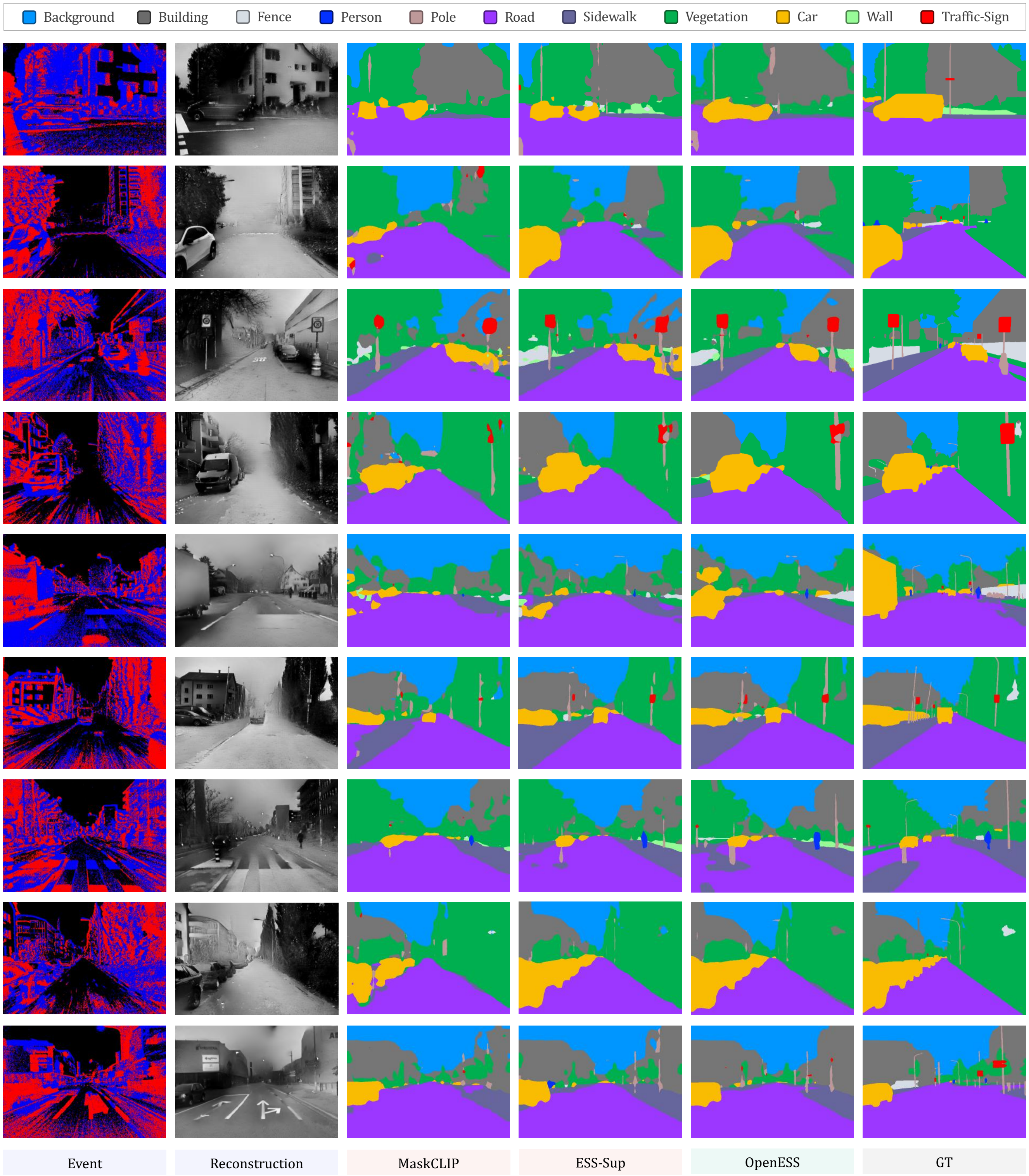


Figure 11. **Qualitative comparisons** (1/2) among different ESS approaches on the *test* set of *DSEC-Semantic* [79]. Best viewed in colors.

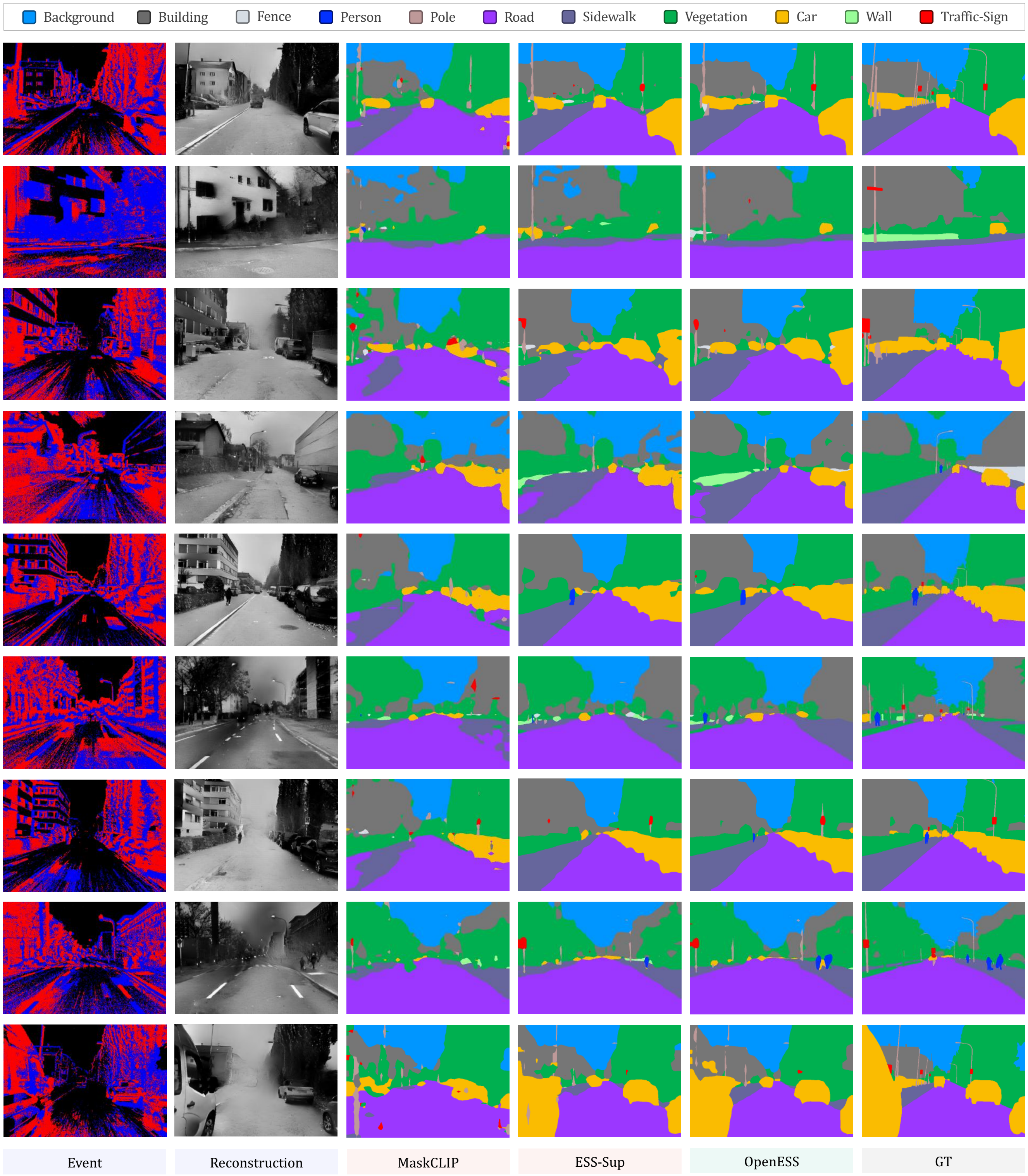


Figure 12. **Qualitative comparisons** (2/2) among different ESS approaches on the *test* set of *DSEC-Semantic* [79]. Best viewed in colors.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. 4, 6, 7, 10, 12, 13

[2] Inigo Alonso and Ana C. Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2019. 1, 2, 5, 6, 8, 9, 10, 11, 13, 16

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 6, 7

[4] Ahmed Nabil Belbachir, Stephan Schraml, Manfred Mayerhofer, and Michael Hofstätter. A novel hdr depth camera for real-time 3d 360 panoramic vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 425–432, 2014. 2

[5] Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. In *International Conference on Machine Learning Workshops*, pages 1–9, 2017. 2, 6, 7, 8, 9

[6] Shristi Das Biswas, Adarsh Kosta, Chamika Liyanagedera, Marco Apolinario, and Kaushik Roy. Halsie: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1, 2, 6

[7] Vincent Brebion, Julien Moreau, and Franck Davoine. Real-time optical flow for vehicular perception with low- and high-resolution event cameras. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15066–15078, 2021. 2

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924, 2020. 4, 5, 8

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 5, 8

[10] Kaiwei Che, Luziwei Leng, Kaixuan Zhang, Jianguo Zhang, Qinghu Meng, Jie Cheng, Qinghai Guo, and Jianxing Liao. Differentiable hierarchical and surrogate gradient search for spiking neural networks. In *Advances in Neural Information Processing Systems*, pages 24975–24990, 2022. 2

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 1

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 801–818, 2018. 1

[13] Runnan Chen, Youquan Liu, Lingdong Kong, Nenglun Chen, Xinge Zhu, Yuexin Ma, Tongliang Liu, and Wenping Wang. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, 2023. 2, 10

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 7

[15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4, 5, 8

[16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9620–9629, 2021. 7

[17] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 2

[18] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *IEEE/CVF International Conference on Computer Vision*, pages 19866–19877, 2023. 2

[19] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 9

[20] Javier Cuadrado, Ulysse Rançon, Benoit R. Cottereau, Francisco Barranco, and Timothée Masquelier. Optical flow estimation from event-based cameras and spiking neural networks. *Frontiers in Neuroscience*, 17:1160034, 2023. 2

[21] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 2

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[23] Burak Ercan, Onur Eker, Aykut Erdem, and Erkut Erdem. Evreal: Towards a comprehensive benchmark and analysis suite for event-based video reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3942–3951, 2023. 2

[24] Burak Ercan, Onur Eker, Canberk Saglam, Aykut Erdem, and Erkut Erdem. Hypere2vid: Improving event-based video reconstruction via hypernetworks. *arXiv preprint arXiv:2305.06382*, 2023. 2

[25] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. 1, 2, 3

[26] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *European Conference on Computer Vision Workshops*, pages 266–282, 2022. 2

[27] Daniel Gehrig and Davide Scaramuzza. Pushing the limits of asynchronous graph-based object detection with event cameras. *arXiv preprint arXiv:2211.12324*, 2022. 2

[28] Daniel Gehrig and Davide Scaramuzza. Are high-resolution event cameras really needed? *arXiv preprint arXiv:2203.14672*, 2022. 1

[29] Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 2, 4

[30] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 2, 6

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

[32] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 5, 9

[33] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *IEEE International Conference on 3D Vision*, pages 197–206, 2021. 2

[34] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from events and frames. *arXiv preprint arXiv:2203.13674*, 2022. 2

[35] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision Workshops*, pages 540–557, 2022. 2

[36] Suman Ghosh and Guillermo Gallego. Multi-event-camera depth estimation and outlier rejection by refocused events fusion. *Advanced Intelligent Systems*, 4(12):2200221, 2020. 2

[37] Renxiang Guan, Zihao Li, Xianju Li, and Chang Tang. Pixel-superpixel contrastive learning and pseudo-label correction for hyperspectral image clustering. *arXiv preprint arXiv:2312.09630*, 2023. 3

[38] Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22867–22876, 2023. 1, 2, 6, 9

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 3, 5

[40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 7

[41] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247, 2023. 2

[42] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *IEEE International Conference on 3D Vision*, pages 534–542, 2020. 2

[43] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2022. 2

[44] Kunping Huang, Sen Zhang, Jing Zhang, and Dacheng Tao. Event-based simultaneous localization and mapping: A comprehensive survey. *arXiv preprint arXiv:2304.09793*, 2023. 1

[45] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 2

[46] Olivier J. Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron Van den Oord, Oriol Vinyals, and Joao Carreira. Efficient visual pretraining with contrastive detection. In *IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 2

[47] Zexi Jia, Kaichao You, Weihua He, Yang Tian, Yongxiang Feng, Yaoyuan Wang, Xu Jia, Yihang Lou, Jingyi Zhang, Guoqi Li, and Ziyang Zhang. Event-based semantic segmentation with posterior attentio. *IEEE Transactions on Image Processing*, 32:1829–1842, 2023. 2, 6, 9

[48] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021. 2

[49] Youngeun Kim, Joshua Chough, and Priyadarshini Panda. Beyond classification: Directly training spiking neural networks for semantic segmentation. *Neuromorphic Computing and Engineering*, 2(4):044015, 2022. 2, 4, 5, 6, 7, 14

[50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 4, 6, 7, 10, 11, 12

[51] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2, 4

[52] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[53] Yijin Li, Zhaoyang Huang, Shuo Chen, Xiaoyu Shi, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Blinkflow: A dataset to push the limits of event-based optical flow estimation. *arXiv preprint arXiv:2303.07716*, 2023. 2

[54] Zhengqin Li and Jiansheng Chen. Superpixel segmentation using linear spectral clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1356–1363, 2015. 4

[55] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2

[56] Daqi Liu, Alvaro Parra, and Tat-Jun Chin. Spatiotemporal registration for event-based visual odometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2021. 2

[57] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, 2023. 3, 10

[58] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[60] Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. 2

[61] Nico Messikommer, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 2, 6

[62] Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Event-intensity stereo: Estimating depth by the best of both worlds. In *IEEE/CVF International Conference on Computer Vision*, pages 4258–4267, 2021. 2

[63] Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36 (6):51–63, 2019. 2

[64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[65] Tianbo Pan, Zidong Cao, and Lin Wang. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. *arXiv preprint arXiv:2309.12842*, 2023. 2

[66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019. 5

[67] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Tai Wang, Xinge Zhu, and Yuexin Ma. Learning to adapt sam for segmenting cross-domain point clouds. *arXiv preprint arXiv:2310.08820*, 2023. 2, 10

[68] Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-based vision. In *IEEE/CVF International Conference on Computer Vision*, pages 6038–6048, 2023. 2

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3, 4, 9

[70] Ulysse Rançon, Javier Cuadrado-Anibarro, Benoit R. Cottereau, and Timothée Masquelier. Stereospike: Depth learning with a spiking neural network. *IEEE Access*, 10: 127428–127439, 2022. 2

[71] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 4

[72] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters*, 2(2):593–600, 2016. 2

[73] 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. 2, 4, 5, 6, 8, 14

[74] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 2

[75] Stephan Schraml, Ahmed Nabil Belbachir, and Horst Bischof. Event-driven stereo matching for real-time 3d panoramic vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 466–474, 2015. 2

[76] Bongki Son, Yunjae Suh, Sungho Kim, Heejae Jung, Jun-Seok Kim, Changwoo Shin, Keunju Park, Kyoobin Lee, Jinman Park, Jooyeon Woo, Yohan Roh, Hyunku Lee, Yibing Wang, Ilia Ovsiannikov, and Hyunsurk Ryu. A 640× 480 dynamic vision sensor with a 9μm pixel and 300meps address-event representation. In *IEEE International Solid-State Circuits Conference*, 2017. 1

[77] Lea Steffen, Daniel Reichard, Jakob Weinland, Jacques Kaiser, Arne Roennau, and Rüdiger Dillmann. Neuromorphic stereo vision: A survey of bio-inspired sensors and algorithms. *Frontiers in Neuroscience*, 13:28, 2019. 2

[78] David Stutz, Alexander Hermans, and Bastian Leibe. Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166:1–27, 2018. 10

[79] 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, 2022. 1, 2, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 20, 21

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 3

[81] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing*, 31:7237–7251, 2022. 2

[82] Jiacheng Wang, Xiaomeng Li, Yiming Han, Jing Qin, Liansheng Wang, and Zhou Qichao. Separated contrastive learning for organ-at-risk and gross-tumor-volume segmentation with limited annotation. In *AAAI Conference on Artificial Intelligence*, pages 2459–2467, 2022. 3

[83] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *IEEE/CVF International Conference on Computer Vision*, pages 2135–2145, 2021. 2, 6

[84] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2, 6

[85] Shu Wang, Huchuan Lu, Fan Yang, and Ming-Hsuan Yang. Superpixel tracking. In *IEEE/CVF International Conference on Computer Vision*, pages 1323–1330, 2011. 4

[86] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 6

[87] Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. *arXiv preprint arXiv:2301.00805*, 2023. 2

[88] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *arXiv preprint arXiv:2306.15880*, 2023. 2

[89] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15254–15264, 2023. 2

[90] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, 12:331, 2018. 2

[91] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12760–12771, 2023. 6

[92] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 2

[93] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[94] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Rui Zhang, Qingyuan Zhou, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *arXiv preprint arXiv:2403.10001*, 2024. 2, 10

[95] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pretraining. In *IEEE/CVF International Conference on Computer Vision*, pages 10699–10709, 2023. 7

[96] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 2

[97] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Advances in Neural Information Processing Systems*, 2023. 2, 4, 6, 8, 15, 16, 17

[98] Zelin Zhang, Anthony J. Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear

inverse problem with deep regularization using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8372–8389, 2023. 2

[99] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023. 1

[100] Chong Zhou, Chen Change Loy, and Bo Da. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712, 2022. 2, 4, 6, 8, 15, 16, 17

[101] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pretraining with online tokenizer. In *International Conference on Learning Representations*, 2021. 7

[102] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. E-clip: Towards label-efficient event-based open-world understanding by clip. *arXiv preprint arXiv:2308.03135*, 2023. 2

[103] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. Rgb-event fusion for moving object detection in autonomous driving. In *IEEE International Conference on Robotics and Automation*, pages 7808–7815, 2023. 2

[104] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *European Conference on Computer Vision Workshops*, 2018. 4

[105] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 2, 4

[106] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *arXiv preprint arXiv:2307.09220*, 2023. 2

[107] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3594–3604, 2022. 2

[108] Xiao-Long Zou, Tie-Jun Huang, and Si Wu. Towards a new paradigm for brain-inspired computer vision. *Machine Intelligence Research*, 19(5):412–424, 2022. 2

[109] Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *IEEE/CVF International Conference on Computer Vision*, pages 12846–128567, 2023. 2