



HAL
open science

OVOSE: Open-Vocabulary Semantic Segmentation in Event-Based Cameras

Muhammad Rameez Ur Rahman, Jhony H. Giraldo, Indro Spinelli, Stéphane Lathuilière, Fabio Galasso

► **To cite this version:**

Muhammad Rameez Ur Rahman, Jhony H. Giraldo, Indro Spinelli, Stéphane Lathuilière, Fabio Galasso. OVOSE: Open-Vocabulary Semantic Segmentation in Event-Based Cameras. 27 th International Conference on Pattern Recognition (ICPR), Dec 2024, Kolkata, INDIA, India. pp.18-33, <10.1007/978-3-031-78444-6_2>. <hal-04794160>

HAL Id: hal-04794160

<https://hal.science/hal-04794160v1>

Submitted on 17 Jul 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

OVOSE: Open-Vocabulary Semantic Segmentation in Event-Based Cameras

Muhammad Rameez Ur Rahman¹, Jhony H. Giraldo², Indro Spinelli¹,
Stéphane Lathuilière², and Fabio Galasso¹

¹ Sapienza University of Rome, Italy

{rahman, spinelli, galasso}@di.uniroma1.it

² LTCI, Télécom Paris, Institut Polytechnique de Paris, France

{jhony.giraldo, stephane.lathuiliere}@telecom-paris.fr

Abstract. Event cameras, known for low-latency operation and superior performance in challenging lighting conditions, are suitable for sensitive computer vision tasks such as semantic segmentation in autonomous driving. However, challenges arise due to limited event-based data and the absence of large-scale segmentation benchmarks. Current works are confined to closed-set semantic segmentation, limiting their adaptability to other applications. In this paper, we introduce OVOSE, the first Open-Vocabulary Semantic Segmentation algorithm for Event cameras. OVOSE leverages synthetic event data and knowledge distillation from a pre-trained image-based foundation model to an event-based counterpart, effectively preserving spatial context and transferring open-vocabulary semantic segmentation capabilities. We evaluate the performance of OVOSE on two driving semantic segmentation datasets DDD17, and DSEC-Semantic, comparing it with existing conventional image open-vocabulary models adapted for event-based data. Similarly, we compare OVOSE with state-of-the-art methods designed for closed-set settings in unsupervised domain adaptation for event-based semantic segmentation. OVOSE demonstrates superior performance, showcasing its potential for real-world applications. The code is available at <https://github.com/ram95d/OVOSE>.

Keywords: Open vocabulary segmentation · Low-level vision · Distillation.

1 Introduction

Event cameras, known for their exceptional temporal resolution, low latency, and motion blur resistance, have transformed various deep learning applications [10]. Use cases include autonomous driving [4], object recognition [8], and semantic segmentation [2]. Their outstanding performance in challenging conditions makes event cameras optimal for capturing reliable visual data [19]. Despite their success, integrating event cameras into existing computer vision models is challenging. Their unique data format, featuring asynchronous event streams without

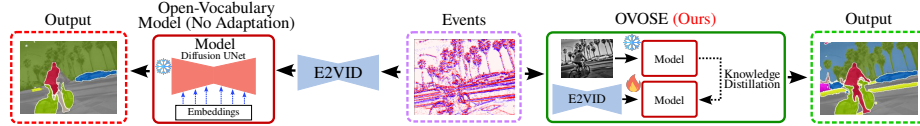


Fig. 1. Output of a regular RGB foundation model for semantic segmentation and OVOSE in event-based data. OVOSE accurately segments person, trees, and (the sky).

traditional image frames [28], necessitates a reassessment of established techniques. While traditional image-based semantic segmentation has made notable progress [17], event cameras, being less prevalent in real-world scenarios, suffer from a scarcity of both raw and labelled data. This fact raises two intertwined issues: the impossibility of collecting internet-scale event datasets and, as a consequence, the difficulty in training data-intensive deep learning techniques. This challenge obstructs effective semantic segmentation model training, especially in scenarios lacking established benchmarks or methodologies.

Recently, powerful foundation models have demonstrated their effectiveness in segmenting conventional images [17]. Some of these models extend the closed-set capabilities of classical semantic segmentation models [5,6], performing well in open-vocabulary settings where the set of classes in training and testing are disjoint. However, applying these powerful open-vocabulary models directly to event-based images is challenging, and retraining them is impractical due to the scarcity of annotated event-based data [15].

The alternative approach of converting events into images using E2VID [25], and pairing it with an open-vocabulary model does not perform satisfactorily as illustrated in Fig. 1. Despite the fact that E2VID is intended to reduce the domain gap between images and events, there still exists a difference between real grayscale and reconstructed-event images. In this paper, we introduce OVOSE, the first Open-Vocabulary Semantic Segmentation algorithm tailored for event-based data. OVOSE operates as a two-branch network, with one branch dedicated to grayscale images and the other to event data. Each branch incorporates a copy of an image foundation model, with OVOSE adapting the event branch for optimal performance in event-based data. Our algorithm integrates text-to-image diffusion [23] and a mask generator. Using a CLIP-style image encoder and MLP, we derive embeddings for conditioning the text-to-image diffusion UNet. We use UNet’s features as input to the mask generator for mask generation. We categorize the mask generator’s outputs using a frozen CLIP-style text encoder for open-vocabulary segmentation. To enhance model performance, we distill knowledge from the image branch to the events branch, using an E2VID model [25] for translating events to reconstructed images. OVOSE, characterized by its simplicity and effectiveness, outperforms existing models in event-based semantic segmentation, including Unsupervised Domain Adaptation (UDA) methods, demonstrating its effectiveness in addressing the open-vocabulary segmentation problem in event-based data.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we present the first open-vocabulary semantic segmentation approach tailored explicitly for event-based data.
- We introduce OVOSE that distills knowledge to transfer semantic insights from a foundation model trained on regular images to enhance open-vocabulary segmentation performance for event-based data.
- To mitigate the effects of sub-optimal reconstructions, we investigate various mask reweighting strategies and introduced a novel dissimilarity network. This network recalibrates the mask loss by leveraging the differences between reconstructed and original images, enabling precise fine-tuning of the segmentation model and thus producing robust and accurate predictions.
- We perform extensive evaluations in open-vocabulary semantic segmentation for three event datasets. OVOSE readily outperforms existing closed-set semantic segmentation methods and straightforward adaptations of open-vocabulary models. A set of ablation studies validates the key components of our algorithm.

2 Related Work

Event Camera Semantic Segmentation. Alonso et al. [1] introduced event camera semantic segmentation an Xception-type network for the DDD17 dataset but suffered from limitations in the quality of generated labels. Gehrig et al. [11] improved performance substantially by utilizing a synthetic event dataset converted from videos. Wang et al. [30] explored knowledge distillation and transfer learning between images and events, though relying on labeled datasets. Mes-sikommer et al. [20] proposed a method aligning image and event embeddings but faced challenges with hallucinations [28]. Sun et al. [28] addressed domain gap reduction between events and images, yet required real event-based data. Yang et al. [35] presented a self-supervised learning framework, but still relied on labeled datasets for fine-tuning. While existing methods are designed for closed-set semantic segmentation with limited known classes, our approach is an open-vocabulary segmentation method tailored for event-based cameras, trained solely on synthetic unlabeled datasets.

Open-Vocabulary Semantic Segmentation. Recent approaches in open-vocabulary semantic segmentation for regular images have centered on embedding spaces linking image pixels to class descriptors [3,32]. Some methods leverage CLIP [22] for text and image embeddings [16], while others combine CLIP with Vision Transformer [9]. OpenSeed is introduced [36] for joint segmentation and detection tasks. ODISE [34] uniquely merges pre-trained text-image diffusion and discriminative models [26], excelling in open-vocabulary panoptic segmentation. However, these methods are designed for image data and don't directly translate to event cameras due to the fundamentally different data representation (continuous event streams vs. static images). This work proposes a novel approach for open-vocabulary semantic segmentation on event cameras. We bridge the gap between image-based methods and event data by transferring

knowledge from a powerful image foundation model to a new model specifically tailored for the event domain.

2.1 Knowledge Distillation

Prior knowledge distillation (KD) [14] methods focus on single modality, *i.e.*, image using logits [33] or features [21] or across different modalities [37] utilizing paired data. KD is also applied to event cameras [31] to distill knowledge from the image-based teacher model to the event-based student model. However, they require labeled image datasets and unlabelled real events and frames for effective transfer in semantic segmentation. Furthermore, [30] employs event-to-image transfer for semantic segmentation, but it gives poor performance [28] when applied for semantic segmentation. Unlike these approaches, we employ a synthetic training dataset and distill knowledge from an image foundation model to a foundation model tailored for open vocabulary semantic segmentation in events. The problem becomes complex as we strive to bridge the gap between events and images, and additionally tackle the synthetic-to-real gap.

3 Open-vocabulary Segmentation in Events

3.1 Preliminaries.

Event Representation. Each pixel in an event camera operates independently and reports brightness changes asynchronously, signaling only when the changes exceed a certain threshold. When a change is detected, an event is generated, capturing the pixel positions (x_i, y_i) , timestamp (t_i) , and polarity (p_i) , indicating whether a change involves an increase or decrease in brightness. Consequently, each event can be represented as $\mathbf{E} = [x_i, y_i, t_i, p_i]$. In this paper, we transform events into grid-like representations [12], such as voxel grid [38,28] to facilitate further processing. We utilize a voxel grid representation of events as input to our model.

Text-to-Image Diffusion UNet. The text-to-image diffusion model [23] generates high-quality images from textual descriptions, utilizing the power of pre-trained encoders like CLIP [22] to encode text into embeddings. The process starts by adding Gaussian noise to images, and then the UNet architecture effectively reverses this noise, guided by cross-attention mechanisms that align the text embeddings with visual features, making the features semantically rich. We incorporated the text-to-image diffusion UNet into our model to extract rich features that are relevant to the text.

Problem statement. Open vocabulary represents a generalization of the zero-shot task in semantic segmentation. In this setting, a model predicts masks for unseen classes \mathcal{C}^{unseen} by learning from labeled data of seen classes \mathcal{C}^{seen} . The sets of seen and unseen classes are separate and do not overlap, *i.e.*, $\mathcal{C}^{unseen} \cap \mathcal{C}^{seen} = \emptyset$. The objective of this work is to train a model F_{θ} with parameters θ to predict the segmentation map of some stream of events. To solve this, we have

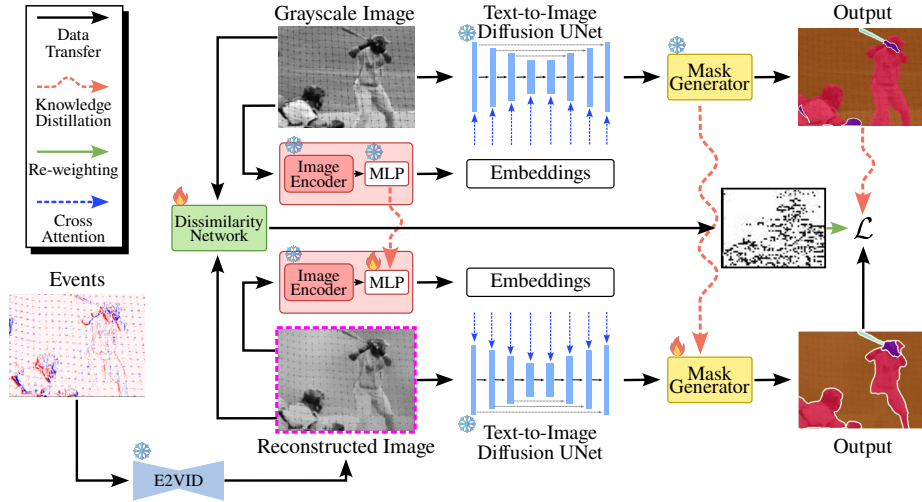


Fig. 2. Overview of OVOSE pipeline. Our algorithm comprises two components: the original grayscale image branch and the event-based branch. Initially, events are transformed into a grayscale image using the E2VID model. Subsequently, both the original and reconstructed grayscale images undergo text embedding through an image encoder and an MLP. The features from a frozen text-to-image diffusion UNet are then extracted for each tuple of image and text embedding. For each branch, a mask generator predicts class-agnostic binary masks and associated mask embedding features. Categorization is achieved through a dot product between mask embedding features and text embeddings. Both branches are initialized with ODISE weights [34], and knowledge distillation occurs from the original image branch to the event-based branch during training. Original and reconstructed images are input into a dissimilarity network to weigh the distillation in the outputs. During the evaluation, only the event-based branch is utilized.

a unlabeled training set $\mathcal{X}_{train} = \{\mathbf{E}_i^{(t)}, \mathbf{X}_i^{(t)}\}_{i=1}^{N_t}$, where $\mathbf{E}_i^{(t)}$ is the i th stream of events, $\mathbf{X}_i^{(t)} \in \mathbb{R}^{H \times W}$ is the i th original grayscale image with H and W the height and width of the image, and N_t is the number of streams in the dataset. For the training set, we additionally need the set of seen classes \mathcal{C}^{seen} . We evaluate F_θ in a testing set $\mathcal{X}_{test} = \{\mathbf{E}_i^{(s)}, \mathbf{Y}_i^{(s)}\}_{i=1}^{N_s}$ where $\mathbf{Y}_i^{(s)} \in \{0, 1\}^{H \times W \times |\mathcal{C}^{unseen}|}$.

3.2 Overview of OVOSE

As shown in Figure 2 OVOSE is divided into two sections: (i) the image branch that takes as input the original grayscale images $\mathbf{X}_i^{(t)} \in \mathcal{X}_{train}$, and (ii) the event branch that takes as input the stream of events $\mathbf{E}_i^{(t)} \in \mathcal{X}_{train}$ or $\mathbf{E}_i^{(s)} \in \mathcal{X}_{test}$. It is worth clarifying that during the evaluation we only use the event branch. During training, the whole image branch is frozen and only used to distill knowledge to the event branch. For the event branch, we use the pre-trained E2VID model with parameters θ_e [25] to transform some stream of events

\mathbf{E} into a reconstructed image $\hat{\mathbf{X}}$, thus $F_{\theta_e}(\mathbf{E}) = \hat{\mathbf{X}}$. E2VID introduces a novel approach by leveraging a convolutional recurrent neural network architecture to process event camera data’s sparse and asynchronous nature, producing high temporal resolution images. Taking as input \mathbf{X} or $\hat{\mathbf{X}}$, the forward pass of each branch is identical, so we explain only one of these in the following.

We first employ a frozen image encoder of type CLIP $\mathcal{V}(\cdot)$ [22] to encode \mathbf{X} or $\hat{\mathbf{X}}$ into an embedding space. Subsequently, a learnable MLP is used to project the image embedding into implicit text embeddings. We use \mathbf{X} or $\hat{\mathbf{X}}$ along with the implicit text embeddings as input to a text-to-image diffusion model [17] for feature extraction. More precisely, we employ a UNet architecture to do the denoising process. Formally for the image branch, we have:

$$\mathbf{f} = F_{\theta_U}(\mathbf{X}, \text{MLP}(\mathcal{V}(\mathbf{X}))), \quad (1)$$

where \mathbf{f} is the feature vector from the diffusion network in the image branch, and θ_U is the parameters of the UNet model. We feed this feature vector \mathbf{f} as input to a Mask2Former model [7] to produce class-agnostic binary masks with their corresponding mask embeddings. For categorization, we use a text encoder of type CLIP $\mathcal{T}(\cdot)$ to embed the categories in \mathcal{C}^{seen} . We thus perform a dot product between text and mask embeddings to categorize each mask.

We use the estimated segmentation map of the image branch as the ground truth of the event branch by computing a loss function between both outputs. However, as we kept E2VID frozen, under poor reconstructions, we weighed this loss function using the output of a dissimilarity network to give more emphasis to the regions where E2VID reconstructs well. We further perform knowledge distillation from the Mask2Former in the image to the Mask2Former in the event branch, and similarly for the MLP networks. The parameters of the two branches are initialized with the weights of the ODISE model [34].

3.3 Distilling Image Embeddings

To address the difference in image embeddings between the output of the MLP for the grayscale image and the corresponding output for the synthetic (reconstructed) image, we implement knowledge distillation. This involves transferring knowledge from the image embeddings of the original image to those of the reconstructed image in the event branch. To do so, we introduce the minimization of the Frobenius norm ($\|\cdot\|_F$) of the matrix of differences between real and reconstructed images encoded by the trainable MLP:

$$\mathcal{L}_t = \|\text{MLP}_{\mathbf{X}}(\mathcal{V}(\mathbf{X})) - \text{MLP}_{\mathbf{E}}(\mathcal{V}(\hat{\mathbf{X}}))\|_F, \quad (2)$$

where $\text{MLP}_{\mathbf{X}}(\cdot)$ is the frozen MLP of the image branch and $\text{MLP}_{\mathbf{E}}(\cdot)$ is the MLP of the event branch. In other words, we leverage the information encoded in the image embeddings of the original image to guide the learning of the image embeddings for the reconstructed image. The Frobenius norm serves as a metric to quantify the dissimilarity between these embeddings, enabling the model to refine its representation of image information and enhance the consistency between the two image modalities.

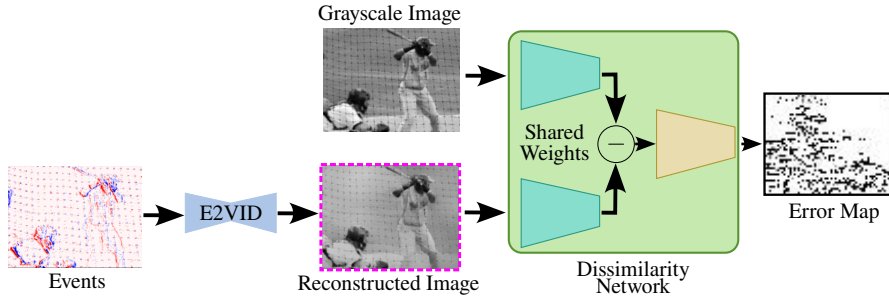


Fig. 3. Dissimilarity network takes the grayscale and reconstructed images as input, and it outputs an error map to reweight the mask loss. E2VID is unable to reconstruct the stripes and hence considered a high error area by the dissimilarity network.

3.4 Feature Distillation

To provide further guidance from the image branch to the event one, we minimize the Frobenius norm of the matrix differences of the outputs of each layer in the transformer decoder of Mask2Former between the original image and reconstructed image:

$$\mathcal{L}_f = \frac{1}{L} \sum_{i=1}^L \|\mathbf{D}_i - \widehat{\mathbf{D}}_i\|_F, \quad (3)$$

where L represents the total number of decoder layers, and \mathbf{D} and $\widehat{\mathbf{D}}$ are the output matrices of each layer in the image and events branch, respectively.

3.5 Mask Re-weighting

While maintaining E2VID frozen, poor reconstructions may occur, prompting the imposition of classification on inadequately reconstructed regions. However, this approach risks compromising model performance where reconstructions are accurate. To address this, we introduce a dissimilarity network to discern differences between the grayscale image and its reconstructed counterpart from events. Illustrated in Figure 3, this network comprises two convolutional layers. The first layer shares weights for grayscale and reconstructed images, followed by a rectified linear unit (ReLU) activation, while the second layer is followed by a sigmoid activation function $\sigma(\cdot)$. The squared error between the grayscale and reconstructed images' outputs feeds into the second convolutional layer, generating a reweighting map. As shown in Figure 3 that a notable discrepancy exists between the grayscale image and the reconstructed image, particularly concerning the stripes on the shirt in this example. Consequently, this discrepancy in the error map indicates reduced importance attributed to that specific area. Mathematically, this process can be expressed as:

$$\mathbf{M} = \sigma(\text{conv}_2(\text{ReLU}(\text{conv}_1(\mathbf{X})) - \text{ReLU}(\text{conv}_1(\widehat{\mathbf{X}}))))^2, \quad (4)$$

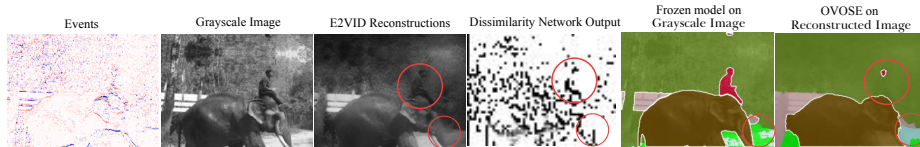


Fig. 4. The impact of reweighting the mask loss, influenced by the dissimilarity between the grayscale and reconstructed images. Poorly reconstructed areas such as the person and the elephant’s trunk lead to their exclusion in the reweighting process.

where \mathbf{X} is the grayscale original image and $\hat{\mathbf{X}}$ is the reconstructed image from E2VID. We use \mathbf{M} to re-weight our distillation loss at the level of segmentation maps \mathcal{L}_m . This loss for the i th stream is given by:

$$\mathcal{L}_m = \mathbf{M} \odot \mathcal{L}_{CE} \left(\mathbf{Y}_i, \hat{\mathbf{Y}}_i \right), \quad (5)$$

where $\mathbf{Y}_i \in \mathbb{R}^{H \times W \times |\mathcal{C}^{seen}|}$ is the output segmentation map of the image branch, $\hat{\mathbf{Y}}_i \in \mathbb{R}^{H \times W \times |\mathcal{C}^{seen}|}$ is the output segmentation map of the event branch, \mathcal{L}_{CE} is the cross-entropy loss, and \odot is the point-wise product between matrices. We illustrate a sample output of the dissimilarity network in Figure 4. It can be seen that OVOSE successfully ignores the areas where the error is high, for example, the person and the elephant’s trunk.

3.6 Category Label Supervision

As we have access to category labels in the training set, we follow [34] to compute a category label loss \mathcal{L}_c . To this end, we compute the probability of a mask belonging to one of the training categories using a cross-entropy loss with a learnable temperature parameter as in [34].

The total loss of OVOSE is given by:

$$\mathcal{L}_{final} = \mathcal{L}_t + \mathcal{L}_f + \lambda_m \mathcal{L}_m + \lambda_c \mathcal{L}_c \quad (6)$$

where $\lambda_m = 5.0$ and $\lambda_c = 2.0$ are regularization parameters.

4 Experiments and Results

In this section, we present an overview of the baseline methods for comparison, details on the training and test data, and a comprehensive analysis of both quantitative and qualitative results. Subsequently, we delve into ablation studies to validate the components of OVOSE.

4.1 Experimental Framework

Baseline methods. As there is no open-vocabulary semantic segmentation method for events, we benchmark OVOSE against leading UDA methods for

semantic segmentation in event-based data. However, this comparison is unfair with OVOSE since UDA methods: (i) know the set of unseen classes, (ii) have access to one or multiple labeled source datasets, and (iii) have access to the unlabeled testing dataset for adaptation. We also compare our algorithm with straightforward adaptations of open-vocabulary semantic segmentation methods in regular images. This adaptation consists of reconstructing a grayscale image from the stream of events with the E2VID model (similar to our event branch) and using this as input to the open-vocabulary model. More precisely, OVOSE is compared with the UDA methods E2VID [25], EV-transfer [20], VID2E [11] and ESS [28]. For the open-vocabulary methods, we compare our algorithm against E2VID+OpenSeed [36] and E2VID+ODISE [34]. We use the mean Intersection over Union (mIoU) and pixel accuracy metrics for the quantitative evaluations.

Training data. As described in the problem statement, OVOSE requires a training dataset \mathcal{X}_{train} where we have the stream of events and their corresponding grayscale images. To address this need, we leverage synthetic training data as introduced by [25]. This synthetic dataset was generated using the event simulator ESIM [24], which simulated MS-COCO images [18]. The dataset comprises 1,000 sequences, each spanning 2 seconds, with grayscale images and events of dimensions 240×180 . For our training purposes, we resize the images and events to dimensions 256×192 , ensuring compatibility and optimization for our network. Following [34], we use MS-COCO classes for category label supervision.

Evaluation datasets. We evaluate the open-vocabulary performance of OVOSE on two popular event camera-based self-driving datasets and TimeLens++:

- *DAVIS Driving Dataset* (DDD17) is a dataset for semantic segmentation in autonomous driving. Alonso et al. [1] pre-trained an Xception network to generate semantic pseudo-labels which were consolidated into six classes: flat (road and pavement), background (construction and sky), object (pole, pole group, traffic light, traffic sign), vegetation, human, and vehicle.
- *DSEC-Semantic* DSEC [13] consists of high-resolution images 1440×1080 , synchronized events 640×480 , and semantic labels [28] are generated using a state-of-the-art semantic segmentation method. The fine-grained labels for 19 classes are consolidated into 11 classes: background, building, fence, person, pole, road, sidewalk, vegetation, car, wall, and traffic sign.
- *TimeLens++* We show qualitative results of applying directly OVOSE in the Time lens++ dataset [29]. Time Lens++ consists of high-resolution events of size 970×625 .

Implementation details. We keep the image branch frozen and fine-tune the MLP and Mask2Former of the event branch. We use convolutions with 3×3 kernel size with a stride of 2 and padding 1 for the dissimilarity network. We set the learning rate as 1×10^{-5} . We train OVOSE using Adam optimizer with a batch size of 4 on Nvidia Ampere GPU with 48GB of RAM. We initialize the weights of text and image encoder from [22], text-to-image diffusion UNet from [23], and MLP and Mask2former from [34]. For qualitative results we use the open source code and weights provided by [28] and [34]. For OpenSeed [36], we use their provided open-source code and weights.

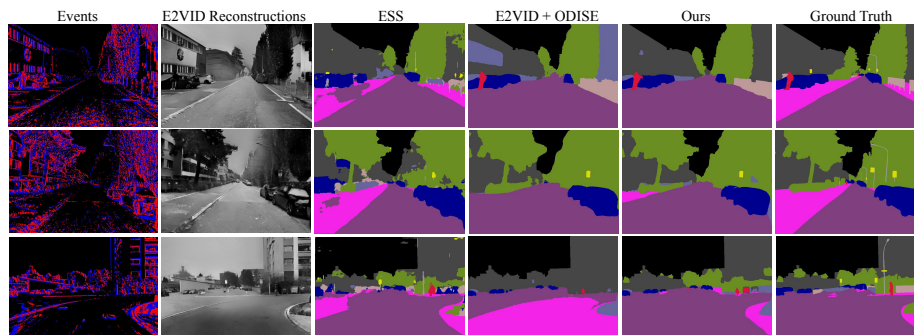


Fig. 5. Qualitative samples from ESS in UDA closed-set, E2VID+ODISE, and OVOSE in open vocabulary setting. As compared to ESS and E2VID+ODISE, OVOSE produce accurate and less noisy predictions even though it is trained on a synthetic dataset.

4.2 Results

Semantic segmentation on DSEC-Semantic. Figure 5 shows a qualitative comparison of OVOSE against ESS and ODISE. We obtain the qualitative results for the ESS in Figure 5 by utilizing the official model provided by the authors. Even though the ESS method is trained on real events in a closed-set setting, its overall semantic segmentation across the entire image appears noisy. Notably, it fails to segment vehicles in several instances accurately. For E2VID+ODISE in an open vocabulary setting, it misclassifies parts of the road and buildings. In contrast, OVOSE excels in recognizing traffic signs and delivers superior overall semantic segmentation across the entire image. This is particularly evident in its ability to discern intricate details and provide accurate segmentations, showcasing its robustness and adaptability in the driving scenario of DSEC.

Table 1 presents the comparison of OVOSE against the baseline methods on the DSEC-Semantic dataset. Even though the UDA methods are trained in the closed-set setting using real events and urban street datasets similar to DSEC, OVOSE surpasses their performance by a significant margin. Notably, our model improves the state-of-the-art UDA method ESS [28] by a substantial 3.57% in the mIoU metric. Similarly, OVOSE outperforms E2VID+ODISE by 4.83% in the mIoU metric, showcasing the superior performance of OVOSE for semantic segmentation in event-based data.

Semantic segmentation on DDD17. Figure 6 presents a qualitative comparison between OVOSE, ESS, and E2VID+ODISE. ESS effectively segments poles and other event-specific objects that may not be available in the ground truth. However, it struggles with inaccuracies and noise when segmenting cars. In contrast, OVOSE operating in an open vocabulary setting showcases the ability to segment traffic signs even when unlabeled in the ground truth images. Notably, OVOSE excels in differentiating between classes, particularly with vehicles, and demonstrates more accurate and noise-resistant semantic segmentation compared to the ESS. OVOSE also outperforms E2VID+ODISE by providing clearer

Table 1. Results on DSEC Semantic in UDA and open-vocabulary setting. OVOSE not only outperforms all the UDA methods even though they are trained in closed-set settings but also translated open-vocabulary methods.

Method	Type	Problem Formulation	Training Data	Input	Acc (%) \uparrow	mIoU (%) \uparrow
EV-Transfer [20]	UDA	Closed-Set	Cityscapes+DSEC events+frames	events+frames	60.50	23.20
E2VID [25]	UDA	Closed-Set	Cityscapes+DSEC events+frames	events+frames	76.67	40.70
ESS [28]	UDA	Closed-Set	Cityscapes+DSEC events+frames	events+frames	84.04	44.87
E2VID+OpenSeed [36]	Translation	Open-Set	MS-COCO	events	65.25	32.82
E2VID+ODISE [34]	Translation	Open-Set	MS-COCO	events	81.24	43.61
OVOSE (Ours)	Distillation	Open-Set	EV-COCO	events	85.67	48.44

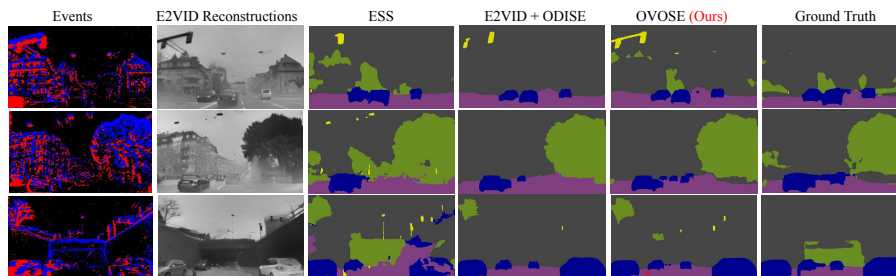


Fig. 6. Qualitative samples from DDD17 in UDA closed-set, E2VID+ODISE, and OVOSE in open vocabulary setting. OVOSE does better overall predictions, especially vehicles, persons, vegetation, and construction.

object recognition and segmentation. This comparison emphasizes the need for knowledge distillation and OVOSE’s capacity to deliver precise semantic segmentation even in scenarios with incomplete or noisy ground truth annotations.

When it comes to quantitative evaluation on the DDD17 dataset, the presence of noisy ground truth labels, as highlighted by [28], poses challenges. Additionally, the previous work by Ev-SegNet [1] merged several labels, further complicating the evaluation process. To address this, we consider the original classes pre-merge for predictions, subsequently merging them to facilitate a comparison with the ground truth. Table 2 summarizes the quantitative comparison of OVOSE with the baseline methods on the DDD17 dataset. OVOSE significantly improves the UDA state-of-the-art ESS method [28] by 0.93% in the mIoU metric. Notably, our algorithm outperforms E2VID+OpenSeed and E2VID+ODISE by large margins, underscoring the significance of knowledge distillation and mask reweighting in semantic segmentation for event cameras.

Qualitative Results on the Time Lens++ Dataset. We use the Time Lens++ [29] dataset to evaluate the open-vocabulary capabilities of OVOSE since no open-vocabulary event dataset is available. Some qualitative results are illustrated in Figure 7. We see in Figure 7 that OVOSE performs high-quality segmentation of various classes such as trees, traffic lights, people, cars, buildings, sidewalk, and roads which demonstrates OVOSE preserved open vocabulary ca-

Table 2. Results on DDD17 in UDA and open-vocabulary setting. OVOSE outperforms all the UDA methods trained in closed-set settings but also translated open vocabulary methods

Method	Type	Problem Formulation	Training Data	Input	Acc (%) \uparrow	mIoU (%) \uparrow
EV-Transfer [20]	UDA	Closed-Set	Cityscapes+DDD17 events+frames		47.37	14.91
E2VID [25]	UDA	Closed-Set	Cityscapes+DDD17 events+frames		83.24	44.77
VID2E [11]	UDA	Closed-Set	Cityscapes+DDD17 events+frames		85.93	45.48
ESS [28]	UDA	Closed-Set	Cityscapes+DDD17 events+frames		87.86	52.46
E2VID+OpenSeed [36]	Translation	Open-Set	MS-COCO	events	33.25	17.95
E2VID+ODISE [34]	Translation	Open-Set	MS-COCO	events	84.63	48.12
OVOSE (Ours)	Distillation	Open-Set	EV-COCO	events	88.84	53.39



Fig. 7. Open-vocabulary performance of OVOSE on the Time Lens++ dataset [29].

pabilities. Furthermore, OVOSE can recognize the bike and train even though the reconstruction from E2VID is very noisy and far from optimal.

4.3 Ablation Studies

Mask Reweighting. We conduct an ablation study to validate the reweighting schemes and knowledge distillation techniques in OVOSE. To that end, We evaluate OVOSE without training; this is equivalent to performing E2VID+ODISE. Then, we analyzed the impact of reweighting by using only the distillation pipeline, with reweighting based upon cosine similarity (CS) on the original and reconstructed image, by using the squared differences of the output of Stable Diffusion (SD) from the image and event branches and finally with our Dissimilarity Network (DN) corresponding to OVOSE. Table 3 shows the results of the ablation study. Distillation resulted in a notable increase in accuracy by 3.8% and mIoU by 3.4%, showcasing the effectiveness of this approach. Moreover, the introduction of the dissimilarity network for reweighting the loss function gives

Table 3. Ablation of distillation strategies of OVOSE on DSEC Semantic dataset.

Method	Acc (%) \uparrow	mIoU (%) \uparrow
Baseline	81.2	43.61
Distillation	85.0	47.01
Distillation+Reweight CS	85.1	48.01
Distillation+Reweight SD	85.1	47.32
Distillation+Reweight DN	85.6	48.44

Table 4. Ablation studies on DSEC-Semantic dataset. MLP’s is MLP layer of OVOSE and MG is the Mask Generator

Ablation	Parameters		Acc (%) \uparrow	mIoU (%) \uparrow
	λ_c	λ_m		
Loss	5.0	5.0	84.47	46.47
	5.0	2.0	84.70	46.44
	2.0	5.0	85.67	48.44
	<u>MLP’s</u>	<u>MG</u>		
Finetuning	\checkmark	\times	82.46	44.34
	\times	\checkmark	83.19	44.68
	\checkmark	\checkmark	85.67	48.44
	<u>FireNet [27]</u>	<u>E2VID [25]</u>		
Image-Reconstructor	\checkmark	\times	82.85	43.94
	\times	\checkmark	85.67	48.44
	<u>DSEC Classes</u>	<u>Ours</u>		
Text Prompts	\checkmark	\times	84.04	46.50
	\times	\checkmark	85.67	48.44

an additional improvement of 1.43% in the mIoU metric. **Loss Parameters.** We study the impact of varying λ_c and λ_m in Equation 6 on model performance. As shown in Table 4, the optimal combination of $\lambda_c = 2.0$ and $\lambda_m = 5.0$ achieved the highest accuracy 85.67% and mean Intersection over Union (mIoU) 48.44%, indicating that a lower weight on caption loss and a higher weight on mask loss enhance performance.

Fine-tuning Ablation. We explore the effects of fine-tuning the MLP layers and the Mask Generator (MG), individually and together. From Table 4, we observe that fine-tuning only the MLP resulted in 82.46% accuracy and 44.34% mIoU, while fine-tuning only MG slightly improved performance to 83.19% accuracy and 44.68% mIoU. However, simultaneous fine-tuning of both MLP’s and MG yielded the best results with 85.67% accuracy and 48.44% mIoU, underscoring the necessity of fine-tuning both components together.

Image Reconstructor Ablation. We replaced E2VID [25] with FireNet [27] and reported results in Table 4. E2VID significantly outperformed FireNet, achieving 85.67% accuracy and 48.44% mIoU which indicates that E2VID’s recurrent neural network handles temporal information well and outputs higher quality reconstructions for our downstream task.

Ablation on Text prompts. We further assess the effectiveness of two text prompt configurations: directly using the DSEC-Class name and our text prompts shown in Table 1 and Table 2 in the supplementary document. Results in Table 4 show that DSEC-Classes configuration resulted in 84.04% accuracy and 46.50% mIoU. In contrast, our text prompts configuration achieved superior performance with 85.67% accuracy and 48.44% mIoU, demonstrating that the detailed prompts enhance the model’s performance more effectively.

5 Conclusion

In this work, we introduced OVOSE, the first open-vocabulary semantic segmentation algorithm designed for event-based data. Comprising grayscale image and event branches, each equipped with a pre-trained foundation model, our approach leverages synthetic data for Knowledge Distillation from regular images to enhance semantic segmentation in events. OVOSE employs distillation at multiple stages of the foundation model, enhancing its effectiveness with a mask reweighting strategy through a dissimilarity network. We evaluate OVOSE in the DDD17 and DSEC-Semantic datasets and compare it against with existing methods in UDA close-set semantic segmentation and foundation models adapted to the event domain with E2VID. Our algorithm outperforms all these models, offering a promising avenue for research in open-vocabulary semantic segmentation tailored for event cameras.

Acknowledgments

We acknowledge financial support from the PNRR MUR project PE0000013-FAIR and from the Sapienza grant RG123188B3EF6A80 (CENTS).

References

1. Alonso, I., Murillo, A.C.: EV-SegNet: Semantic segmentation for event-based cameras. In: IEEE/CVF CVPRW (2019)
2. Binas, J., Neil, D., Liu, S.C., Delbrück, T.: Ddd17: End-to-end davis driving dataset. ArXiv [abs/1711.01458](#) (2017)
3. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems* **32** (2019)
4. Chen, G., Cao, H., Conradt, J., Tang, H., Rohrbein, F., Knoll, A.: Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine* **37**(4), 34–49 (2020)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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 (2018)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Computer Vision – ECCV 2018*. p. 833–851. Springer-Verlag, Berlin, Heidelberg (2018)

7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF CVPR. pp. 1290–1299 (2022)
8. Cho, H., Kim, H., Chae, Y., Yoon, K.J.: Label-free event-based object recognition via joint learning with image reconstruction from events. In: Proceedings of the IEEE/CVF ICCV. pp. 19866–19877 (October 2023)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**(1), 154–180 (2020)
11. Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Video to events: Recycling video datasets for event cameras. In: Proceedings of the IEEE/CVF CVPR. pp. 3586–3595 (2020)
12. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF ICCV. pp. 5633–5643 (2019)
13. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* **6**, 4947–4954 (2021)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
15. Jian, D., Rostami, M.: Unsupervised domain adaptation for training event-based networks using contrastive learning and uncorrelated conditioning. In: Proceedings of the IEEE/CVF ICCV. pp. 18721–18731 (October 2023)
16. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022)
17. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. 2023 IEEE/CVF CVPR pp. 7061–7070 (2022)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
19. Messikommer, N., Georgoulis, S., Gehrig, D., Tulyakov, S., Erbach, J., Bochicchio, A., Li, Y., Scaramuzza, D.: Multi-bracket high dynamic range imaging with event cameras. In: 2022 IEEE/CVF CVPRW. pp. 546–556 (jun 2022)
20. Messikommer, N., Gehrig, D., Gehrig, M., Scaramuzza, D.: Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters* **7**(2), 3515–3522 (2022)
21. Park, S., Kwak, N.: Feed: Feature-level ensemble for knowledge distillation. arXiv preprint arXiv:1909.10754 (2019)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
23. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)

24. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on Robot Learning (2018)
25. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)* (2019)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF CVPR. pp. 10684–10695 (2022)
27. Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D.: Fast image reconstruction with an event camera. In: IEEE Winter Conf. Appl. Comput. Vis. (WACV). pp. 156–163 (2020)
28. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: Ess: Learning event-based semantic segmentation from still images. In: ECCV (2022)
29. Tulyakov, S., Bochicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D.: Time Lens++: Event-based frame interpolation with non-linear parametric flow and multi-scale fusion. *IEEE/CVF CVPR* (2022)
30. Wang, L., Chae, Y., Yoon, K.J.: Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In: Proceedings of the IEEE/CVF ICCV. pp. 2135–2145 (2021)
31. Wang, L., Chae, Y., Yoon, S.H., Kim, T.K., Yoon, K.J.: Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 608–619 (2021)
32. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF CVPR. pp. 8256–8265 (2019)
33. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: European Conference on Computer Vision. pp. 588–604. Springer (2020)
34. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF CVPR. pp. 2955–2966 (2023)
35. Yang, Y., Pan, L., Liu, L.: Event camera data pre-training. In: Proceedings of the IEEE/CVF ICCV. pp. 10699–10709 (October 2023)
36. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF ICCV. pp. 1020–1031 (2023)
37. Zhao, L., Peng, X., Chen, Y., Kapadia, M., Metaxas, D.N.: Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. *2020 IEEE/CVF CVPR* pp. 6527–6536 (2020)
38. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF CVPR. pp. 989–997 (2019)

Supplementary

Muhammad Rameez Ur Rahman¹, Jhony H. Giraldo², Indro Spinelli¹,
Stéphane Lathuilière², and Fabio Galasso¹

¹ Sapienza University of Rome, Italy

² LTCI, Télécom Paris, Institut Polytechnique de Paris, France

In this supplementary document, we provide details about the text prompts used in our model, fully supervised results, and comprehensive ablation studies of our method.

1 Text Prompts

Tables 1 and 2 present the textual prompts used to create CLIP text embeddings for two datasets: DSEC-Semantic and DDD17. Table 1 outlines the classes and corresponding text prompts utilized for the DSEC-Semantic dataset. The prompts are designed to cover a wide range of visual entities and contexts to enhance the model’s understanding. Each class is associated with various descriptive terms to ensure comprehensive semantic coverage. For instance, the "person" class includes prompts like "man," "women," "person," "pedestrian," and "children," which cater to different possible appearances and contexts of individuals in the dataset. Similarly, Table 2 lists the classes and their corresponding text prompts used in the DDD17 dataset.

Table 1. The textual prompts utilized in the DSEC-Semantic dataset [3] for creating the CLIP text embeddings.

#	Classes and the corresponding Text Prompts
0	background: sky
1	building: building, skyscraper, house,
2	fence: fence
3	person: man, women, person, pedestrian, children
4	pole: pole
5	road: road, highway, expressway
6	sidewalk: sidewalk, track, footpath, foot pavement pedestrian zone
7	vegetation: vegetation, tree, trees, palm tree, bushes, plants, trees, lawn
8	car: car, bicycle, bike, truck, locomotive, bus, motorcycle
9	wall: wall, stone wall
10	traffic-sign: traffic sign, stop sign, traffic light

Table 2. The textual prompts utilized in the DDD17 dataset [1] for creating the CLIP text embeddings.

#	Classes and the corresponding Text Prompts
0	flat: road, highway, pavement, street, lane track
1	background: building, sky, house, wall, fence
2	object: pole, light pole, street light, traffic light, traffic signal, traffic sign
3	vegetation: vegetation, tree, grass, meadow, plants
4	human: human, person, rider, people
5	vehicle: vehicle, car, truck, bus, motorcycle, bicycle, train, van, locomotive

2 DSEC supervised results

We compare OVOSE against the previous methods EV-SegNet [1], ESS [3] and HMNET [2] in a fully supervised setting. To that end, we train OVOSE in a supervised closed-set setting on the DSEC-Semantic dataset [3]. We report the results in Table 3 and it can be seen that OVOSE outperforms ESS by a large margin of 3.62% mIoU in Table 3.

Table 3. Comparison of OVOSE with state-of-the-art methods on DSEC Semantic dataset in a supervised setting. OVOSE outperforms state-of-the-art by 3.62% mIoU

Method	Acc (%)	mIoU (%)
EV-SegNet [1]	88.61	51.76
ESS [3]	89.25	51.57
ESS [3]	89.37	53.29
HMNET-B3 [2]	88.70	51.20
HMNET-L3 [2]	89.90	55.00
OVOSE (ours)	89.45	58.62

References

- Alonso, I., Murillo, A.C.: Ev-segnet: Semantic segmentation for event-based cameras. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1624–1633 (2019)
- Hamaguchi, R., Furukawa, Y., Onishi, M., Sakurada, K.: Hierarchical neural memory network for low latency event processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22867–22876 (2023)
- Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: Ess: Learning event-based semantic segmentation from still images. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 341–357. Springer Nature Switzerland, Cham (2022)