Know Where You Are From: Event-Based Segmentation via Spatio-Temporal Propagation

Ke Li[1], Gengyu Lyu[1], Hao Chen[2], Bochen Xie[3], Zhen Yang[1], Youfu Li[3], Yongjian Deng[1]*

[1]College of Computer Science, Beijing University of Technology
[2]School of Computer Science and Engineering, Southeast University
[3]Department of Mechanical Engineering, City University of Hong Kong
{tokeli@emails., lyugengyu@, yangzhen@, yjdeng@}bjut.edu.cn, haochen303@seu.edu.cn
{boxie4-c@my.,meyfli@}cityu.edu.hk

Abstract

Event cameras have gained attention in segmentation due to their higher temporal resolution and dynamic range compared to traditional cameras. However, they struggle with issues like lack of color perception and triggering only at motion edges, making it hard to distinguish objects with similar contours or segment spatially continuous objects. Our work aims to address these often overlooked issues. Based on the assumption that various objects exhibit different motion patterns, we believe that embedding the historical motion states of objects into segmented scenes can effectively address these challenges. Inspired by this, we propose the ESS framework "Know Where You Are From" (KWYAF), which incorporates past motion cues through spatio-temporal propagation embedding. This framework features two core components: the Sequential Motion Encoding Module (SME) and the Event-Based Reliable Region Selection Mechanism (ER^{2}SM). SMEs construct prior motion features through spatio-temporal correlation modeling for boosting final segmentation, while ER^{2}SM adapts to identify high-confidence regions, embedding motion more precisely through local window masks and reliable region selection. A large number of experiments have demonstrated the effectiveness of our proposed framework in terms of both quantity and quality.

Code — https://github.com/SchuckLee/KWYAF

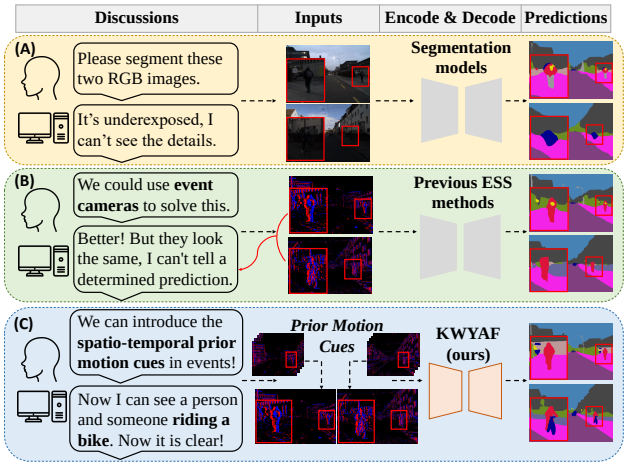

Figure 1: Illustration of our motivation. Event cameras can overcome the inherent limitations of RGB frames; however, the sparse data from event cameras lack color information and often display similar textures along object edges. By incorporating historical motion trajectories, we aim to equip models with prior motion states of each object for enhancing the semantic prediction capabilities.

Introduction

In recent years, progress in deep learning has led to the development of semantic segmentation models such as (Ronneberger, Fischer, and Brox 2015; Xie et al. 2021; Xu et al. 2024), significantly improving segmentation performance. These methods mainly rely on traditional frame-based cameras, which are difficult to handle the challenges of motion blur, over/underexposure in scenarios with high-speed or extreme lighting conditions. To tackle this issue, event cameras (Lichtsteiner, Posch, and Delbruck 2006; Suh et al. 2020) have emerged. Event cameras generate sparse data streams by capturing brightness changes of moving pixels only (Gallego et al. 2020), demonstrating superior performance in high-speed motion and complex lighting environments(Deng, Chen, and Li 2024; Liu et al. 2024). Exploring event-based semantic segmentation methods can significantly enhance system adaptability in complex dynamic environments, presenting significant research value.

Despite the aforementioned advantages of event cameras, current methods for event-based semantic segmentation (ESS) remain limited. Most existing methods tend to enhance the learning of sparse event data by adapting them to mature learning architectures designed (Alonso and Murillo 2019; Jia et al. 2023), employing domain adaptation techniques (Sun et al. 2022c; Jing et al. 2024), proposing cross-domain feature synthesis approaches (Jiang et al. 2023; Das Biswas et al. 2024; Yao et al. 2024) or introducing new types of activation functions (Kim, Chough, and Panda 2022; Zhang et al. 2024).

Generally, these approaches treat event data as a special

*Corresponding Author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

form of images for model customization, while neglecting the nature of event cameras that their outputs are sparse, lack of color perception and commonly triggered at moving edges only. These peculiar properties of event cameras comparing to RGB images would hinder models make correct prediction when come across situations such as segmenting spatially continuous or similarly shaped objects. In this work, we suggest that supplementing prior motion knowledge to segmentation scenes may address above problems, since the motion modes and states of objects are generally differs from each other. As shown in Figure 1, it is challenging to distinguish a person with or without a bicycle due to the sparse and edge-similar output format while we solve this issue by providing their historical motion information.

In this work, we propose a novel event-based semantic segmentation framework called Know Where You Are From (KWYAF), which integrates the historical spatio-temporal information of event signals to impart prior motion cues to the scene being segmented, thereby enhancing semantic prediction. We first uniformly divide the entire input event sequence into multiple segments along the temporal dimension and stack them into voxel grids as inputs to the framework. Subsequently, we introduce the Sequential Motion Encoding module (SME), where a series of SMEs are utilized to embed the motion process of the segmentation scene to construct joint spatio-temporal features. Within each SME, it starts by calculating the visual correlation volume (Corr) between features extracted from earlier motion features and the current event segment, then updating motion features containing current spatio-temporal cues and refining the event features jointly by cross-attention mechanisms. As time progresses, subsequent features gradually incorporate more dynamic context from earlier features, and finally form the prior-motion-awareness representation of the segmentation scene for final prediction.

Furthermore, considering the noisy and sparse nature of event data, we propose an Event-based Reliable Region Selection Mechanism (ER^{2}SM) to ensure the reliability of the extracted motion cues within SMEs. Particularly, the ER^{2}SM consists of two selective processes such as local window masking (LWM) and reliable region selection (RRS), which aims to avoid confusion from distant noisy areas and recognize high-confidence regions during motion embedding $w.r.t$ the event density. Finally, we integrate the spatio-temporal features across all input event segments and feed them into an off-the-shelf decoder CFM (Sun et al. 2022a), for the final prediction.

In summary, our contributions are as follows:

- We propose a novel ESS framework named KWYAF. KWYAF aims to inject the prior spatio-temporal propagation clues into the segmentation scene, thereby alleviating the segmentation accuracy limitations caused by the lack of color perception and high sparsity of event cameras.
- By introducing the Sequential Motion Encoding module (SME) , the model can endow the historical motion states to the segmentation scene, contributing the spatio-temporal information of event data to segmentation tasks.
- We design an Event-based Reliable Region Selection Mechanism (ER^{2}SM). By combining local window masking and reliable region selection based on event density, ER^{2}SM can effectively select high-confidence regional features for precise motion capturing with SMEs.
- Experimental results on DDD17 (Binas et al. 2017) and DSEC-Semantic (Gehrig et al. 2021) datasets demonstrate that our method achieves state-of-the-art performance $w.r.t$ event-based semantic segmentation.

Related Work

Semantic Segmentation

Semantic segmentation can be seen as a pixel-level classification task (Long, Shelhamer, and Darrell 2015). Initial works such as U-Net (Ronneberger, Fischer, and Brox 2015) and SegNet (Badrinarayanan, Kendall, and Cipolla 2017) present huge potential of this task in practice. Benefiting from the progress in Transformer-based networks (Xie et al. 2021; Cheng et al. 2022; Jain et al. 2023) and unsupervised learning techniques (Lan et al. 2023; Niu et al. 2024), models for segmentation achieve further performance improvement. To make segmentation models more general to dynamic scenes in real world, several video semantic segmentation (VSS) studies are proposed aiming to utilize temporal context for better understanding scenarios with complex motion conditions. Methods such as EFC (Ding et al. 2020) and ETC (Liu et al. 2020) adopt frame-by-frame inference, considering temporal consistency between frames as an additional constraint; TMANet (Wang, Wang, and Liu 2021) and TDNet (Hu et al. 2020) use attention-based approaches to aggregate features from sequential frames, enhancing target segmentation. Recently, MRCFA (Sun et al. 2022b) and CFFM++ (Sun et al. 2024) recognize the multi-scale interactive effects between reference frames and target frames and achieve SOTA performance on this task. This paper generalizes the motivation of VSS to the ESS field. Considering the disadvantages of events being sparse and generated only at edges, we believe that incorporating motion context of the scene can assist the network in making more accurate predictions. In addition, the SME and ER^{2}SM designed according to the imaging nature of event cameras assure the effectiveness and reliability in the process of forming event-based motion representations.

Event-based Semantic Segmentation (ESS)

The initial work (Alonso and Murillo 2019) in ESS presents event streams as frame-based representations and adapt Xception (Chollet 2017) structure for final prediction, in which the first ESS dataset DDD17 is provided. To further exploit the potential of event data, multiple perspectives of studies has been proposed by researchers. For example, methods in (Kim, Chough, and Panda 2022; Zhang et al. 2024) try to fit the asychronous nature of event data using spiking neural networks; approaches in (Jia et al. 2023; Smith, Doe, and Lee 2023) take advantages of powerful encoding ability in long-range features of Transformer-based networks for the ESS task. Another popular stream aims to utilize knowledge from other domains (Wang et al. 2021;

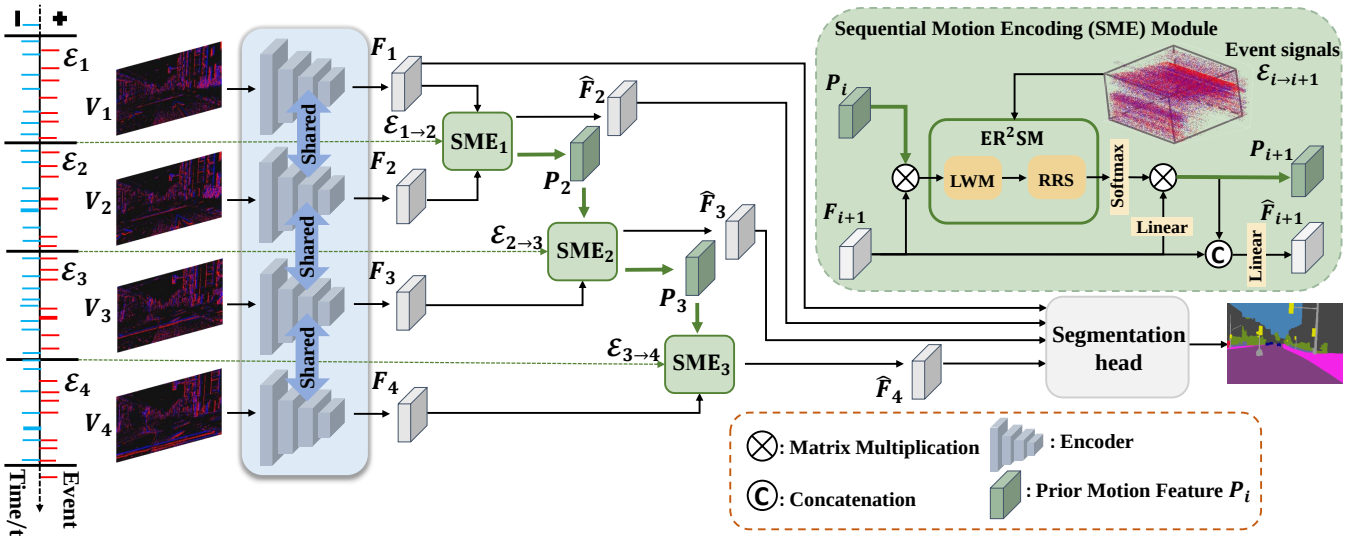

Figure 2: Overview of the proposed Know Where You Are From (KWYAF) framework. The green region on the right illustrates the Sequential Motion Encoding (SME) module, which infers and integrates spatio-temporal propagation messages between event segments. Within the SME module, we also highlight the Event-based Reliable Region Selection Mechanism (ER^{2}SM), specifically addresses the negative effects caused by event noise and sparsity. Through the combined effects of Local Window Masking (LWM) and Reliable Region Selection (RRS), ER^{2}SM empowers SME to identify high-confidence regions and aggregate reliable motion information.

Sun et al. 2022c; Das Biswas et al. 2024; Kong et al. 2024) such as RGB images or introduce self-supervised learning pipelines (Klenk et al. 2024) to enhance the optimization of ESS models. While large performance gain is achieved by previous works, they commonly overlook the issues of highly sparse events triggered mainly at motion edges and the lack of color perception in event cameras, resulting in difficulties of distinguishing objects with similar contours (Figure 1). To address this, we propose the KWYAF framework consisting of the SME and ER^{2}SM, endowing the segmented scene with historical motion knowledge, thereby enhancing its semantic prediction capabilities.

Method

Problem Formulation

Event Representation. Event data are typically defined as a set $\mathcal{E} = \{e_j\}_N = \{(x_j, y_j, t_j, p_j)\}_N$, where x, y are pixel coordinates, t represents the event's timestamp, and p indicates the polarity (either positive or negative), reflecting an increase or decrease in brightness changes. In this work, we divide the input event stream $\mathcal{E}$ into multiple segments equally by time interval. Following prior work (Wang et al. 2021), each segment $\mathcal{E}_k$ is converted into a voxel grid $V_k \in \mathbb{R}^{B\times H\times W}$ to be used as input for the network, where $\{H, W\}$ is the resolution of sensors and the integration channel B is defined as 3 in our approach.

Consequently, we obtain a sequence of voxel grids $\{V_k\}_{k=1}^{S}$, where we set S as 4 and focus on segmenting the latest one V_4.

Learning Pipeline. As shown in Figure 2, the proposed KWYAF takes a sequence of event representations ($\{V_k\}_{k=1}^{4}$) as input to extract dense features $\{F_i \in \mathbb{R}^{C\times H_F\times W_F}\}_{i=1}^{4}$ which represent the spatio-temporal semantics behind each event segments.

Here, C (equals to 256) represents the number of feature channels, and the size of $\{H_F, W_F\}$ equals to $\{\frac{H}{8}, \frac{W}{8}\}$. Then, these features are fed into the Sequential Motion Encoding modules (SMEs) for embedding spatio-temporal propagation to construct prior motion representations for enhancing the final prediction on V_4. Through combining with the proposed Event-based Reliable Region Selection Mechanism (ER^{2}SM), SME can extract accurate motion messages by identifying high-confidence regions across event segments, discarding the coding of motion regions that are likely to have errors. Finally, we input the refined features from all event segments into an off-the-shelf decoder CFM (Sun et al. 2022a) for segmentation. In the following, we will detail two core designs in our method, *i.e.,* SME and ER^{2}SM.

SME: Sequential Motion Encoding

As shown in Figure 2, the input to the i-th SME module is the prior motion embedding P_i output from the previous layer and the feature F_{i+1} of the current event segment $\mathcal{E}_{i+1}$. Our goal is to enable F_{i+1} to perceive how it has been evolved from earlier event segment scene. To this end, we first calculate the correlation volume (Teed and Deng 2020) between P_i and F_{i+1} to acquire spatio-temporal relationships between them as formulated in Eq. 1.

$$Corr_i = F_{i+1}P_i^T \in \mathbb{R}^{H_F\times W_F\times W_F\times H_F}, \quad (1)$$

where the first two dimensions (H_F, W_F) of $Corr$ represent the feature point locations in F_{i+1}, and the last two dimen-

sions (W_F, H_F) represent the similarity between each feature point in F_{i+1} $w.r.t$ all feature locations in P_i. However, as event data is noisy and commonly reliable only at motion edges, inevitable erroneous within the $Corr$ related to these locations, leading to imprecise spatio-temporal propagation modeling. To this end, we apply the ER^{2}SM to Corr for regularizing prior motion knowledge by recognizing and selecting high-confidence regions for feature aggregation, resulting in rectified correlation volume $\widehat{Corr}_i$. The details of ER^{2}SM are placed in the next section.

Next, similar to most re-weighting mechanisms (Zhang et al. 2022; Shen et al. 2024), we apply SoftMax to obtain a probability distribution and multiply it by the linearly projected F_{i+1}, resulting in the updated motion feature P_{i+1} with newly embedded spatio-temporal cues from $\mathcal{E}_{i+1}$ as shown in Eq. 2.

$$P_{i+1} = \text{SoftMax}\left(\widehat{Corr}_i / \sqrt{C}\right) \text{Linear}\left(F_{i+1}\right). \quad (2)$$

In this way, $P_{i+1} \in \mathbb{R}^{C \times H_F \times W_F}$ can perceive motion transforming process from the first event segment $\mathcal{E}_1$ to current segment $\mathcal{E}_{i+1}$. Intuitively, we can then update P_{i+1} as P_{i+2} with the next SME module for fusing the long-range motion propagation from $\mathcal{E}_1$ to $\mathcal{E}_{i+2}$. Notably, we suggest that although the calculated feature P_{i+1} contains sufficient motion information across event segments, it may lose some spatial contextual features that are also critical to segmentation. Therefore, to retain the semantics of F_{i+1} itself, we concatenate P_{i+1} with the original feature F_{i+1} and perform dimensional reduction, resulting in $\hat{F}_{i+1} \in \mathbb{R}^{C \times H_F \times W_F}$ as the input to the CFM for final prediction:

$$\hat{F}_{i+1} = \text{Linear}\left(\text{Concat}\left(P_{i+1}, F_{i+1}\right)\right). \quad (3)$$

By applying our SME module across the input event segments, we encode the motion propagation of the scene, enabling subsequent event features to build motion awareness from historical motions, thereby endowing the scene to be segmented with prior motion awareness.

ER^{2}SM: Event-based Reliable Region Selection

Considering the capturing characteristics of event cameras, the resulting motion correlation volume ($Corr$) may be impaired by event noise or regions with low event densities. To address this, we propose the Event-based Reliable Region Selection Mechanism (ER^{2}SM), illustrated in Figure 3, which focuses the network on aggregating high-confidence feature regions through the combined effect of local window masking and reliable region selection strategies, thereby enhancing the fidelity of extracted motion messages.

Local Window Masking (LWM). Event cameras naturally lack of color capturing and with relatively low resolution, so it is challenging for models to evaluate similarities between pixels during motion correlation calculation, *e.g.,* distinguishing whether these pixels are truly related or just have similar appearances. To alleviate these issues, we set an assumption that the motion messages are mostly related to or propagated from its spatio-temporal neighborhood. Based on this assumption, we propose a Local Window Masking

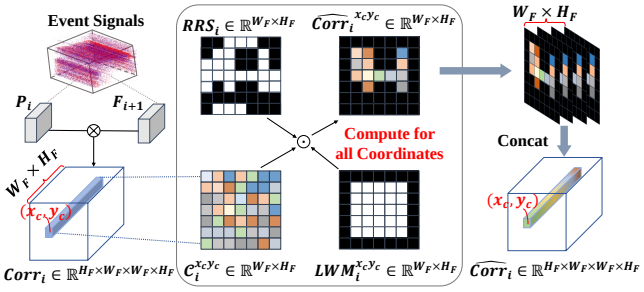

Figure 3: Illustration of our ER^{2}SM module. The correlation volume are refined jointly with the proposed local window masking and reliable region selection strategies.

(LWM) operation applied to the correlation volume ($Corr$) calculated by Eq. 1, making motion embedding in each location focus on its neighborhood and isolated with distant unrelated messages. As shown in Figure 3, for any coordinate (x_c, y_c) in $Corr_i$, we can get a corresponding feature map ($\mathcal{C}_i^{x_c y_c} \in \mathbb{R}^{W_F \times H_F}$), in which each value represents the correlation level of F_{i+1} and P_i. Next, we introduce the $\text{LWM}_i^{x_c y_c}$ with the same size of $\mathcal{C}_i^{x_c y_c}$, taking the location (x_c, y_c) as the center, assigning 1 to the adjacent region ($R_a \in \mathbb{R}^{m \times m}$) of the location (x_c, y_c) and 0 to the distant region. In this way, while may lose some distant relevant assistance, our model can largely benefits from avoiding the influence of distant noise and unrelated pixels with similar appearances. This work sets m as 7 for all datasets experimentally as referred in Table 5.

Reliable Region Selection (RRS). We observe that although event inputs are very sparse, due to pooling and aggregation of receptive fields, such inputs become dense high-level semantic features after being encoded by multi-layer neural networks. Such dense features are undoubtedly necessary for the goal we want to achieve, not only because we need to get dense semantic segmentation prediction, but also because we also want to find the association relationship from the joint representation of local appearance and high-level semantics in motion coding. However, we also realize that it is not intuitive or reasonable to search for motion association in the area where there is no event trigger in the input. Therefore, we also propose a simple yet effective reliable region selection (RRS) strategy to achieve high confidence region selection based on sparse event density besides the LWM. Specifically, as shown in Figure 3, we first use the events $\mathcal{E}_{i \to i+1}$ to generate an event histogram (Hist_i^e, Eq. 4) then obtain the RRS_i mask as formulated in Eq. 5. Notably, different from the LWM that requires to be computed for each coordinate in $Corr$, the RRS_i mask is consistent for all locations.

$$\text{Hist}_i^e(x, y) = \sum_{e_j \in \mathcal{E}_{i \to i+1}} \delta(x - x_j, y - y_j), \quad (4)$$

$$\text{RRS}_i(x, y) = \begin{cases} 1 & \text{if Hist}_i^e(x, y) \neq 0 \\ 0 & \text{if Hist}_i^e(x, y) = 0 \end{cases}. \quad (5)$$

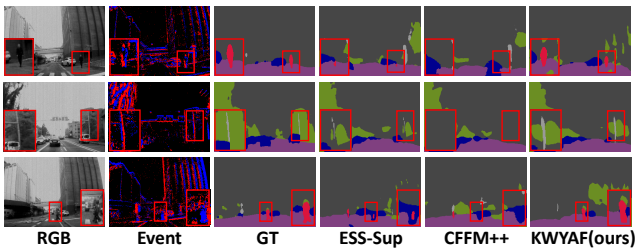

Figure 4: Qualitative results on the DDD17 dataset. Compared to leading ESS and VSS methods, our approach demonstrates reliable predictions for moving targets.

Finally, we can combine $\text{RRS}_i \in \mathbb{R}^{H_F \times W_F}$ with $\text{LWM}_i \in \mathbb{R}^{H_F \times W_F \times W_F \times H_F}$ for reliably selecting motion embedding regions by jointly considering sparse & noisy nature of event cameras, where the ER^{2}SM in Eq. 6 can be detailed as follows.

$$\begin{aligned} \widehat{Corr_i} &= \text{ER}^2\text{SM}\left(Corr_i\right) \\ &= Corr_i \cdot \text{LWM}_i \cdot \text{RRS}_i. \end{aligned} \tag{6}$$

Segmentation Head & Losses

This stage aims to integrate multi-segment spatio-temporal features for final prediction, and we achieve this target by adopting an off-the-shelf VSS decoder CFM (Sun et al. 2022a). This module enhances the target feature to be segmented through cross-attention operation. As for training, we choose commonly used cross-entropy (CE) loss and Dice loss in this work for model optimization.

Experiment

We conduct experiments on two commonly used ESS datasets, DDD17 (Binas et al. 2017) and DSEC-Semantic (DSEC) (Gehrig et al. 2021), and compare our method with state-of-the-art (SOTA) supervised ESS methods as well as the baseline model SegFormer (Xie et al. 2021). Among them, EV-SegNet (Alonso and Murillo 2019), ESS-Sup (Sun et al. 2022c), EvSegFormer (Jia et al. 2023), HMNet (Smith, Doe, and Lee 2023), and SpikingEDN (Zhang et al. 2024) perform ESS training only with events, consistent with our approach. Unlike these methods, E2VID (Rebecq et al. 2019), EvDistill (Wang et al. 2021), Vid2E (Gehrig et al. 2020) and ESS (Sun et al. 2022c) use knowledge transfer techniques from RGB to event domains. Additionally, to demonstrate the effectiveness of our KWYAF method, we also compare it with advanced VSS methods such as CFFM (Sun et al. 2022a), MRCFA (Sun et al. 2022b), and CFFM++ (Sun et al. 2024). We adopt mean Accuracy (Acc), mean Intersection over Union (mIoU) as evaluation metrics.

Evaluation on the DDD17 Dataset

Dataset and Training Details. DDD17 (Binas et al. 2017) is the first event-based semantic segmentation dataset captured by a DAVIS sensor. Due to the relatively low resolution (346×260), multiple categories are merged and provide labels with six classes. We use the lightweight SegFormer backbone MiT-B0 as the network encoder. As for input, we build each voxel gird using events within $50ms$ interval following (Alonso and Murillo 2019; Smith, Doe, and Lee 2023; Zhang et al. 2024). During training, we employ data augmentation techniques such as random resizing and flipping, and train for 60 epochs with the batch size of 32. We utilize the AdamW and "poly" learning rate schedule, with an initial learning rate of 1e-3. All experiments are implemented using Pytorch on two RTX 3090s.

Method	Venue	Backbone	Acc	mIoU
VSS methods				
CFFM	CVPR'22	MiT-B0	89.39	56.67
MRCFA	ECCV'22	MiT-B0	89.46	56.32
CFFM++	TPAMI'24	MiT-B0	89.72	57.03
ESS methods				
Ev-SegNet	CVPRW'19	Xception	89.76	54.81
E2VID	TPAMI'19	U-Net	85.84	48.47
Vid2E	CVPR'20	Xception	<u>90.19</u>	56.01
EvDistill	CVPR'21	ResNet34	-	**58.02**
ESS	ECCV'22	E2VID	88.43	53.09
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	NIPS'20	MiT-B0	89.34	52.76
Ours		MiT-B0	90.04	<u>57.69</u>
ESS-Sup*	ECCV'22	E2VID	91.08	61.37
Ours*		MiT-B0	**91.32**	**62.41**

Table 1: Results on the DDD17 dataset. All scores are given in percentage (%). *: Constructing Voxel grids $w.r.t$ event numbers following (Sun et al. 2022c).

Results. The quantitative results are shown in Table 1. Compared to the baseline model and other SOTA event-only trained ESS methods, our KWYAF framework demonstrates a notable improvement (1.02% mIoU gain over ESS-Sup). This result underscores the efficacy of incorporating spatio-temporal motion information, effectively mitigating the limitations of event data, which is noisy, sparse and lacks color awareness. As demonstrated in Figure 4, our method is capable of accurately identifying moving humans and poles, as well as precisely segmenting challenging boundaries between humans and vehicles.

Furthermore, our framework outperforms the SOTA VSS method, CFFM++, by 0.66% mIoU. We attribute this improvement to our proposed ER^{2}SM, which addresses the sparse and noisy nature of event signals. Through the combined effects of LWM and RRS, ER^{2}SM filters out noise and extracts high-confidence motion information. Compared to methods based on RGB transfer learning, our approach achieves competitive results. While we slightly trail behind the EvDistill method, it is noteworthy that EvDistill's model parameters (58.64M) are several times larger than ours (4.84M), which highlights the practical value of our

Method	Venue	Backbone	Acc	mIoU
VSS methods				
CFFM	CVPR'22	MiT-B0	90.51	55.97
MRCFA	ECCV'22	MiT-B0	90.67	55.74
CFFM++	TPAMI'24	MiT-B0	90.74	56.75
ESS methods				
Ev-SegNet	CVPRW'19	Xception	88.61	51.76
E2VID	TPAMI'19	U-Net	80.06	44.08
ESS	ECCV'22	E2VID	84.17	45.38
ESS-Sup	ECCV'22	E2VID	89.08	52.30
HMNet	CVPR'23	HMNet-L1	89.80	55.00
EvSegFormer	TIP'23	MiT-B1	<u>90.38</u>	<u>56.04</u>
SpikingEDN	TNNLS'24	SNN	-	53.17
SegFormer	NIPS'20	MiT-B0	90.12	54.19
Ours		MiT-B0	**90.87**	**57.75**
ESS-Sup*	ECCV'22	E2VID	89.37	53.29
Ours*		MiT-B0	**90.47**	**57.45**

Table 2: Results on the DSEC dataset.

method. Finally, in the ESS-Sup* study, the use of event counts to construct voxel grids effectively mitigates the issue of uneven event distribution in the DDD17 dataset. Therefore, we conduct experiments using the same input method, and the results were consistent with our previous findings. With balanced event distribution, our method better performance over ESS-Sup* by 1.04% mIoU.

Evaluation on the DSEC Dataset

Dataset and Training Details. The DSEC dataset (Sun et al. 2022c) includes 53 driving sequences captured by event cameras with a resolution of 640×480. Compared to the DDD17, DSEC has more refined labels with 11 classes. This paper splits the training and testing sets following (Sun et al. 2022c). The training process is similar to DDD17, yet with an initial learning rate of 6e-5.

Results. The conclusions drawn from the DSEC align with those from the DDD17 dataset. Notably, our method consistently achieves superior performance across all compared approaches in Table 2, including both ESS methods relying solely on events and those utilizing RGB knowledge during training. We attribute this outcome to the broader range of scenes and category labels present in the DSEC dataset. This diversity enables our method, enhanced by historical spatio-temporal awareness, to excel in more challenging scenarios. Furthermore, the higher resolution of the DSEC enables our proposed ER^{2}SM module to maximize its potential in high-confidence feature selection. As demonstrated in Figure 5, our method achieves clear identification of small moving objects like humans and precise delineation of roads and pole boundaries.

Variants	SME	ER^{2}SM	Acc	mIoU
A	×	×	90.12	54.19
B	✓	×	90.63	56.93
C (Ours)	✓	✓	**90.87**	**57.75**

Table 3: Ablation study results showing the impact of SME and ER^{2}SM.

Variants	Acc	mIoU
A	90.49	56.38
B	90.67	57.46
C (Ours)	**90.87**	**57.75**

Table 4: Efficacy comparisons of different designs in motion embedding.

Model Analysis

We conduct the a series of experiments on the DSEC dataset to validate the effectiveness of our designs.

Impacts of SME and ER^{2}SM. We first study the contributions of the SME and ER^{2}SM modules in our framework. As shown in Table 3, with only SMEs, our method achieves mIoU of 56.93%, which surpasses the baseline by 2.74%, demonstrating the efficacy of SMEs in supplementing motion propagation cues for the segmentation scene, and thereby assisting the model's decision making. After applying ER^{2}SM, the method further gains improvement in mIoU with 0.82%. As shown in Figure 6, after employing the ER^{2}SM (C), small moving objects can be segmented with more details than setting B does. We attribute this enhancement to ER^{2}SM's ability in selecting high-confidence regions through the combined effects of LWM and RRS, ensuring accurate and reliable motion estimation during the SME aggregation process.

Designs for Motion Embedding. Here, we aim to evaluate the efficacy and rationality of the proposed SME (setting C presented in Table 4 and Figure 7) with other two heuristic approaches A (Sun et al. 2022b) and B (Su et al. 2023). Specifically, settings A&B both optimize the features of the previous (A) or the next segment (B) only by using the motion correlation of two adjacent event segments. Unlike them, our method introduces P to encode all prior motion propagation and injects itself to the target event segment, so that the segmentation scene can obtain multi segment and continuous motion association clues.

The quantitative results in Table 4 show that our approach achieves better performance on this task. We attribute this to that the separate processing of motion features (P_{i+1}) and features used for segmentation prediction ($\hat{F}_{i+1}$) can enable the network to better disentangle the prior motion propagation and adaptively apply it to downstream tasks. The visual comparisons in Figure 7 also verify our point, in which our approach achieves more accurate prediction and locating no matter the dynamic part (walking person) and static regions (traffic sign).

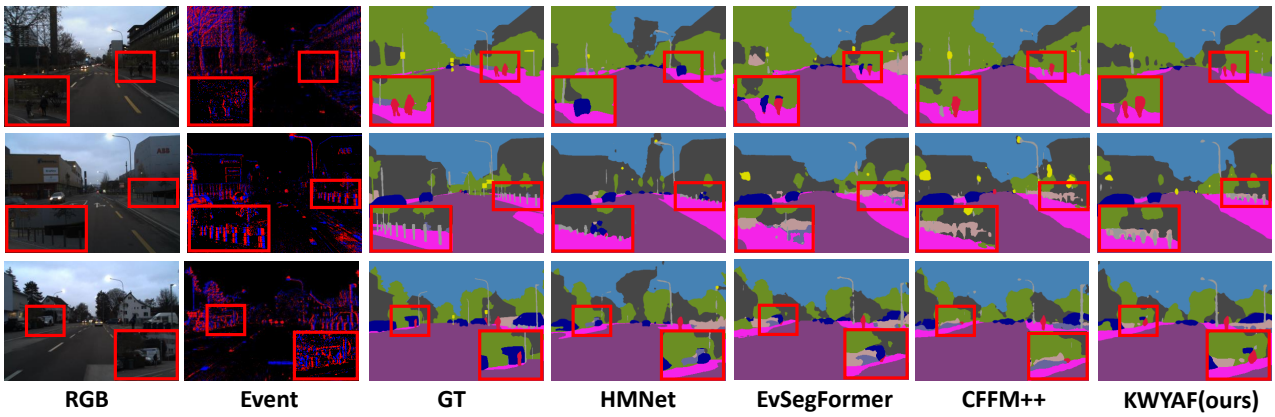

Figure 5: Qualitative results on the DSEC dataset. We visually compare our proposed method with leading ESS and VSS methods. The highlighted regions for comparison are within bounding boxes.

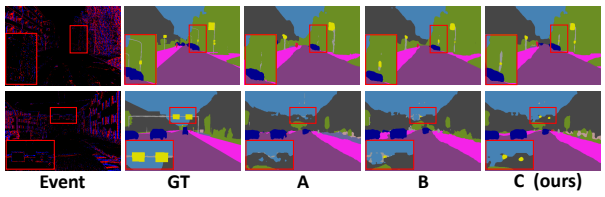

Figure 6: Impact of our SME and ER^{2}SM modules. A represents the baseline without these modules, B with SME only, and C with both modules applied.

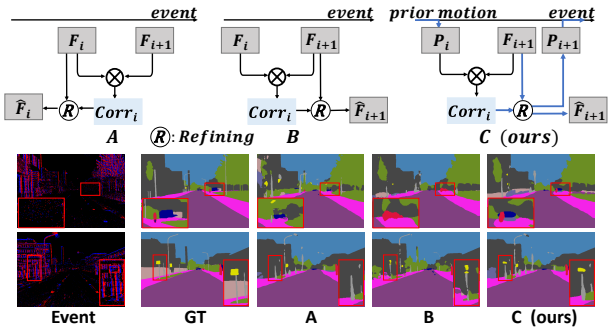

Figure 7: The schematic diagrams and visual results $w.r.t$ different designs of motion embedding.

Efficacy of LWM and RRS. This section explores the efficacy of two main component LWM and RRS inside the ER^{2}SM, the results are shown in Table 5. As can be seen, adding the Local Window Masking (LWM) resulted in a 0.4% mIoU improvement, indicating that the inclusion of LWM made the calculation of the correlation volume between event features more precise, helping the network avoid the impact of distant event noise and irrelevant features. Additionally, when the LWM window size is too small (as seen in settings B and C), the Reliable Region Selection (RRS) mechanism appears to fail. We speculate that the small window size, after undergoing double filtering, fails to capture sufficient regional semantic cues. As the window size increases, the benefits of RRS become apparent, as it can accurately select high-confidence region features within larger windows, ensuring the reliability of motion information aggregation. Based on the results, we select a window size of $m = 7$ as our choice experimentally.

Variants	LWM	RRS	Size	Acc	mIoU
A	×	×	-	90.63	56.93
B	✓	×	3	90.78	57.33
C	✓	✓	3	90.60	57.16
D	✓	×	7	90.75	57.40
E	✓	✓	7	**90.87**	**57.75**
F	✓	×	11	90.52	57.29
G	✓	✓	11	90.73	57.39

Table 5: Ablation study results showing the impact of different settings of the LWM and RRS modules in ER^{2}SM.

Conclusion

There are common issues identified in the event-based segmentation task (ESS), *i.e.,* it is challenging to distinguish different object with similar contour appearances especially when they are spatially continuous. This paper proposes a novel ESS framework named Know Where You Are From (KWYAF) to address the above problem. Based on the assumption that objects within a scene normally hold various motion modes or states, our framework equips a series of SMEs to model the spatio-temporal propagation and represent these priors with spatial semantics jointly for final prediction. Meanwhile, acknowledging the noisy and motion-edge-triggered nature of event data, we introduced the ER^{2}SM to enhance the robustness of motion encoding to event data with different distribution, in which local window masking and reliable region selction operations are imposed together to locate reliable regions during motion encoding. Extensive experiments on DDD17 and DSEC datasets validate the effectiveness of our framework and designs.

Acknowledgments

This work is jointly supported by National Key Research and Development Program of China (2022YFF0610000), 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.

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

Cheng, B.; Misra, I.; Schwing, A. G.; et al. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Das Biswas, S.; Kosta, A.; Liyanagedera, C.; et al. 2024. HALSIE: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5964–5974.

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.

Ding, M.; Wang, Z.; Zhou, B.; Shi, J.; Lu, Z.; and Luo, P. 2020. Every frame counts: Joint learning of video segmentation and optical flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10713–10720.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 154–180.

Gehrig, D.; Loquercio, A.; Derpanis, K. G.; 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.; et al. 2021. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954.

Hu, P.; Caba, F.; Wang, O.; et al. 2020. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jain, J.; Singh, A.; Orlov, N.; Huang, Z.; Li, J.; Walton, S.; and Shi, H. 2023. SeMask: Semantically Masked Transformers for Semantic Segmentation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 752–761.

Jia, Z.; You, K.; He, W.; et al. 2023. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32: 1829–1842.

Jiang, J.; Zhou, X.; Duan, P.; and Shi, B. 2023. EvPlug: Learn a Plug-and-Play Module for Event and Image Fusion. *arXiv preprint arXiv:2312.16933*.

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.

Kim, Y.; Chough, J.; and Panda, P. 2022. Beyond classification: Directly training spiking neural networks for semantic segmentation. *Neuromorphic Computing and Engineering*, 2(4): 044015.

Klenk, S.; Bonello, D.; Koestler, L.; Araslanov, N.; and Cremers, D. 2024. Masked event modeling: Self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2378–2388.

Kong, L.; Liu, Y.; Ng, L. X.; Cottereau, B. R.; and Ooi, W. T. 2024. OpenESS: Event-based Semantic Scene Understanding with Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15686–15698.

Lan, M.; Wang, X.; Ke, Y.; Xu, J.; Feng, L.; and Zhang, W. 2023. SmooSeg: Smoothness Prior for Unsupervised Semantic Segmentation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 11353–11373. Curran Associates, Inc.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2006. A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, 2060–2069. IEEE.

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.

Liu, Y.; Shen, C.; Yu, C.; and Wang, J. 2020. Efficient semantic video segmentation with per-frame inference. In *Proceedings of the European Conference on Computer Vision*.

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.

Niu, D.; Wang, X.; Han, X.; Lian, L.; Herzig, R.; and Darrell, T. 2024. Unsupervised Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22744–22754.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. Events-to-Video: Bringing Modern Computer Vision to Event Cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3857–3866.

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.

Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; and Yang, W. 2024. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913.

Smith, J.; Doe, J.; and Lee, C. 2023. Hierarchical Neural Memory Network for Low Latency Event Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1234–1243.

Su, J.; Yin, R.; Zhang, S.; and Luo, J. 2023. Motion-state Alignment for Video Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3571–3580.

Suh, Y.; Choi, S.; Ito, M.; Kim, J.; Lee, Y.; Seo, J.; Jung, H.; Yeo, D.-H.; Namgung, S.; Bong, J.; et al. 2020. A 1280$\times$ 960 dynamic vision sensor with a 4.95-μm pixel pitch and motion artifact minimization. In *2020 IEEE international symposium on circuits and systems (ISCAS)*, 1–5. IEEE.

Sun, G.; Liu, Y.; Ding, H.; Wu, M.; and Van Gool, L. 2024. Learning local and global temporal contexts for video semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, G.; Liu, Y.; Ding, H.; et al. 2022a. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sun, G.; Liu, Y.; Tang, H.; et al. 2022b. Mining relations among cross-frame affinities for video semantic segmentation. In *Proceedings of the European Conference on Computer Vision*.

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

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, 402–419. Springer.

Wang, H.; Wang, W.; and Liu, J. 2021. Temporal memory attention for video semantic segmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2254–2258. IEEE.

Wang, L.; Chae, Y.; Yoon, S. H.; et al. 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, 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, G.; Jia, W.; Wu, T.; Chen, L.; and Gao, G. 2024. HAFormer: Unleashing the Power of Hierarchy-Aware Features for Lightweight Semantic Segmentation. *IEEE Transactions on Image Processing*, 33: 4202–4214.

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.

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, Y.; Yang, Y.; Xiong, C.; Sun, G.; and Guo, Y. 2022. Attention-based dual supervised decoder for RGBD semantic segmentation. *arXiv preprint arXiv:2201.01427*.