ESEG: Event-Based Segmentation Boosted by Explicit Semantic-Edge Guidance

Yucheng Zhao[1], Gengyu Lyu[1], Ke Li[1], Zihao Wang[1], Hao Chen[2], Zhen Yang[1], Yongjian Deng[1]*

[1]College of Computer Science, Beijing University of Technology
[2]School of Computer Science and Engineering, Southeast University
yzhao836@gatech.edu, {lyugengyu@, tokeli@emails., yangzhen@, yjdeng@}bjut.edu.cn, rex.wangzihao@gmail.com
haochen303@seu.edu.cn

Abstract

Event-based semantic segmentation (ESS) has attracted researchers' attention recently, as event cameras can solve problems such as under/over-exposure or motion blur that are difficult for RGB cameras to handle. However, event data are noisy and sparse, resulting in difficulties for the model to locate and extract reliable cues from their sparse representations, especially when performing pixel-level tasks. In this paper, we propose a novel framework ESEG to alleviate the dilemma. Given that event signals relate closely to moving edges, instead of proposing complex structures to expect them to recognize those reliable edge regions behind event signals on their own, we introduce the explicit edge-semantic supervision as a reference to let the ESS model globally optimize semantics, considering the high confidence of event data in edge regions. In addition, we propose a fusion module named Density-Aware Dynamic-Window Cross Attention Fusion (D^2CAF), in which the density perception, cross-attention, and dynamic window masking mechanisms are jointly imposed to optimize edge-dense feature fusion, leveraging the characteristics of event cameras. Experimental results on DSEC and DDD17 datasets demonstrate the efficacy of the ESEG framework and its core designs.

Code — https://github.com/cheesewawa/ESEG

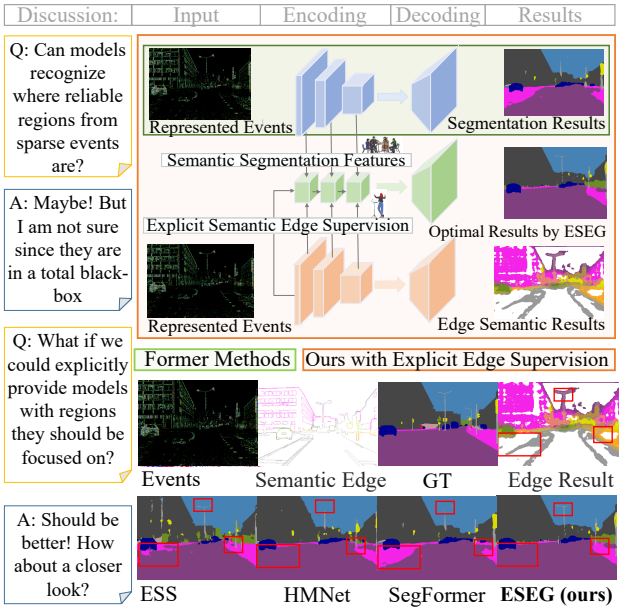

Figure 1: Our framework aims to assist models in focusing on reliable regions of event-based representations by providing edge-semantic guidance, instead of expecting networks to learn this ability without explicit assistance.

Introduction

Semantic segmentation plays a significant role in computer vision (Hao, Zhou, and Guo 2020), and has been successfully applied to numerous fields including autonomous driving (Papadeas et al. 2021; Muhammad et al. 2022), defect detection (Usamentiaga et al. 2022; Xu et al. 2022) and computer-aided diagnosis (Huang et al. 2022; Qureshi et al. 2023). Recently, large-scale models like SAM (Kirillov et al. 2023) and DINOv2 (Oquab et al. 2023) have made remarkable progress in performance gains and scene generalization.

However, these semantic segmentation models usually face challenges when the input data has quality issues, *e.g.,* failure in extreme light conditions or high-speed motion scenes. Event cameras show their potential to alleviate the above issues (Gallego et al. 2020). Due to the special working principles such as asynchronously capturing brightness changes at each pixel, these cameras have higher temporal resolution and dynamic range, and limited power requirements, and new studies demonstrated that they are promising to offer more dependable data in extreme conditions for better performance (Deng, Chen, and Li 2024; Liu et al. 2024).

Therefore, researchers turn their focus to exploring the design of segmentation models for event data which hold peculiar data formats with noise and sparsity. Ev-SegNet (Alonso and Murillo 2019) is the first model to perform the task based on event data, followed by research employing new mechanisms such as posterior attention (Jia et al. 2023) and SNNs (Hareb and Martinet 2024). Other attempts include cross-modal distillation (Wang et al. 2021), transfer learning from still images (Sun et al. 2022), adapting large-scale pre-trained models (Yao et al. 2024), and hybrid pseudo-labeling (Jing et al. 2024). Recognizing that the event-based repre-

*Corresponding Author

sentation is sparse and commonly conveys valuable knowledge at limited regions, previous studies aim to let the ESS model learn how to locate these regions and extract reliable messages from them implicitly by introducing novel learning architectures or transfer learning techniques. Based on the above observation, a question is raised naturally: *whether we could explicitly inform models which regions they should pay more attention to?*

Inspired by the above observation, this work introduces a novel ESS learning framework (Figure 1), ESEG, in which the edge-semantic messages are directly imposed as supervision to make the network aware of the important and reliable regions that should be paid more attention to explicitly. First, considering there is no existing semantic edge label in the event-based datasets, the SELSAM (Semantic Edge Label with SAM) algorithm is proposed to generate ground truth labels. The algorithm is based on SAM rather than other models, because it can filter out most of the internal texture edges of objects by returning clear object boundary contour according to the regional semantics. The internal texture edges of objects may harm the semantic segmentation model in understanding different object boundaries. Second, instead of using edge semantics just as a supplementary for dense feature enhancement (Chen et al. 2020; Yuan et al. 2020), we intend to arrange two roles for the extracted edge semantics, *i.e.,* an information density indicator and edge locating distinguisher. Specifically, D^2CAF (Density-Aware Dynamic-Window Cross-Attention fusion) combines the DIM (Density Indicator Matrix) for feature mapping, dynamic window masking for edge-aware aggregation, and cross-attention for feature fusion, which allows edge information to offer more assistance while not letting irrelevant non-edge information harm the feature encoding during the optimization of the network.

By setting up the three branches including edge-semantic, dense-semantic, and fusion branches, the ESEG can learn high-confidence features under explicit edge-semantic supervision and refine the exploitation of event data across various regions and levels.

The main contributions are summarized in four folds. (1) The ESEG structure leverages the event camera's nature in sensitivity to moving edges by introducing explicit edge-semantic supervision. (2) This work provides a new type of label for common ESS datasets, such as DSEC and DDD17, generated from a SAM-based pipeline. (3) The introduced D^2CAF module with density recognition, cross-attention fusion, and dynamic window masking optimizes the fusion of dense-semantic features under the guidance of sparser edge-semantic features, making two types of features contribute jointly for final segmentation. (4) Our approach achieves state-of-the-art performance on multiple well-recognized and widely used datasets DSEC and DDD17.

Related Work

Semantic Segmentation

Earlier semantic segmentation models employed FCNs (Long, Shelhamer, and Darrell 2015) and CRFs(Chen et al. 2014). Inspired by the encoder-decoder architecture (Ronneberger, Fischer, and Brox 2015; Badrinarayanan, Kendall, and Cipolla 2017), several later models continued upgrading the architecture. For instance, to enlarge the receptive field, the DeepLab family (Chen et al. 2017a,b, 2018) employs dilated convolutional models, and several studies (Xie et al. 2021; Strudel et al. 2021; Gu et al. 2022; Zhang et al. 2022) employs Visual Transformers for the task. Motivated by large-scale models, some powerful approaches such as SAM (Kirillov et al. 2023) and DINOv2 (Oquab et al. 2023) are proposed and driven by predefined prompts. The other stream that is related to ours considers edge factors in semantic segmentation, *e.g.,* adding additional edge detection branches (Chen et al. 2020; Bertasius, Shi, and Torresani 2016), or focusing on edge refinement (Yuan et al. 2020). However, edge-related methods encounter their bottleneck in RGB-based segmentation as the colorful and high-resolution RGB images can handle the object boundaries well in normal cases thus limiting the efficacy of supplementary edge features. Our work gets inspiration from these works and provides the ESS field with new insights. Through combining information density indicators, edge-aware distinguishing, and cross-attention mechanisms, our method can fully exploiting the edge-semantic cues while fitting the data nature of event cameras.

Semantic Edge Detection

Semantic edge detection aims to identify boundaries and their semantic meaning. Most methods related to this task are based on deep learning, *e.g.,* CASENet (Yu et al. 2017) employs a skip-layer architecture for fusion and a multi-label loss function for supervision. Some recent research continues the attempt to optimize the task such as employing lightweight networks (DFF) (Hu et al. 2019) or using diverse deep supervision (DDS-R) (Liu et al. 2022). Due to DFF's characteristics of being lightweight and easy optimization, this model is adopted in our work as the semantic edge detection model.

Event-Based Semantic Segmentation

Event-based semantic segmentation research is still in its infancy. Ev-SegNet (Alonso and Murillo 2019) is the first research that performs the task using event input only and provides the commonly used DDD17 dataset through an RGB-event simulator. Subsequent work has made various attempts to improve the learning capability of the ESS model, such as transfer knowledge from RGB domains (Sun et al. 2022; Wang et al. 2021; Rebecq et al. 2019), combining with additional modalities (Zhang et al. 2023; Xie et al. 2024; Ghasemzadeh and Shouraki 2023), introducing temporal-sensitive learning architectures (Hamaguchi et al. 2023; Zhang et al. 2024) or augmenting the training data with generative models (Gehrig et al. 2020). The mainstream methods with leading performance primarily focus on introducing novel network architectures or performing transfer learning to enhance the models' ability to recognize reliable messages behind event data in a black-box mode. In this work, we directly introduce edge information that correlates to event trigger distributions as explicit supervision instead. With additional edge supervision, our model can optimize

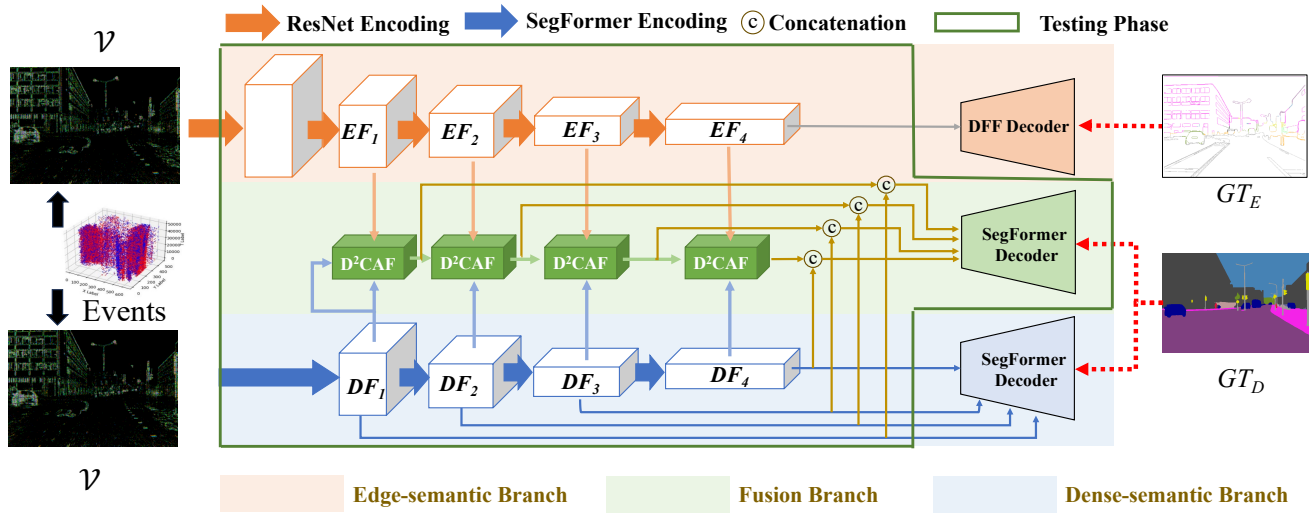

Figure 2: The design of the ESEG structure. There are three steps exhibited sequentially. (i) The edge-semantic branch is trained alone and supervised by GT_E. (ii) The edge-semantic branch is frozen whereas the rest of the learning architecture is used to be optimized by GT_D supervision. (iii) Only the network inside the green box is utilized during the testing phase.

semantic representations through recognizing more reliable motion edge regions, and thereby enhance the segmentation performance.

Approach

Problem Formulation

Event Representation. Event streams $(\{x_i, y_i, t_i, p_i\}_N)$ can be integrated as a grid-like representations $(V \in \mathbb{R}^{B\times H\times W})$ following the approach in (Zhou and Tuzel 2018; Messikommer et al. 2022), where B represents three integration bins. In this work, we take $50ms$ events for constructing $\mathcal{V}$ and set B as 3 for all experiments and datasets.

Pipeline. The overall pipeline of the ESEG is shown in Figure 2, where three branches can be distinguished: the edge-semantic, dense-semantic and fusion branches.

The edge-semantic branch takes the event representation $\mathcal{V}$ as input to aware semantic edge regions from event data, where the DFF model (Hu et al. 2019) with ResNet34 backbone (He et al. 2016) is adopted for this branch. Since there are no currently available labels for this task, we generate the semantic edge supervision GT_E based on given still images and corresponding segmentation labels. As shown in Figure 2 (i), the ResNet34 is composed of 5 ResNet blocks, and we utilize the last four output features $(\{EF_i\}_{i=1}^4)$ to guide the feature fusion and refinement for semantics learned from the dense-semantic branch.

The dense-semantic branch is supervised by original semantic segmentation labels GT_D with identical input $\mathcal{V}$ as the edge-semantic branch, where the SegFormer architecture is employed for this branch. We divide the encoder of Segformer into four parts and note the output of these four blocks as dense-semantic features $(\{DF_i\}_{i=1}^4)$.

In the fusion branch, a series of fusion modules, D^{2}CAF, are equipped to exploit edge-semantic features (EF) as guidance to refine and fuse multi-level dense-semantic features (DF). For i-th fusion module D^{2}CAF$_\text{i}$, it simultaneously refines the dense-semantic features (DF_i) and fuses it with the output of D^{2}CAF$_{i-1}$ as formulated in Eq. 1.

$$DF'_i = \text{D}^2\text{CAF}_\text{i}(EF_i, DF'_{i-1}, DF_i). \quad (1)$$

After obtaining fused features from four stages $(\{DF'_i\}_{i=1}^4)$, we concatenate them with dense-semantic features $(\{DF_i\}_{i=1}^4)$ accordingly and then input these concatenated features into the simple Segformer decoder for final prediction.

Problems. From the above description of the working pipeline of the ESEG framework, it is easy to find there are two problems that have to be addressed before leveraging the powerful edge information for event-based segmentation enhancement. (i) Lack of required labels. Currently, there are no existing event-based segmentation datasets which includes semantic edge labels (GT_E) for training. (ii) The challenge to exploit edge-semantic features. For performing a dense semantic prediction task, inappropriate usage of edge relative features even impairs the final results since it may force the network to focus on edge regions while losing decision fidelity on smooth regions. This issue is even more serious for event-based models since their inputs are in a sparse format as well.

In this work, we introduce the Semantic Edge Label based on SAM algorithm (**SELSAM**) and Density-Aware Dynamic-Window Cross Attention Fusion module (**D^{2}CAF**) to alleviate the label shortage and edge-dense fusion dilemmas, respectively. In the following, we will detail these two contributions in sequence.

SELSAM: Semantic Edge Label with SAM

The SELSAM algorithm aims to generate the semantic edge labels as the edge-semantic labels GT_E to train the edge-

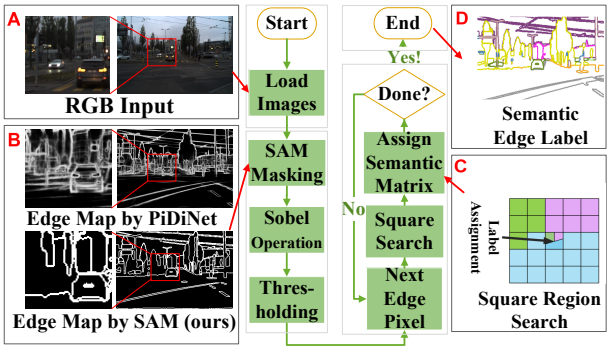

Figure 3: Procedure of the edge-semantic label generation.

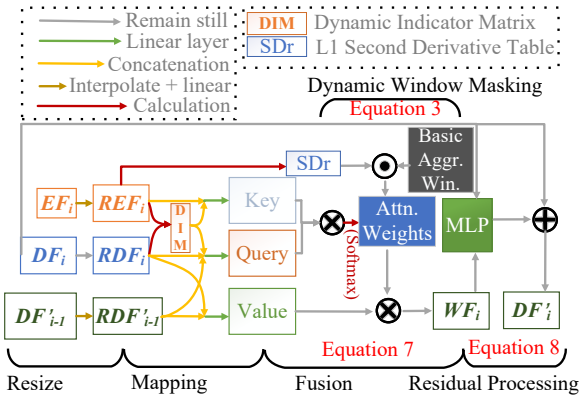

Figure 4: Structure of the D^{2}CAF (Density-Aware Dynamic-Window Cross-Attention Fusion) Module

semantic branch. Specifically, SAM is adopted as the base model for the SELSAM since the edge maps produced by SAM include fewer internal textures within each object than other methods, *e.g.,* PiDiNet (as shown in Figure 3.B). We attribute these advantages to the strong robustness to different application scenarios provided by SAM while most edge detection methods may fail in unseen samples. Internal edges are normally irrelevant to the detection of segmentation boundaries so learning from them might turn the model's focus to the irrelevant internal parts instead of important segmentation boundaries, impairing the performance.

There are several stages to generate edge-semantic labels: (i) Edge maps without semantic information are produced first. For instance, the input images (Figure 3.A) are segmented by SAM, followed by converting the masks to a grayscale image, computing gradients, combining gradients into an edge strength image, and transforming into a binary edge map (Figure 3.B). (ii) Next, the semantic edge map is generated by iterating over every edge pixel and counting the semantic labels within the 5×5 neighbor area (Figure 3.C). Finally, a label image with the size of $H \times W \times Cls$ for each input image is produced by the SELSAM, where H and W are the input resolution and Cls represents the segmentation categories of datasets. The visualization can be seen in Figure 3.D.

D^{2}CAF: Density-Aware Dynamic-Window Cross Attention Fusion

The detailed architecture of D^{2}CAF is illustrated in Figure 4, where an i-th fusion module D^{2}CAF$_\text{i}$ aims to fuse DF_i and DF'_{i-1} under the guidance of EF_i to form the refined output DF'_i. In particular, DF'_{i-1} and EF_i are first resized to the same size as $DF_i \in \mathbb{R}^{C_i^d \times H_i^d \times W_i^d}$, as RDF'_{i-1} and REF_i, via a linear layer followed by a pooling operation. In the mapping phase, the DIM (Density Indicator Matrix) is first calculated. Mapping is then performed by concatenating different features together with the DIM and going through linear layers as shown in the figure. The cross-attention fusion phase allows the model to fuse the dense features (DF_i, DF'_{i-1}) guided by the reference EF_i. During fusion, the dynamic window masking mechanism is applied to allow the aggregating process to consider neighbor information $w.r.t$ edge or non-edge regions adaptively. Finally, the residual processing phase consists of a simple MLP network and shortcut connections. In the following, we will detail the motivation and calculation of the introduced DIM and dynamic window masking sequentially.

DIM: Density Indicator Matrix Edge features ($REF_i \in \mathbb{R}^{C_i^e \times H_i^d \times W_i^d}$) are commonly with a relatively sparse format, while features (DF_i) for dense segmentation are considered denser. The former carries valuable information mainly at edge areas while the latter carries information within all areas of the feature maps. Based on these observations, we tend to provide a reference so that our model can adaptively identify which regions should be given more attention for optimization after considering the differences in information distribution of features from different sources (REF_i, DF_i) in the same region. To achieve this, we build a density information indicator for features from different sources. For any location of these features, L1-distance, L2-distance, and information entropy across all feature channels are computer and result in two indicators both with the size of $(H_i^d, W_i^d, 3)$. Then, we concatenate these two indicators for building the final DIM $w.r.t$ $REF_i \& DF_i$ with the size of $(H_i^d, W_i^d, 6)$.

In particular, L1-&L2-distance and information entropy can indicate information density at the position from different perspectives, where a higher value represents higher density. Therefore, introducing the DIM that contains inductive biases such as L1-&L2-distance and the information entropy can give the fusion module insights into which positions might contain vital edge information or have a large discrepancy in information distributions, aiding our model in identifying edge areas, segmenting boundaries or recognizing regions that require to be more concentrated.

Dynamic Window Masking During cross-attention aggregation, we argue that a larger range of neighbor information should be considered when coming across edge regions, while less such information should be included for non-edge regions. Intuitively, edge regions commonly contain semantic boundaries from different objects. For such regions, it is

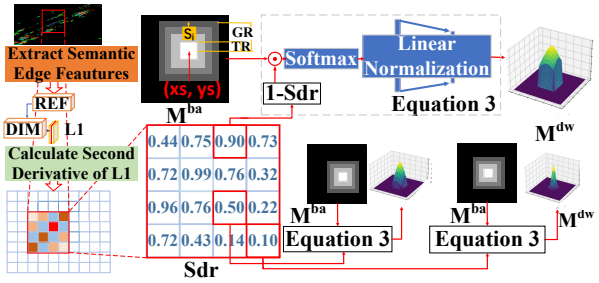

Figure 5: Procedure of dynamic-window masking.

necessary to increase the receptive field and clearly judge the surrounding situation, to obtain the accurate location of the boundaries of several objects. Instead, a non-edge smooth area is usually located inside the object, thus considering a large range may include edges from other objects and introduce harmful noise.

In this paper, we introduce a dynamic window mask mechanism (Figure 5) to respond to the above heuristic issues. For a specific coordinate (x_s, y_s), the mechanism first introduces a basic aggregation window ($M^{ba}_{(x_s,y_s)} \in \mathbb{R}^{H^d_i \times W^d_i}$) based on approximate Gaussian function, and changes the influence degree of neighborhood on the central coordinate based on this window and the calculated edge region probability. To generate $M^{ba}_{(x_s,y_s)}$, transparent radius TR and gradient radius GR are defined first. Next, we fill this mask with values that conform to the normal distribution $S(loc)$. The visualization is referred to in Figure 5.

$$S(loc) = f(U(loc)) = \begin{cases} e^{(-\frac{U(loc)^2}{2})} & \text{if } loc \leq GR \\ 0 & \text{if } loc > GR \end{cases}, \quad (2)$$

where $U(loc) = \frac{max(loc-TR,0)}{GR-TR}$ and loc is a positive integer calculated by $max(|x - x_s|, |y - y_s|)$, (x, y) is an arbitrary coordinate in $M^{ba}_{(x_s,y_s)}$. This paper uses TR and GR values as two and six consistently.

After getting the basic aggregation window $w.r.t\ (x_s, y_s)$, we try to adjust the attention proportion of its neighborhood features according to the probability that the location is located in the edge region. The purpose of this mechanism is to make the attention mechanism get more help from the neighbor information in the edge region, and less interference from the neighbor region in the non-edge region. In specific, to estimate the edge probability, we first retrieve a feature matrix (F^{L1}_i) from DIM that corresponds to the L1-distance of REF_i, then compute the Sdr_i matrix in which each value is derived from the second derivative of F^{L1}_i. The design concept here is that F^{L1}_i can reflect the density of edge features, and the presence of feature edges will cause the density change nearby to be more obvious than that of other parts. Therefore, using the second derivative operator to detect the sharp change of F^{L1}_i value can reflect the probability of edge existence to a certain extent. Finally, we can achieve the dynamic window mask as formulated in Eq. 3.

$$M^{dw}_{(x_s,y_s)} = \text{Norm}(\text{Softmax}(100(1 - Sdr_i(x_s, y_s)M^{ba}_{(x_s,y_s)})), \quad (3)$$

where Norm represents the normalization operation. With Eq. 3, we can obtain $M^{dw}_i \in \mathbb{R}^{H^d_i \times W^d_i \times W^d_i \times H^d_i}$ for the masking calculations in the cross-attention mechanism that integrates $M^{dw}_{(x,y)} \in \mathbb{R}^{H^d_i \times W^d_i}$ masks from all features positions. As shown in Figure 5, due to the variations of second derivative values, the basic aggregation masks will be multiplied by a different value before the Softmax operation and normalization, making the neighbor areas more or less influential in the attention mechanism as the surface smoother or sharper.

Cross-Attention Fusion The query, key, and value mapping processes for the cross-attention mechanism in the D^2CAF_i module can be described as:

$$Q_i = Linear_Q([REF_i, DIM]), \quad (4)$$

$$K_i = Linear_K([DF_i, RDF'_{i-1}, DIM]), \quad (5)$$

$$V_i = Linear_V([DF_i, RDF'_{i-1}]), \quad (6)$$

where $[\cdot, \cdot]$ represents the concatenation operation and $Linear$ represents linear projection layer. The cross-attention mechanism with dynamic window masking in the module can be described as:

$$WF_i = \text{Softmax}(\frac{(Q_i \times K^T_i) \cdot M^{dw}_i}{\sqrt{d_k}})V_i, \quad (7)$$

where WF is the attention-weighted feature. Then, a shortcut connection is employed to the weighted feature as formulated in Eq. 8 for obtaining the refined dense-semantic features under the edge information guidance.

$$DF'_i = \text{MLP}([WF_i, DF_i]) + DF_i, \quad (8)$$

where MLP is the multi-layer perceptron network inside the fusion module.

In this way, the fusion process based on the cross-attention mechanism can adjust its attention to vital edge areas under the guidance of edge-semantic reference and use high-quality edge information for optimization. As for the entire structure of our learning framework, multiple sequential fusion modules (D^2CAF) allow the model to use high-quality edge information for multi-level and multi-scale feature refinement and fusion, ultimately increasing the model's accuracy in locating object boundaries in the segmentation scenes and therefore improve the overall performance.

Experiments

Experiments for this paper are conducted on two commonly used datasets for event-based semantic segmentation, DSEC-Semantic (Gehrig et al. 2021) and DDD17 (Binas et al. 2017). The ESEG method is compared with numerous SOTA methods in this section. They can be categorized into two folds: training and inference with only event data, *e.g.,* Ev-SegNet (Alonso and Murillo 2019), EvSegFormer (Jia et al. 2023), SpikingEDN (Zhang et al. 2024), HMNet (Hamaguchi et al. 2023) and ESS-Sup (Sun et al. 2022); Learning or combining knowledge from RGB data domain such as E2VID (Rebecq et al. 2019), Vid2E (Gehrig et al. 2020), EvDistill (Wang et al. 2021) and ESS (Sun et al. 2022).

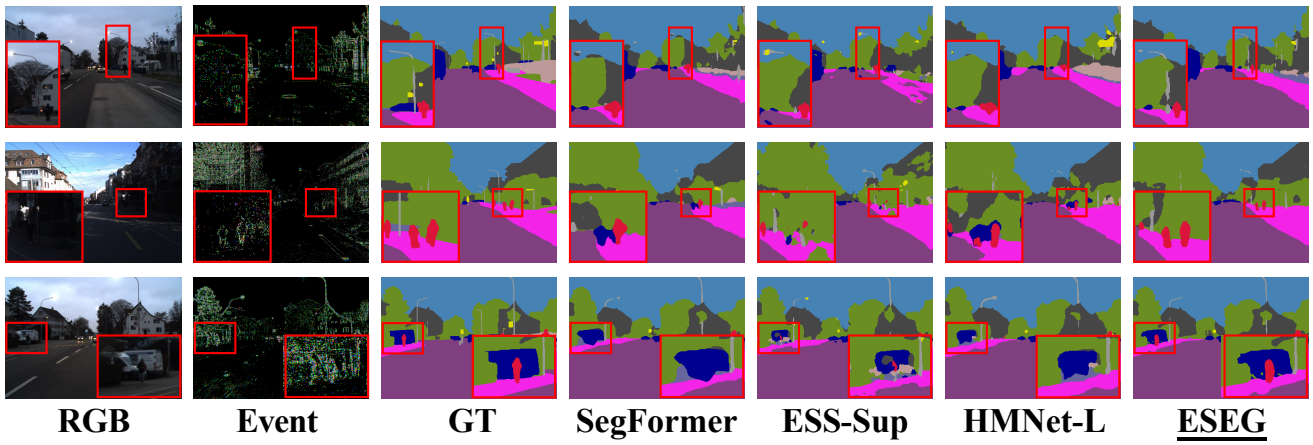

Figure 6: Visualization comparisons on the DSEC between ESEG and other representative methods.

Evaluation on the DSEC-Semantic Dataset

Dataset and Experiment Details DSEC is a dataset for driving scenarios with both event and RGB data. DSEC-Semantic provides semantic labels for eleven sequences from the DSEC dataset, where 11-class labels are used for training and evaluation. The DFF method with ResNet34 backbone is adopted for the edge-semantic branch and the SegFormer architecture is for the dense-semantic branch. We deploy two versions of the ESEG structure here: a baseline version (ESEG-B) uses MiT-b0 as Segformer's backbone, whereas the large version (ESEG-L) uses MiT-b1.

During training, data augmentation techniques including random resizing and flipping are employed. The AdamW optimizer and the Polynomial LR scheduler are used with the initial learning rate as $6e{-}5$. The model was trained for 40 epochs with the batch size as 4. The mean Intersection over Union (mIoU) and Accuracy (Acc) are used as the evaluation metrics. As for the input, we stack voxel girds ($\mathcal{V}$) based on the fixed time duration ($50ms$) following (Alonso and Murillo 2019; Jia et al. 2023; Zhang et al. 2024). All experiments are implemented using Pytorch on an RTX 3090.

Comparison with Other Methods on DSEC Table 1 shows that both ESEG-B and ESEG-L hold leading or comparable performance over other methods.

First, the ESEG-B significantly outperforms the baseline model SegFormer by 1.7%, indicating that the introduction of explicit edge-semantic supervision with the new fusion method effectively optimize the feature extraction and refinement processes.

Second, the performance enhancement by ESEG compared to methods that use event data only such as Ev-SegNet, EvSegFormer, SpikingEDN, ESS-Sup, and HMNet, suggests that the unique approach of leveraging event data's sensitivity and higher dependability around edges is effective, and promising to offer new research path on this specific task. Moreover, ESEG holds comparable accuracy with approaches that learn both event and RGB knowledge domains such as Vid2E, E2VID, and ESS. Such results indicate that even without extra modal assistance during learning, an explicit edge awareness design that fully leverages the nature of event cameras can also exploit semantics from event data comprehensively. Visualization results in Figure 6 also validate our conclusion, *e.g.,* the proposed ESEG can segment polls, traffic signs, and traffic lights better than other compared methods obviously, with reasonably smoother and more refined contours for buildings, sidewalks, and vehicles.

Method	Venue	Backbone	Acc	mIoU
E2VID	TPAMI'19	U-Net	80.06	44.08
ESS	ECCV'22	E2VID	84.17	45.38
Ev-SegNet	CVPRW'19	Xception	88.61	51.76
ESS-Sup*	ECCV'22	E2VID	89.08	52.30
HMNet-B	CVPR'23	HMNet-B1	88.70	51.20
HMNet-L	CVPR'23	HMNet-L1	89.80	55.00
SpikingEDN	TNNLS'24	SNN	-	53.17
SegFormer	NeurIPS'20	MiT-b0	90.02	54.19
ESEG-B	-	MiT-b0	90.22	55.93
ESEG-L	-	MiT-b1	91.47	**57.55**

Table 1: Comparison results of methods on the DSEC. *: The input is the same as our method while the original ESS-Sup builds their input $w.r.t$ fixed event numbers.

Dataset and Experiment Details The DDD17 Dataset includes multiple modalities such as event-based vision data, RGB images, and IMU data. The dataset also provides semantic labels with six categories for driving scenarios and objects, supporting tasks like object detection and semantic segmentation. Training setups including learning architectures, data augmentation techniques, the optimizer, batch size, number of epochs, and evaluation metrics are the same as the experiments on DSEC. The learning rate scheduler is the same as DSEC but the initial learning rate is 1×10^{-3} for DDD17.

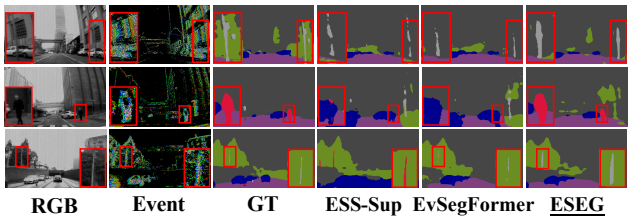

Figure 7: Visualizations on the DDD17 of ESEG and other representative methods.

Method	Venue	Backbone	Acc	mIoU
E2VID	TPAMI'19	U-Net	85.84	48.47
Vid2E	CVPR'20	Xception	90.19	56.01
EvDistill	CVPR'21	DeepLabV3+	-	58.02
ESS	ECCV'22	E2VID	88.43	53.09
Ev-SegNet	CVPRW'19	Xception	89.76	54.81
ESS-Sup*	ECCV'22	E2VID	89.12	56.67
EvSegFormer	TIP'23	MiT-b1	**94.72**	54.41
SpikingEDN	TNNLS'24	SNN	-	53.15
SegFormer	NeurIPS'20	MiT-b0	89.34	52.76
ESEG-B	-	MiT-b0	89.64	57.01
ESEG-L	-	MiT-b1	90.68	**59.97**

Table 2: Comparison results of methods on the DDD17

Evaluation on the DDD17 Dataset

Comparison with Other Methods on DDD17 Table 2 exhibits the results of comparison experiments with ours and other ESS methods with leading performances on DDD17. Basically, the conclusion of the performance comparisons on the DDD17 dataset is consistent with which summarized from the DSEC dataset. For instance, our approach achieves a significant performance gain compared to the baseline model Segformer. Also, our work is able to outperform other models that are trained using event data only such as Ev-SegNet, SpikingEDN, ESS-Sup, and EvSegFormer. As shown in Figure 7, boosted by edge-semantic features, the proposed method can achieve better performance in segmenting challenging scenarios such as pedestrians and traffic poles in the distance. Interestingly, we find that the mIoU of ESEG-B is slightly lower than EvDistill which leverages transfer learning techniques. We attribute this problem to insufficient training of our model due to the relatively poor image quality of ddd17. Instead, benefiting from powerful RGB priors, EvDistill could optimize the model easier. Therefore, when equipped the ESEG with higher capable backbones, our proposed method outperform the EvDistill by a large margin.

Ablation Studies

This section aims to analyze the efficacy of our proposed fusion strategy and report the performance of different settings in Table 3. First, it is obvious that all fusion settings (B-F)

Var.	Fusion Strategy	Cross Attn.	DIM Mapping	Window	Acc	mIoU
A	No Fusion	-	-	none	89.42	54.19
B	Concatenation	-	-	none	89.90	54.48
C	Cross Attn.	+	-	none	89.95	54.83
D	DIM Cross Attn.	+	+	none	90.02	55.22
E	DIM Cross Attn.	+	+	M^{ba}	90.04	55.33
F	**D^{2}CAF**	+	+	M^{dw}	**90.22**	**55.93**

Table 3: Ablation studies with ESEG-B on the DSEC.

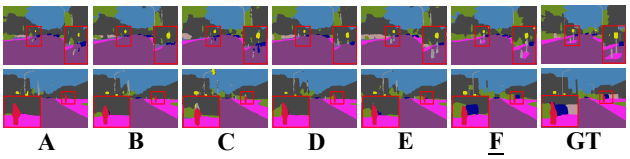

Figure 8: Results of ablation experiments

outperform A, which does not use edge-semantic supervision, indicating that explicit edge guidance is a valuable reference that can improve the feature representation. We can find from comparisons between C and D that the introduction of DIM enhances the fusion effects and we attribute this to the information density cues that might provide the attention process with more beneficial indications. Next, the better performance of E compared to D demonstrates that masking distant weights can prevent irrelevant distant harmful interference, *e.g.,* the setting D falsely judge a part of poles as road while E corrects this mistake (Figure 8). Finally, when equipped the setting E with dynamic window masking, our proposed D^{2}CAF module achieves the highest accuracy and provides the best visualized results, suggesting its capability to fuse and refine dense-semantic for performance enhancements by fully exploiting the references conveyed from edge features.

Conclusion

This paper proposes a novel learning framework named ESEG that leverages explicit edge-semantic supervision for boosting event-based semantic segmentation. Considering the characteristics of event data that are commonly triggered by moving edges, the edge-semantic information is believed to be a reference to guide the model to be aware of which regions are reliable or require more attention. To achieve the above target, we first generate edge-semantic labels via a SAM-based pipeline, then introduce the edge-dense fusion module D^{2}CAF with mechanisms derived from information density and edge awareness for better dense-semantic representation. Extensive experiments validate the efficacy and rationality of the ESEG and its core designs. Notably, since our framework has the potential for migrating to other visual tasks, we argue that it may open new research avenues for event-based model learning.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (62203024, 92167102, 61873220, 62102083, 62173286, 61875068, 62177018, 62306020), the Natural Science Foundation of Jiangsu Province (BK20210222), the R&D Program of Beijing Municipal Education Commission (KM202310005027), the Research Grants Council of Hong Kong (CityU11206122) and the Young Elite Scientist Sponsorship Program by BAST (BYESS2024199).

References

Alonso, I.; and Murillo, A. C. 2019. EV-SegNet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Bertasius, G.; Shi, J.; and Torresani, L. 2016. Semantic segmentation with boundary neural fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3602–3610.

Binas, J.; Neil, D.; Liu, S.-C.; et al. 2017. DDD17: End-to-end DAVIS driving dataset. *arXiv preprint arXiv:1711.01458*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017a. 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.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, X.; Lian, Y.; Jiao, L.; Wang, H.; Gao, Y.; and Lingling, S. 2020. Supervised edge attention network for accurate image instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, 617–631. Springer.

Deng, Y.; Chen, H.; and Li, Y. 2024. A Dynamic GCN with Cross-Representation Distillation for Event-Based Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1492–1500.

Gallego, G.; Delbruck, T.; Orchard, G. M.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A.; Conradt, J.; Daniilidis, K.; and Scaramuzza, D. 2020. Event-based Vision: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3586–3595.

Gehrig, M.; Aarents, W.; Gehrig, D.; and Scaramuzza, D. 2021. DSEC: A Stereo Event Camera Dataset for Driving Scenarios. *IEEE Robotics and Automation Letters*.

Ghasemzadeh, M.; and Shouraki, S. 2023. Semantic Segmentation Using Events and Combination of Events and Frames. In *International Conference on Artificial Intelligence and Smart Vehicles*, 167–181. Springer.

Gu, J.; Kwon, H.; Wang, D.; Ye, W.; Li, M.; Chen, Y.-H.; Lai, L.; Chandra, V.; and Pan, D. Z. 2022. Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12094–12103.

Hamaguchi, R.; Furukawa, Y.; Onishi, M.; and Sakurada, K. 2023. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22867–22876.

Hao, S.; Zhou, Y.; and Guo, Y. 2020. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 406: 302–321.

Hareb, D.; and Martinet, J. 2024. EvSegSNN: Neuromorphic Semantic Segmentation for Event Data. *arXiv preprint arXiv:2406.14178*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Y.; Chen, Y.; Li, X.; and Feng, J. 2019. Dynamic feature fusion for semantic edge detection. *arXiv preprint arXiv:1902.09104*.

Huang, S.-Y.; Hsu, W.-L.; Hsu, R.-J.; and Liu, D.-W. 2022. Fully convolutional network for the semantic segmentation of medical images: A survey. *Diagnostics*, 12(11): 2765.

Jia, Z.; You, K.; He, W.; Tian, Y.; Feng, Y.; Wang, Y.; Jia, X.; Lou, Y.; Zhang, J.; Li, G.; et al. 2023. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32: 1829–1842.

Jing, L.; Ding, Y.; Gao, Y.; Wang, Z.; Yan, X.; Wang, D.; Schaefer, G.; Fang, H.; Zhao, B.; and Li, X. 2024. HPL-ESS: Hybrid Pseudo-Labeling for Unsupervised Event-based Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23128–23137.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Liu, Y.; Cheng, M.-M.; Fan, D.-P.; Zhang, L.; Bian, J.-W.; and Tao, D. 2022. Semantic edge detection with diverse

deep supervision. *International Journal of Computer Vision*, 130(1): 179–198.

Liu, Y.; Deng, Y.; Chen, H.; and Yang, Z. 2024. Video Frame Interpolation via Direct Synthesis with the Event-based Reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8477–8487.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Messikommer, N.; Gehrig, D.; Gehrig, M.; and Scaramuzza, D. 2022. Bridging the Gap between Events and Frames through Unsupervised Domain Adaptation. *IEEE Robot. Autom. Lett.*, 7(2): 3515–3522.

Muhammad, K.; Hussain, T.; Ullah, H.; Del Ser, J.; Rezaei, M.; Kumar, N.; Hijji, M.; Bellavista, P.; and de Albuquerque, V. H. C. 2022. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 22694–22715.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Papadeas, I.; Tsochatzidis, L.; Amanatiadis, A.; and Pratikakis, I. 2021. Real-time semantic image segmentation with deep learning for autonomous driving: A survey. *Applied Sciences*, 11(19): 8802.

Qureshi, I.; Yan, J.; Abbas, Q.; Shaheed, K.; Riaz, A. B.; Wahid, A.; Khan, M. W. J.; and Szczuko, P. 2023. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90: 316–352.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 1964–1980.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Sun, Z.; Messikommer, N.; Gehrig, D.; and Scaramuzza, D. 2022. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, 341–357. Springer.

Usamentiaga, R.; Lema, D. G.; Pedrayes, O. D.; and Garcia, D. F. 2022. Automated surface defect detection in metals: a comparative review of object detection and semantic segmentation using deep learning. *IEEE Transactions on Industry Applications*, 58(3): 4203–4213.

Wang, L.; Chae, Y.; Yoon, S.-H.; Kim, T.-K.; and Yoon, K.-J. 2021. 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*, 608–619.

Xie, B.; Deng, Y.; Shao, Z.; and Li, Y. 2024. EISNet: A Multi-Modal Fusion Network for Semantic Segmentation with Events and Images. *IEEE Transactions on Multimedia*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xu, H.; Yan, Z.; Ji, B.; Huang, P.; Cheng, J.; and Wu, X. 2022. Defect detection in welding radiographic images based on semantic segmentation methods. *Measurement*, 188: 110569.

Yao, B.; Deng, Y.; Liu, Y.; Chen, H.; Li, Y.; and Yang, Z. 2024. SAM-Event-Adapter: Adapting Segment Anything Model for Event-RGB Semantic Segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 9093–9100. IEEE.

Yu, Z.; Feng, C.; Liu, M.-Y.; and Ramalingam, S. 2017. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5964–5973.

Yuan, Y.; Xie, J.; Chen, X.; and Wang, J. 2020. Segfix: Model-agnostic boundary refinement for segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 489–506. Springer.

Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; and Stiefelhagen, R. 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*.

Zhang, R.; Leng, L.; Che, K.; Zhang, H.; Cheng, J.; Guo, Q.; Liao, J.; and Cheng, R. 2024. Accurate and Efficient Event-based Semantic Segmentation Using Adaptive Spiking Encoder-Decoder Network. *IEEE Transactions on Neural Networks and Learning Systems*, 1–1.

Zhang, W.; Huang, Z.; Luo, G.; Chen, T.; Wang, X.; Liu, W.; Yu, G.; and Shen, C. 2022. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12083–12093.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4490–4499.