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

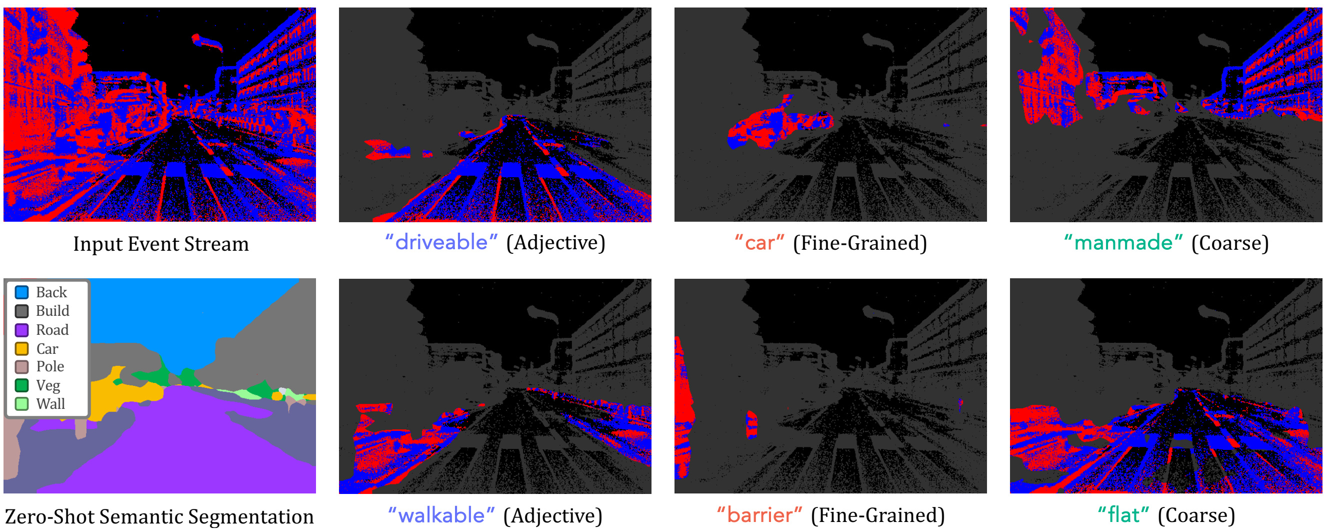

Figure 1. **Open-vocabulary event-based semantic segmentation (OpenESS)**. Our framework is capable of performing zero-shot semantic segmentation of event data streams with open vocabularies. Given raw events and text prompts as inputs, OpenESS outputs semantically coherent open-world predictions across 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 [27, 43, 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 [24].

Event-based semantic segmentation (ESS) emerges as one of the core event perception tasks and has gained increasing attention [2, 6, 37, 78]. ESS inherits the challenges of traditional image segmentation [10, 11, 18, 38, 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 [46, 48, 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, 78]. 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 [29, 61, 78] 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 [12, 49, 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 [24, 60, 77, 108]. A large variety of event-based recognition, perception, localization, and reconstruction tasks have been established, encompassing object recognition [17, 28, 47, 68], object detection [26, 30, 103, 109], depth estimation [16, 35, 41, 62, 65, 70], optical flow [19, 32, 33, 53, 80, 81, 105], intensity-image reconstruction [22, 23, 73, 98, 107], visual odometry and SLAM [42, 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, 78]. 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 [29]. EvDistill [84] and DTL [83] utilized aligned frames to enhance event-based learning. EV-Transfer [61] and ESS [78] leveraged domain adaptation to transfer knowledge from existing image datasets to events. Recently, HALSIE [6] and HMNet [37] 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 [9, 48, 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 [12, 88, 106]. Such trends encompass image-based zero-shot and open-vocabulary detection [25, 52, 89, 96], as well as semantic [34, 50, 55, 97, 100], instance [44, 87], and panoptic [20, 40, 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 [17] 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 [45, 74], where pre-processed superpixels are used to

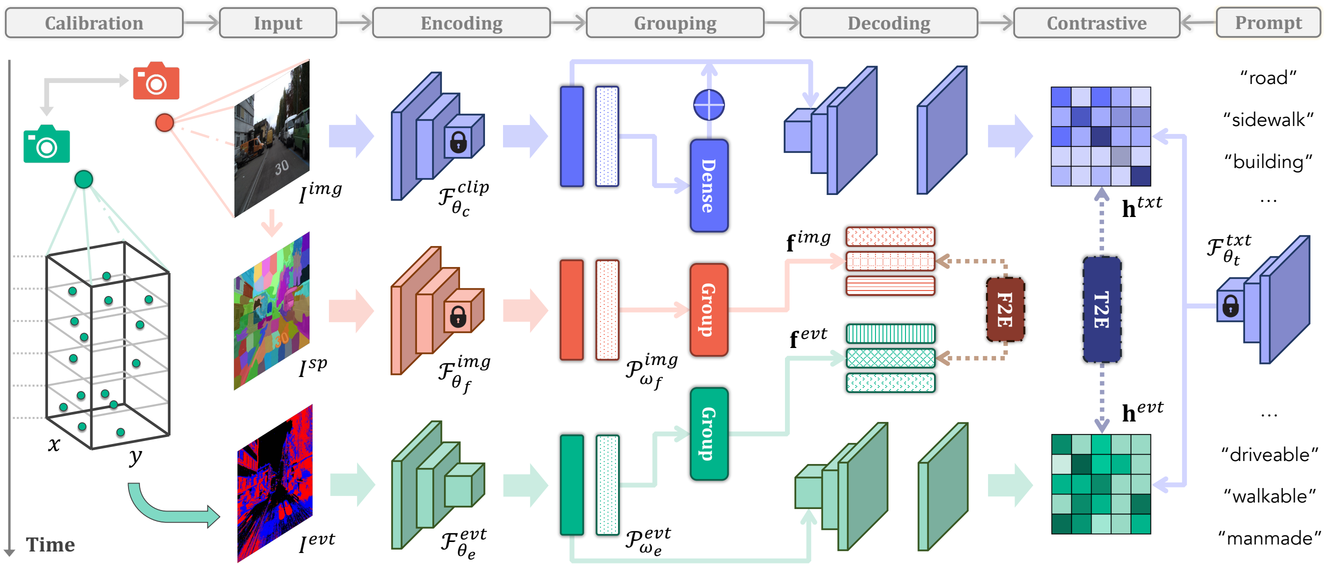

Figure 2. **Architecture overview of the OpenESS framework**. We distill off-the-shelf knowledge from vision-languages models to event representations (*cf*. Sec. 3.1). Given a calibrated event I^{evt} and a frame I^{img}, we extract their features from the event network $\mathcal{F}^{evt}_{\theta_e}$ 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 [36], 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 [38] or Vision Transformer [21]) and a text encoder (using a Transformer architecture [79]) 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 [24], 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 [28, 104, 105], frame-like reconstructions [73], and bio-inspired spikes [48]. 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) [49] 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 [49] 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 [7, 14, 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:

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

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*** [78] provides semantic labels for 11 sequences in the DSEC [31] 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 [38], and SpikingFCN [48], respectively, with an AdamW [59] optimizer with cosine learning rate scheduler. The frame branch uses a pre-trained ResNet-50 [7, 8, 14] 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* [78] datasets. All scores are in percentage (%). The **best** score from each learning setting is highlighted in **bold**.

Method	Venue	DDD17 Acc	DDD17 mIoU	DSEC Acc	DSEC 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 [29]	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 [48]	NCE'22	-	34.20	-	-
EV-Transfer [61]	RA-L'22	51.90	15.52	63.00	24.37
ESS [78]	ECCV'22	88.43	53.09	84.17	45.38
ESS-Sup [78]	ECCV'22	91.08	**61.37**	89.37	53.29
P2T-FPN [91]	TPAMI'23	94.57	54.64	-	-
EvSegformer [46]	TIP'23	**94.72**	54.41	-	-
HMNet-B [37]	CVPR'23	-	-	88.70	51.20
HMNet-L [37]	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* [78] 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* [78], 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

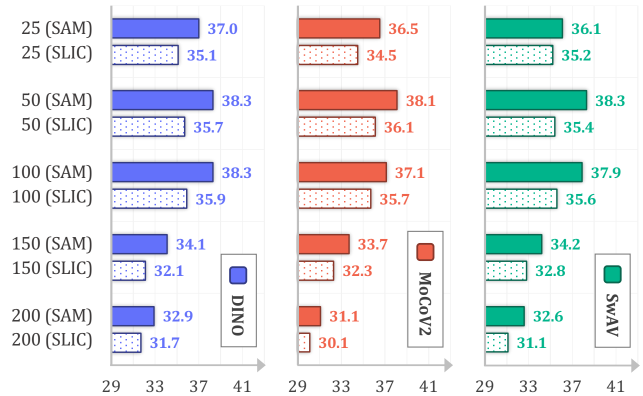

Figure 3. **Ablation study** on the number of superpixels (provided by either SAM [49] or SLIC [1]) involved in calculating the frame-to-event contrastive loss. Models after pre-training are fine-tuned with 1% annotations. All mIoU scores are in percentage (%).

previously best-performing methods, OpenESS is 1.63% and 2.21% better in terms of mIoU scores on *DDD17-Seg* [2] and *DSEC-Semantic* [78], 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.

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* [78]. 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,

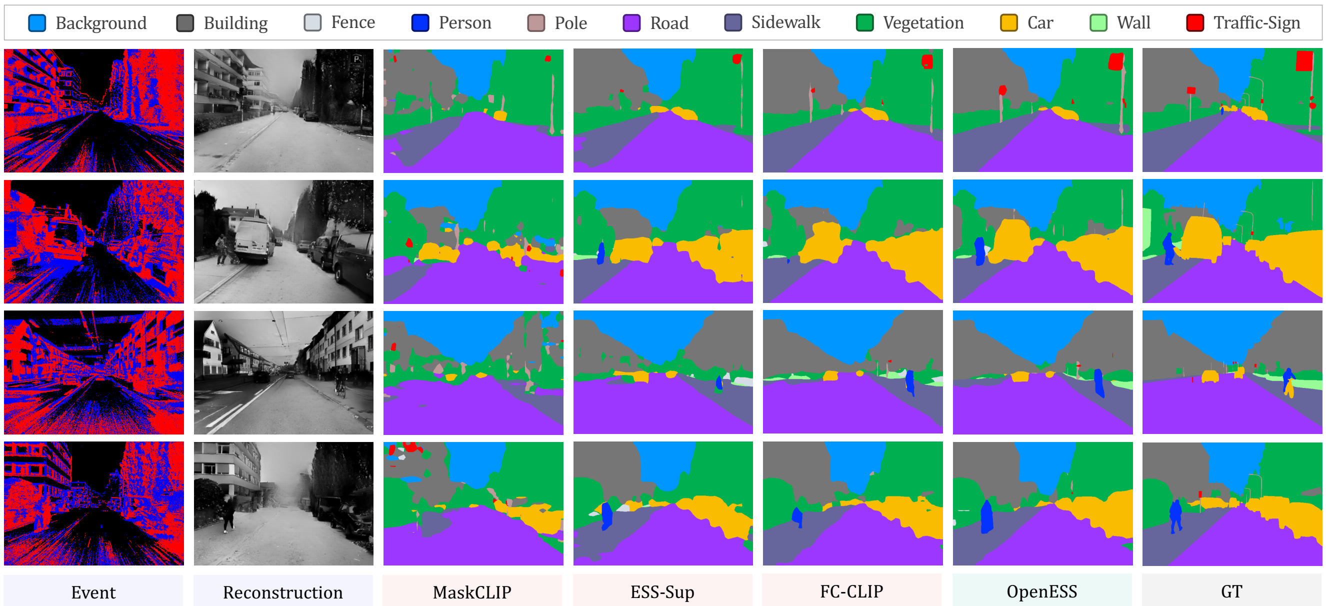

Figure 4. **Qualitative comparisons** of state-of-the-art ESS approaches on the *test* set of *DSEC-Semantic* [78]. Each color corresponds to a distinct semantic category. GT denotes the ground truth semantic maps. Best viewed in colors and zoomed-in for additional details.

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 [15]	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 [39]	CVPR'22	ViT-B/16	✗	52.36	47.56
Random	-	ResNet-50	✗	56.96	57.60
SimCLR [13]	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

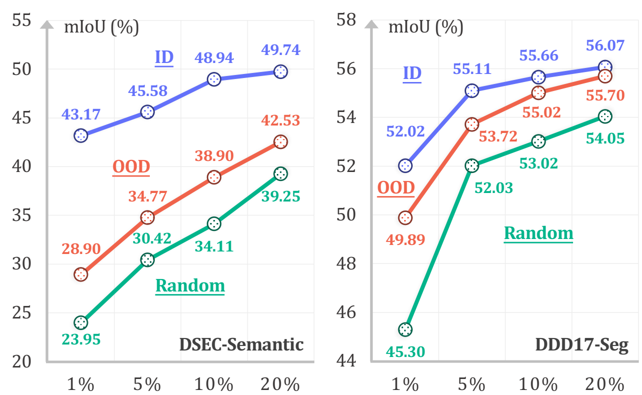

Figure 5. **Cross-dataset representation learning** results of comparing OpenESS pre-training using in-distribution (ID) and out-of-distribution (OOD) data in-between the *DDD17-Seg* [5] and *DSEC-Semantic* [78] datasets. Models after pre-training are fine-tuned with 1%, 5%, 10%, and 20% annotations, respectively.

39, 95, 101] and contrastive learning-based [13, 15] 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 [48], 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 [49] 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* [78] datasets. All mIoU scores are given in percentage (%). The **best** mIoU scores from each learning configuration are highlighted in **bold**.

Method	Configuration	DSEC-Semantic LP	1%	5%	10%	20%	Full	DDD17-Seg 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]	Voxel Grid	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]		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]	Reconstruction	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]		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 [8], MoCoV2 [14], or SwAV [7], 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* [78] 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.

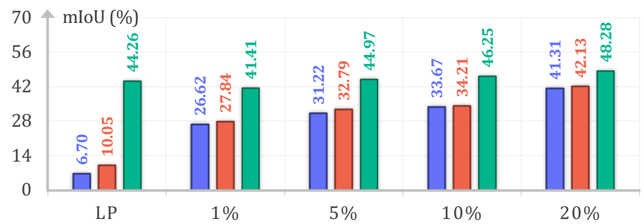

Figure 6. **Single-modality OpenESS representation learning study** on the *DSEC-Semantic* [78] dataset. The results are from models of random initialization (■), `recon2voxel` pre-training (■), and `frame2voxel` pre-training (■), respectively, after linear probing (LP) and annotation-efficient fine-tuning.

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.

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

[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

[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

[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] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

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

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

[28] 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

[29] 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

[30] 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

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

[32] 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

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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

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

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

[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

[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

[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

[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] 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

[79] 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

[80] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. 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

[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

[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

[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

[101] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training 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