Intra-Class Consistency Guided Class-Agnostic Event Segmentation

Zhipeng Sui, Haiqing Hao, Weihua He, and Wenhui Wang*

State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing, China

Abstract. Class-agnostic segmentation enables open-world object segmentation beyond predefined categories, and extending it to the event domain helps enhance perception in high-speed motion and extreme illumination scenarios. However, the sparsity and low redundancy of event data lead to weak intra-class consistency, making event-based segmentation typically underperform compared to RGB-based counterparts. To address this issue, we propose IC^2-ESeg, an Intra-class Consistency guided, Class-agnostic Event Segmentation framework, including three modules: intra-class consistency mining (ICM), lightweight information injection (LIJ), and cross-modal distillation (CMD). Grounded in the event generation mechanism, the ICM module constructs noise, motion, and brightness representations by analyzing statistical noise frequency, spatio-temporal distribution, and brightness gradients, thereby enriching the informational content of event data. The LIJ module efficiently integrates these representations into the image encoder, while the CMD module transfers the general segmentation capability of Segment Anything Model (SAM) to the event domain. Experiments on RGBE-SEG demonstrate that IC^2-ESeg achieves state-of-the-art (SOTA) performance in class-agnostic segmentation. We also construct ComScene, a high resolution RGB-event segmentation dataset encompassing diverse scenarios. Results on MVSEC and ComScene demonstrate the strong generalization of IC^2-ESeg.

Keywords: Event cameras · Class-agnostic segmentation · Intra-class consistency

1 Introduction

Class-agnostic segmentation [18, 27, 41, 42] aims to segment objects based on general "objectness" rather than predefined semantic categories. Distinct from closed-set semantic segmentation, class-agnostic approaches empower systems to perceive and localize arbitrary objects in diverse scenarios. Consequently, this task has emerged as a fundamental component of open-world perception [26, 27], finding widespread applications in data annotation [29, 43] and image editing [11, 40]. Recently, the advent of vision foundation models has revolutionized this

* Corresponding author: wwh@tsinghua.edu.cn

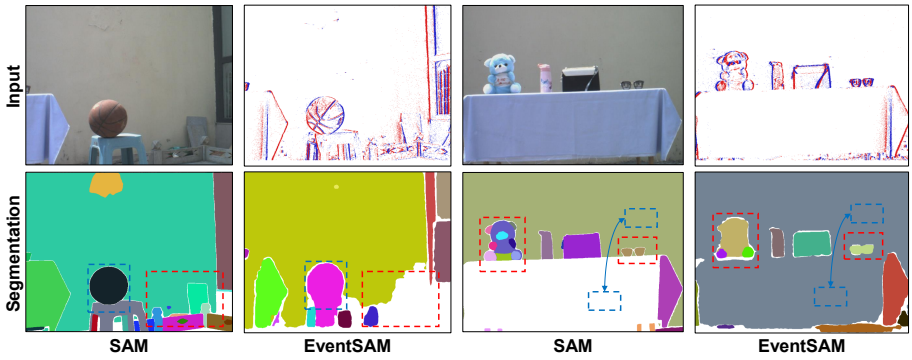

Fig. 1: Comparison of segmentation results between RGB images and event data: RGB images exhibit strong intra-class consistency, whereas event data, due to weaker intra-class consistency, show noticeably inferior segmentation performance compared to RGB images. Segmentation degradation manifests as foreground-background confusion or object adhesion (blue boxes), and the loss of fine-grained details (red boxes).

field. In particular, the Segment Anything Model (SAM) [18] has demonstrated unprecedented zero-shot generalization capabilities. By training on the massive SA-1B dataset, SAM effectively captures universal structural priors and geometric information, establishing a new benchmark for class-agnostic segmentation.

Event cameras [8,38] are characterized by high dynamic range, high temporal resolution, and low latency [19], making event-based class-agnostic segmentation promising for enhancing perception in high-speed motion and extreme illumination scenarios, yet its realization remains challenged by two key issues. First, due to the lack of large-scale real-world event-based datasets comparable to ImageNet [9] or SA-1B [18], training event-based models from scratch to learn generalized object concepts is difficult. Second, event data are inherently asynchronous and spatially sparse, exhibiting substantial differences from spatially dense RGB images. This modality gap hinders the direct transfer of RGB-based class-agnostic segmentation methods to the event domain.

To reduce dependence on large-scale real-world event-based datasets, several studies have explored introducing pretrained models such as SAM into the event domain to enable cross-modal knowledge transfer [7, 12, 39, 44]. However, most of them focus on semantic segmentation rather than class-agnostic segmentation [12,39,44]. EventSAM presents the first event-based class-agnostic segmentation method [7]. However, this approach concentrates solely on feature-level similarity while neglecting modality discrepancies at the input level, thereby limiting segmentation performance.

Our core insight is that the spatial sparsity and low redundancy of event data result in weak intra-class consistency, thereby limiting the performance of event-based class-agnostic segmentation. Unlike dense RGB images, where intra-class consistency is explicitly manifested through similar colors and textures, event data are primarily represented as sparse edges (as shown in Fig. 1), which

increases the difficulty of representation learning. Since class-agnostic segmentation is inherently a dense prediction task, weak intra-class consistency poses a substantial challenge. To address this issue, we exploit the event generation mechanism to mine and model intra-class consistency in event data.

In this paper, we propose IC^2-ESeg, an Intra-class Consistency guided, Class-agnostic Event Segmentation framework. IC^2-ESeg consists of three modules: intra-class consistency mining (ICM), lightweight information injection (LIJ), and cross-modal distillation (CMD). Our observations indicate that the frequency of noise events encodes scene intensity information, the spatio-temporal distribution of events encodes motion information, and event density reflects the coupling between brightness gradients and the motion field. Based on these insights, the ICM module constructs noise, motion, and brightness representations in a learning-free manner to recover scene information. The LIJ module efficiently integrates these representations into the image encoder through a combination of residual connections and visual prompts. The CMD module follows a teacher-student paradigm and transfers the class-agnostic segmentation capability of SAM from the image modality to the event modality through feature-level distillation and pseudo-label supervision.

The experimental results demonstrate that our proposed IC^2-ESeg achieves significant improvements in segmentation performance. With only a 0.08% increase in model parameters, IC^2-ESeg improves mIoU/aIoU by 4.9%/3.6% in automatic mode and by 32.7%/16.4% in prompt mode compared with the baseline. To evaluate zero-shot transfer capability, we construct a new dataset, ComScene, captured using a higher-resolution event camera distinct from the training setup and covering diverse unseen scenarios. Experimental results on MVSEC [45] and ComScene demonstrate strong generalization ability. Our contributions are summarized as follows:

- We propose an intra-class consistency mining (ICM) module to recover scene information in a learning-free manner.
- Building on the ICM module, we propose IC^2-ESeg framework to achieve class-agnostic event segmentation.
- We present ComScene, a high resolution RGB-event segmentation dataset encompassing diverse scenarios.
- Experiments on RGBE-SEG demonstrate that IC^2-ESeg achieves state-of-the-art (SOTA) performance, and results on MVSEC and ComScene further validate its strong generalization ability.

2 Related Work

2.1 Class-Agnostic Image Segmentation

Class-agnostic image segmentation aims to partition all visual entities, including both objects and background regions, without predicting specific semantic labels. Early research often coupled class-agnostic segmentation with few-shot scenarios; notably, CANet utilizes a dense comparison module and iterative optimization to

segment novel classes based on limited support examples [41]. Moving towards open-world perception, the Entity Segmentation task was formally defined to treat "things" and "stuff" uniformly, demonstrating that a fully convolutional framework equipped with a global kernel bank can effectively segment entities without class labels [27]. A paradigm shift occurred with the introduction of the Segment Anything Model (SAM), which leverages massive-scale pre-training and promptable queries to achieve unprecedented zero-shot generalization across diverse domains [18]. To improve performance on out-of-distribution (OOD) and zero-shot tasks, the CSL framework was proposed to mitigate class bias in existing models through mask splitting preprocessing and a soft assignment mechanism during inference [42]. Most recently, a novel bottom-up approach was developed that supervises features on a projective sphere, enabling the segmentation of arbitrary entities via simple mean-shift clustering without relying on top-down proposals [11]. Although notable progress has been made in the image domain, event-based class-agnostic segmentation remains largely underexplored, which is the primary focus of our work.

2.2 Event-Based Segmentation

EV-SegNet utilized an Xception-based encoder-decoder network to perform semantic segmentation directly from event data [1]. EvDistill addressed the challenge of lacking labeled event data by distilling knowledge from unpaired labeled images to events and employed a bidirectional modality reconstruction module to bridge the domain gap [35]. Dual Transfer Learning (DTL) introduced a framework that shares an encoder between the primary segmentation task and an auxiliary event-to-image translation branch, transferring both feature-level and prediction-level knowledge to enhance feature learning [34]. ESS proposed an unsupervised domain adaptation framework that utilizes the recurrent E2VID encoder to generate motion-invariant event embeddings, enabling task transfer from labeled images to unlabeled events [31].

Given the strong segmentation capability of SAM, several studies have adapted SAM for event-based segmentation. SAM-Event-Adapter adapted a frozen SAM for event-RGB semantic segmentation by inserting lightweight adapters into the architecture [39]. EventSAM employed knowledge distillation to transfer knowledge from SAM to an event-based student network and adopted a weighted adaptation strategy that calibrates pivotal token embeddings between modalities to manage high-level embedding discrepancies [7]. ESEG utilized SAM to generate explicit semantic edge labels offline, which serve as supervision to guide the network in focusing on reliable edge regions via a density-aware dynamic-window cross-attention fusion module [44]. EvSAM introduced a novel RGB-event semantic segmentation framework that incorporates an event feature injector and a multi-scale spatiotemporal prompt attention block to achieve event-RGB complementarity and efficient spatiotemporal encoding of event data, respectively [12]. Most of these methods primarily focus on event-based semantic segmentation, while class-agnostic segmentation remains largely underexplored. EventSAM explored class-agnostic event segmentation, but it only considered

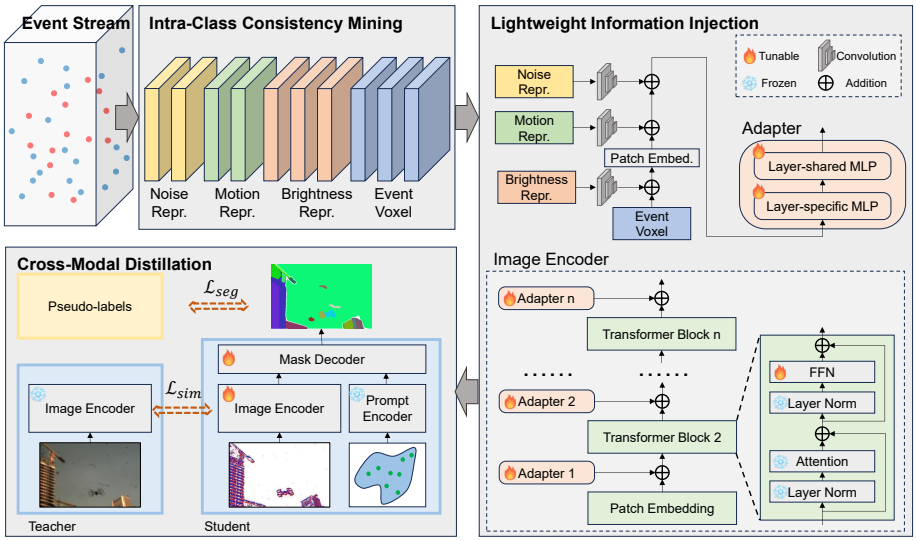

Fig. 2: IC^2-ESeg framework. (i) The intra-class consistency mining module constructs noise, motion, and brightness representations from event data, (ii) the lightweight information injection module integrates these representations into the image encoder, and (iii) the cross-modal distillation module transfers the segmentation capability of the SAM model from the image domain to the event domain.

feature-level similarity, resulting in suboptimal performance. In contrast, our work emphasizes mining intra-class consistency from event data at the input level and efficiently injecting such information into the image encoder.

3 Method

3.1 IC^2-ESeg Overview

This study aims to achieve event-based class-agnostic segmentation. Specifically, given an event sequence, the goal is to predict a dense segmentation map in which pixels are assigned to different object regions without involving explicit semantic class labels. To this end, we propose the IC^2-ESeg framework, as illustrated in Fig. 2, which consists of three modules: intra-class consistency mining (ICM), lightweight information injection (LIJ), and cross-modal distillation (CMD).

3.2 Intra-Class Consistency Mining

This section describes the construction of noise, motion, and brightness representations based on the physical mechanism of event generation. First, the event stream is divided into noise events and valid events. Noise events encode information about scene intensity, and a noise representation is constructed through

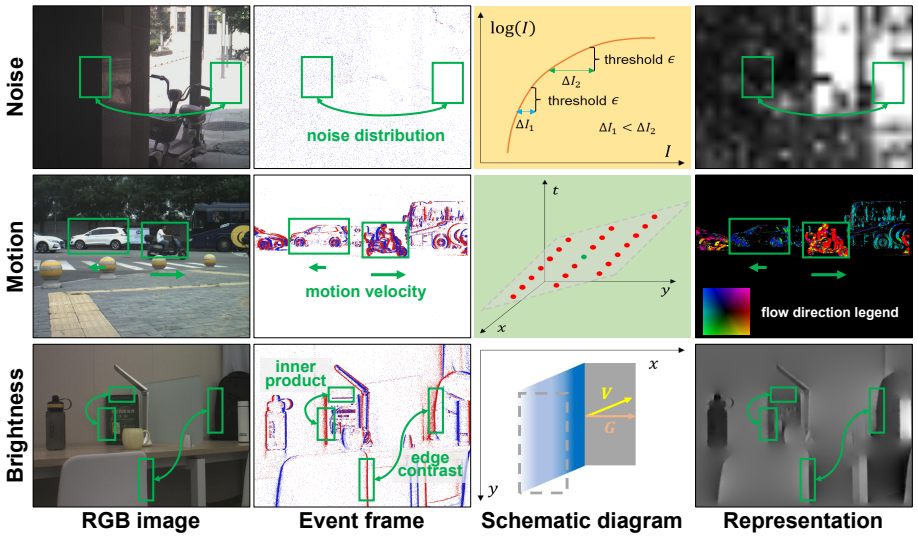

Fig. 3: Intra-class consistency mining based on the physical mechanism of event generation. Observations indicate that intra-class consistency in event data is primarily manifested in noise, motion, and brightness information, motivating us to construct corresponding representations to complement the event data.

statistical analysis. Valid events are generated by the combined effect of brightness gradients and the motion field in the scene. An improved local plane fitting method is employed to estimate the motion representation, resulting in a sparse motion map. The motion representation is further used to estimate brightness gradients, which are then used to construct the brightness representation via Poisson reconstruction, ultimately yielding a dense brightness map.

Noise Representation All vision sensors are affected by the shot noise of photons, and this phenomenon is even more pronounced in event cameras [13]. We regard isolated events in the spatio-temporal domain as noise, as they neither contribute to the formation of scene contours nor exhibit temporal continuity.

By searching the local neighborhood of each event, we can determine whether it belongs to isolated noise. Each event in the event stream $\mathcal{E}$ is encoded as $(\mathbf{u}, t, p)$, where $\mathbf{u}$ denotes the pixel coordinates (x, y), t is the timestamp, and $p \in \{-1, +1\}$ indicates the polarity (decrease or increase in brightness). For event e_k, we consider its local neighborhood:

$$\mathcal{N}(e_k) = \{e_j \in \mathcal{E} | \ \|\mathbf{u}_j - \mathbf{u}_k\|_\infty \leq s, |t_j - t_k| \leq \Delta t\} \ , \tag{1}$$

where s and Δt represent the spatial and temporal search radii, respectively. If the number of events $|\mathcal{N}(e_k)|$ in its neighborhood is less than the threshold $N_{\min}$, we consider this event to be noise. To enhance search efficiency, we index events using a KD-Tree [3]. More details can be found in Supplementary Material S1.1.

Fig. 3 illustrates that the number of noise events decreases as the scene intensity increases. Intuitively, this is because regions with high intensity require larger intensity fluctuations to reach the temporal contrast threshold ϵ. Since the triggering probability of noise events is correlated with scene intensity, it is feasible to estimate the scene intensity based on the noise frequency.

Unlike [5], which relies on long-term event accumulation and a U-Net architecture to estimate scene intensity, we adopt a learning-free method to avoid additional latency and computational overhead. Due to the randomness in the photon arrival process (shot noise), the photon counts at different time steps, denoted as n_t and n_{t-1}, can be regarded as independent and identically distributed random variables. These quantities can be approximated by a Gaussian distribution $\mathcal{N}(\lambda, \lambda)$, where the average photon count λ is proportional to the intensity I [5]. Through theoretical derivation, the probability of triggering a positive event can be expressed as

$$\Pr(n_t - n_{t-1} \cdot e^{\epsilon} > 0) = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\sqrt{\lambda} \cdot (e^{\epsilon} - 1)}{\sqrt{2 \cdot (1 + e^{2\epsilon})}}\right), \tag{2}$$

where $\mathrm{erf}(\cdot)$ denotes the error function. Eq. (2) indicates that as λ increases, the probability of triggering a positive event decreases. Eq. (2) provides a quantitative formulation for estimating the intensity based on the frequency of noise events. We estimate the intensity for the two polarities separately and take their average to obtain the noise representation. More details can be found in Supplementary Material S1.1.

Motion Representation We adopt the method proposed in [2] to construct the motion representation. [2] introduces the Surface of Active Events (SAE) and its quantitative relationship with motion velocity, which can be formulated as

$$\begin{cases} t = \Sigma_e(x, y) \\ \nabla \Sigma_e = \left(\frac{1}{v_x}, \frac{1}{v_y}\right)^T \end{cases}, \tag{3}$$

where $\Sigma_e(x, y)$ denotes the SAE, which is a two-dimensional surface ($\mathbb{R}^2 \rightarrow \mathbb{R}$), and t denotes the timestamp of the last event at position (x, y). (v_x, v_y) represents the velocity and is equal to the inverse of the gradient of the SAE. As shown in Fig. 3, the velocity of each event can be calculated by local plane fitting all events within its spatio-temporal neighborhood Ω_e.

The computation involves selecting the spatio-temporal neighborhood Ω_e, defined by an $L \times L$ spatial window and a temporal window $[t - \Delta t, t + \Delta t]$. We observe that the accuracy of velocity computation is significantly influenced by these parameters. Drawing inspiration from the pyramid multi-scale approach in computer vision [20], we use three different spatio-temporal scales of Ω_e to jointly construct the motion representation, thereby enhancing its robustness. More details can be found in Supplementary Material S1.2.

Brightness Representation We propose an efficient brightness gradient estimation method and construct the brightness representation by Poisson reconstruction of brightness gradients. Scene brightness reconstruction is inherently an ill-posed problem. Existing methods typically rely on recurrent neural network architectures [4,28,30], which often require long time span event sequences as continuous inputs, and undergo several frames of initialization to achieve satisfactory reconstruction results, while also being prone to introducing reconstruction artifacts.

As illustrated in Fig. 3, the event density is determined by the edge contrast and the angle between the edge normal direction and the motion direction. The first-order Taylor expansion of the brightness constancy assumption [32] is given by

$$\frac{\partial L(\mathbf{u},t)}{\partial t} \approx -\nabla L(\mathbf{u},t)^T \cdot \mathbf{v}(\mathbf{u},t), \tag{4}$$

where $\frac{\partial L}{\partial t}$ is the rate of brightness change (event density), ∇L is the brightness gradient (edge contrast), and $\mathbf{v}$ is the motion velocity. Eq. (4) can be further simplified as $E = -\mathbf{G} \cdot \mathbf{v}$. We use Event Count [22] to characterize the event density E and compute the brightness gradient $\mathbf{G}$ based on the constructed motion representation $\mathbf{v}$. We sum the brightness gradient maps computed from the two polarities and further construct the dense brightness representation via Poisson reconstruction [17]. The above procedure requires only a single event segment to produce satisfactory results, thereby eliminating the need for continuous inputs and avoiding initialization issues. More details can be found in Supplementary Material S1.3.

3.3 Lightweight Information Injection

We design an efficient information injection strategy to incorporate noise, motion, and brightness representations into the image encoder. Since these representations encode different physical properties, simple channel concatenation would increase the difficulty of representation learning. Instead, we adopt a staged information injection strategy. Specifically, for the brightness representation, which is more closely related to images, we introduce it in a residual manner. The noise and motion representations are converted into visual prompts and injected into the image encoder.

We adopt a residual paradigm to inject the brightness representation x_b. We choose the commonly used Voxel Grid [46] as the initial event representation, and the duration spanned by the events is discretized into 3 temporal bins. The residual paradigm can be formalized as

$$y = \mathcal{F}(x_b; \Theta) + x_v, \tag{5}$$

where $\mathcal{F}(\cdot)$ and Θ denote the mapping function and its parameters, respectively. Specifically, $\mathcal{F}(\cdot)$ consists of two convolutional layers that implement a channel expansion-compression process ($3 \rightarrow 16 \rightarrow 3$). The residual is additively fused

with x_v, thereby supplementing the spatial brightness information of the event data.

Inspired by SAM-Adapter [6], we incorporate noise and motion representations as visual prompts into each layer of the image encoder in SAM. We first reduce the input embedding dimension through a convolutional layer to ensure low computational cost, obtaining the patch embedding $F_{pe} \in \mathbb{R}^{N \times L \times C/s}$, where N, L, and C denote the batch size, sequence length, and embedding dimension, respectively, and s denotes the scaling factor. The motion representation x_m and the noise representation x_n are processed into patch embeddings of the same dimension, denoted as F_m and F_n, respectively. These embeddings are then summed to obtain the fused feature, i.e., $F_i = F_{pe} + F_m + F_n$. The fused features are further processed by a lightweight adapter consisting of two multilayer perceptrons (MLPs) with a nonlinear activation. The process can be formulated as

$$P_i = \mathrm{MLP}_{up}(\mathrm{GELU}(\mathrm{MLP}^i_{tune}(F_i))), \tag{6}$$

where MLP^i_{tune} extracts guiding information from the fused features and enhances nonlinear expressiveness through the GELU activation function [15]. MLP_{up} is shared across layers and maps the processed features back to the original embedding dimension C. The resulting prompt P_i is injected into each layer of the image encoder.

3.4 Cross-Modal Distillation

We adopt a teacher-student paradigm to enable cross-modal knowledge distillation from images to events, as shown in Fig. 2. To endow the model with strong segmentation capability and generalization, we leverage the SAM model with pretrained weights. SAM consists of an image encoder, a prompt encoder, and a mask decoder [18]. During inference, SAM supports a prompt mode, enabling instance-level segmentation via point, box, or mask prompts, as well as an automatic mode that performs dense segmentation over the entire image. Considering training efficiency and computational cost, we select the ViT-B variant with 91M parameters. The teacher and student networks are used to process image and event data, respectively. The image encoder of the teacher network is frozen, while the patch embedding layer in the student network is trainable to ensure adaptability to event-based inputs. To preserve the generalization capability of SAM and prevent overfitting, we fine-tune only the Feed-Forward Networks (FFNs) within each Transformer block of the student network.

The teacher network consists solely of the image encoder, while the student network adopts the full SAM architecture. The prompt encoder in the student network remains frozen, and the mask decoder is frozen during the early training stage and subsequently updated with a smaller learning rate in the later stage. Pseudo-labels are generated offline using the SAM ViT-H variant with 636M parameters on images paired with the event data and are used to supervise the predictions of the student network. The overall objective function used in the

training process can be expressed as

$$\mathcal{L} = \mathcal{L}_{sim} + \alpha \mathcal{L}_{seg}, \tag{7}$$

where $\mathcal{L}_{sim}$ denotes the feature similarity loss, $\mathcal{L}_{seg}$ denotes the segmentation loss, and α is the hyper-parameter to balance these two loss functions. We adopt the feature similarity loss proposed in [7], which is a correlation-aware weighted loss focusing on aligning pivotal token embeddings. The segmentation loss employs a linear combination of focal loss [21] and dice loss [23] in a 20:1 ratio of focal loss to dice loss, following SAM [18]. During the training process, we randomly select 8 points from each mask in the pseudo-labels as prompts, prompting the model to output the corresponding prediction mask.

4 Experiments

4.1 Experimental Setup

Implementation Details The model is implemented using the PyTorch framework [25] and trained on two NVIDIA RTX 3090 GPUs, using the Adam optimizer. Initial learning rates (LR) are 2×10^{-4} for the image encoder and 1×10^{-4} for the LIJ module. The mask decoder is frozen for the first two epochs, then updated with an LR of 2×10^{-7}. We set $\alpha = 0.1$ and train for 10 epochs with a batch size of 16. An ExponentialLR scheduler ($\gamma = 0.9$ every 3 epochs) is used to stabilize training.

Datasets The RGBE-SEG dataset [7] is constructed from two event-based object-tracking datasets, VisEvent [36] and COESOT [33]. RGBE-SEG contains 65,957 RGB-event pairs with a resolution of 346×260, of which 64,957 are used for training and 1,000 for testing. The MVSEC dataset [45] is a large-scale event-camera dataset designed for 3D perception with RGB-event spatial resolution of 346 $\times$260. Following the protocol of [7], we sample 500 RGB-event pairs from the indoor_flying1 and outdoor_day1 sequences, respectively, to assess scene generalization performance. To address the limited spatial resolution of existing datasets and to cover a broader range of unseen scenarios, we construct a new RGB-event dual-modality dataset named ComScene. ComScene is collected using a dual-modality data acquisition device, equipped with a Teledyne FLIR Blackfly-S as the frame-based (RGB) camera and a Prophesee EVK-4 as the event camera. ComScene contains 15,481 aligned RGB-event pairs with a resolution of 964×696, covering four representative scene categories: life, office, sports, and traffic. The download link and additional details are provided in Supplementary Material S2.

Evaluation Metrics Since the generated masks are class-agnostic, we adopt the evaluation protocol in [7], which assigns each predicted mask to the ground-truth mask with the highest intersection over union (IoU). Based on these matches, the

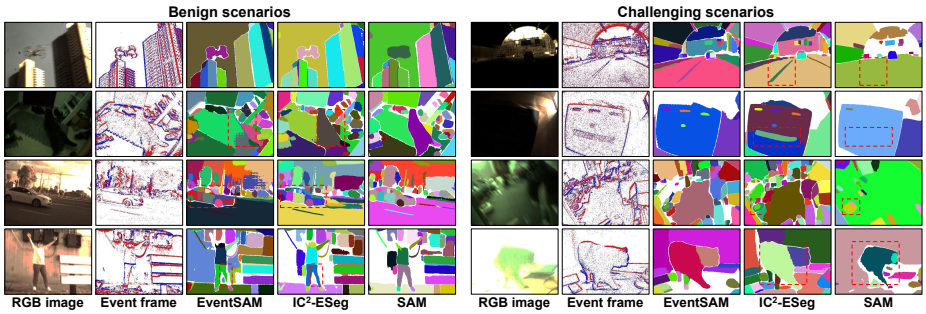

Fig. 4: Segmentation performance on the RGBE-SEG dataset. EventSAM and IC^2-ESeg use event data as input, whereas SAM operates on RGB images. IC^2-ESeg demonstrates superior segmentation performance in both benign and challenging scenarios, with segmentation differences highlighted by red boxes.

mean precision (mP), mean recall (mR), mean intersection over union (mIoU), and area-weighted intersection over union (aIoU) can be further computed. During our experiments, we observe that SAM in automatic mode can produce ambiguous outputs. Specifically, a region may be treated either as an independent segmentation mask or as part of a larger object mask, which significantly affects the objectivity of evaluation metrics. To mitigate the impact of this ambiguity on performance evaluation, we additionally conduct experiments in prompt mode, sampling points from the ground truth as prompts.

Comparing Methods We compare the proposed method with the existing class-agnostic segmentation approach EventSAM [7] in both automatic and prompt modes. Following [7], we also test other event-based segmentation approaches, which can be broadly categorized into three types. The first category involves image reconstruction-based methods, which convert raw events into grayscale frames using E2VID [28] or ETNet [37], followed by inference with a pre-trained SAM. The second category relies on pre-trained E2VID encoders, such as EVDistill [35], DTL [34], and ESS [31], to reduce the domain gap between RGB and event data, and then feeds the resulting features into a segmentation network. The third category employs knowledge distillation methods, such as NGA [16] and LFI [10], which transfer knowledge from RGB backbone networks to event-based data.

4.2 Event Segmentation

Fig. 4 presents a qualitative comparison of segmentation results obtained by different methods in both benign and challenging scenarios. The experimental results indicate that, in benign scenarios, IC^2-ESeg generates more accurate segmentation masks than EventSAM, achieving performance that is visually comparable to that of SAM on RGB images. The challenging scenarios include

Table 1: Performance of different segmentation models on RGBE-SEG dataset. The "*" symbol indicates segmentation results obtained in prompt mode. Compared with mP and mR, mIoU and aIoU are more critical evaluation metrics, as they provide a more comprehensive assessment of model performance. The best performance is highlighted in **bold** and dark green.

Method	Easy				Medium				Hard				All			
	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU
SAM [18]	0.52	0.75	0.39	0.58	0.38	0.71	0.25	0.41	0.26	0.74	0.15	0.29	0.39	0.73	0.26	0.43
E2VID [28]	0.62	0.60	0.38	0.54	0.55	0.58	0.32	0.44	0.46	0.64	0.26	0.36	0.54	0.61	0.32	0.45
ETNet [37]	0.68	0.56	0.40	0.51	0.64	0.52	0.35	0.42	0.60	0.52	0.31	0.35	0.63	0.53	0.35	0.42
NGA [16]	0.56	0.72	0.40	0.58	0.44	0.70	0.29	0.45	0.34	0.73	0.21	0.35	0.45	0.71	0.30	0.45
LFI [10]	0.56	0.71	0.38	0.56	0.42	0.71	0.26	0.44	0.31	0.76	0.18	0.33	0.42	0.72	0.27	0.44
DTL [34]	0.57	0.71	0.40	0.59	0.45	0.70	0.28	0.46	0.32	0.75	0.20	0.35	0.44	0.71	0.29	0.46
EVDistill [35]	0.72	0.46	0.32	0.40	0.64	0.46	0.27	0.34	0.53	0.51	0.23	0.30	0.63	0.47	0.27	0.34
ESS [31]	0.55	0.73	0.40	0.57	0.41	0.73	0.27	0.45	0.29	0.77	0.18	0.32	0.42	0.74	0.28	0.45
EventSAM [7]	0.66	0.73	0.49	0.65	0.61	0.69	0.40	0.54	0.50	0.72	0.34	0.47	0.59	0.71	0.41	0.55
EventSAM [7] *	0.63	**0.85**	0.56	0.73	0.57	**0.83**	0.49	0.60	0.50	**0.81**	0.43	0.53	0.57	**0.83**	0.49	0.61
Ours	0.64	0.78	0.51	0.67	0.57	0.76	0.43	0.57	0.47	0.77	0.36	0.49	0.56	0.77	0.43	0.57
Ours *	**0.84**	0.79	**0.67**	**0.74**	**0.83**	0.77	**0.65**	**0.71**	**0.83**	0.76	**0.63**	**0.69**	**0.83**	0.77	**0.65**	**0.71**

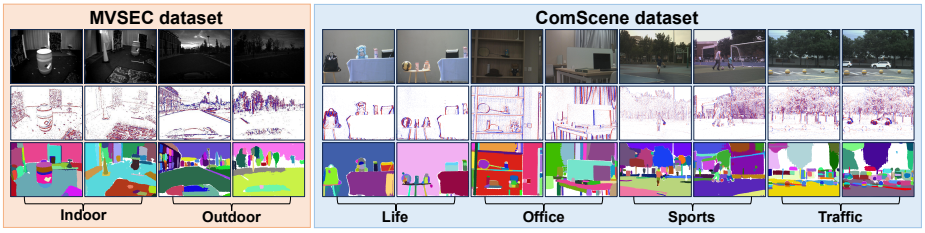

Fig. 5: Generalization performance of IC^2-ESeg on the MVSEC (left) and ComScene (right) datasets. Zoom in for details.

complex conditions such as low/high illumination, and high-speed motion. Due to the severe deterioration of the input RGB images, SAM exhibits a significant degradation in segmentation performance. In contrast, IC^2-ESeg maintains superior segmentation performance over EventSAM in challenging scenarios, demonstrating stronger robustness.

The quantitative experimental results on the RGBE-SEG dataset are provided in Tab. 1. Although it is slightly inferior to several methods in terms of mP and mR, our method attains new SOTA performance on the more critical mIoU and aIoU metrics, outperforming the second-best approach by 4.9% and 3.6%, respectively. We further evaluate the segmentation results under the prompt mode. The results show that in prompt mode, our strategy exhibits a significant lead, improving mIoU and aIoU by 32.7% and 16.4%, respectively, over the second-best result.

Table 2: Generalization across datasets and scenes.

Method	MVSEC dataset				ComScene Dataset							
	Indoor		Outdoor		Life		Office		Sports		Traffic	
	mIoU	aIoU	mIoU	aIoU	mIoU	aIoU	mIoU	aIoU	mIoU	aIoU	mIoU	aIoU
EventSAM (p)	0.45	0.62	0.55	0.63	0.40	0.61	0.49	0.65	0.38	0.51	0.25	0.40
EventSAM (p+b)	0.61	0.70	0.58	0.68	0.57	0.73	0.61	0.75	0.54	0.66	0.44	0.47
Ours (p)	0.64	0.70	0.64	0.72	0.57	0.71	**0.63**	0.74	0.56	0.69	0.49	0.62
Ours (p+b)	**0.69**	**0.75**	**0.69**	**0.77**	**0.58**	**0.74**	**0.63**	**0.77**	**0.57**	**0.71**	**0.52**	**0.64**

4.3 Zero-Shot Transfer Experiments

To evaluate the zero-shot transfer capability of the proposed IC^2-ESeg, we conduct experiments on different datasets encompassing diverse scenes. Qualitative visualizations and quantitative results are presented in Fig. 5 and Tab. 2, respectively. Our method demonstrates strong generalization across datasets and scenes. When using point prompts only (p), our approach achieves an average improvement of 28.0% in mIoU and 13.6% in aIoU on MVSEC, and 48.0% in mIoU and 27.2% in aIoU on ComScene. When both points and boxes are used as prompts (p+b), our method yields average improvements of 16.0% in mIoU and 10.1% in aIoU on MVSEC, and 6.5% in mIoU and 9.6% in aIoU on ComScene. The results further indicate that combining different types of prompts can enhance segmentation performance.

4.4 Computational Overhead and Latency

Compared to the original SAM, the lightweight information injection module introduces additional parameters and computation, stemming from the residual connection (Eq. (5)) and adapter-based visual prompts (Eq. (6)). Experimental results show that these modifications introduce only an additional 76.7K parameters (0.08%) and 32B FLOPs (13.2%) in inference computation, which is considered acceptable.

The intra-class consistency mining module introduces a certain degree of latency. By leveraging parallel acceleration and input downsampling, the computational efficiency is substantially improved without compromising accuracy. Statistical results indicate that the average computation times for the noise, motion, and brightness representations are 0.85 ms, 24.1 ms, and 24.3 ms, respectively. Since these representations can be computed in parallel, the introduced latency is 24.3 ms. Given that the typical time span of input events is around 40 ms, this latency is considered acceptable.

4.5 Ablation Study

We conduct ablation studies on the RGBE-SEG dataset using the prompt mode, and the results are shown in Tab. 3. The results indicate that all representations

Table 3: Component analysis of IC^2-ESeg.

Intra-Class Consistency Mining	Noise Repr.	✗	✔	✔	✔	✔	✔	✔	✔
	Motion Repr.	✔	✗	✔	✔	✔	✔	✔	✔
	Brightness Repr.	✔	✔	✗	✔	✔	✔	✔	✔
Lightweight Information Injection	Residual Connection	✔	✔	✔	✗	✔	✔	✔	✔
	Visual Prompts	✔	✔	✔	✔	✗	✔	✔	✔
Cross-Modal Distillation	$\mathcal{L}_{sim}$	✔	✔	✔	✔	✔	✗	✔	✔
	$\mathcal{L}_{seg}$	✔	✔	✔	✔	✔	✔	✗	✔
Metric	mIoU	0.64	0.61	0.55	0.62	0.61	0.57	0.48	**0.65**
	aIoU	0.68	0.65	0.63	0.68	0.66	0.60	0.59	**0.70**

Table 4: Effectiveness validation of ICM and quantification of intra-class consistency.

(a) Effectiveness validation of ICM.

	CNN	Transformer	Handcrafted ICM
mIoU ↑	0.63	0.60	**0.65**
aIoU ↑	0.68	0.65	**0.70**

(b) Quantification of intra-class consistency.

	RGB Image	Event (Baseline)	Event (ICM)
mCS ↑	**0.85**	0.77	0.80
mFV ↓	**0.12**	0.18	0.16

in the intra-class consistency mining module contribute positively to performance. Removing the brightness representation leads to a substantial decrease in mIoU from 0.65 to 0.55, highlighting its critical role. After removing the motion representation and the noise representation, the mIoU decreases by 0.04 and 0.01, respectively, indicating that the noise representation provides relatively limited information. Removing the residual connection and the adapter-based visual prompts leads to mIoU decreases of 0.03 and 0.04, respectively, indicating that staged injection outperforms direct channel concatenation. The ablation study on the loss function demonstrates that using a single loss significantly degrades model performance. In contrast, jointly optimizing L_{sim} and L_{seg} to perform both feature-level distillation and pseudo-label supervision effectively improves segmentation performance.

In addition, we replace the handcrafted ICM with lightweight CNN- and Transformer-based learnable modules for controlled comparison. As shown in Tab. 4(a), ICM consistently achieves better mIoU and aIoU, validating the effectiveness of event-mechanism-guided representation design.

To further validate whether ICM improves intra-class consistency, we quantitatively measure feature compactness within each ground-truth mask using mean Cosine Similarity (mCS) and mean Feature Variance (mFV). As shown in Tab. 4(b), event features exhibit weaker intra-class consistency than RGB features, while ICM improves mCS by 3.9% and reduces mFV by 11.1% over the event baseline, demonstrating its effectiveness in enhancing intra-class compactness. Details are provided in Supplementary Material S3.

We further analyze the noise representation under different motion and illumination conditions. As shown in Tab. 5, the noise representation yields larger gains in slow-motion and high-light cases, while maintaining stable performance

Table 5: Ablation study of noise representation (using mIoU).

	Slow Motion	Fast Motion	Low Light	High Light
w/o Noise	0.612	0.653	0.672	0.621
w/ Noise	**0.636**	**0.658**	**0.678**	**0.639**

under fast-motion and low-light cases. These results demonstrate that the noise representation serves as an effective complementary auxiliary cue. Details are provided in Supplementary Material S3.

We further analyze whether IC^2-ESeg depends on the SAM-specific encoder. We replace the SAM encoder with encoders from other foundation models, including DINOv2 [24] and MAE [14]. IC^2-ESeg achieves 0.64 and 0.67 mIoU with DINOv2 and MAE, respectively, which is comparable to the SAM-based result of 0.65 mIoU, demonstrating the generality of our framework across different foundation model encoders.

5 Limitations

The intra-class consistency mining module operates in the preprocessing stage, which incurs additional computational overhead and latency. Furthermore, the architecture of IC^2-ESeg adopts a feedforward paradigm, performing segmentation on discrete event segments. We consider the incorporation of recurrent structures for temporal modeling to be a promising direction for future work, with the potential to further improve segmentation performance.

6 Conclusion

We propose IC^2-ESeg, an Intra-class Consistency guided, Class-agnostic Event Segmentation framework. Grounded in the event generation mechanism, IC^2-ESeg constructs noise, motion, and brightness representations to enrich event data, integrates them via lightweight information injection, and transfers SAM's general segmentation capability to the event domain through cross-modal distillation. Experimental results demonstrate that IC^2-ESeg achieves SOTA performance on the RGBE-SEG dataset. We also propose ComScene, a high resolution RGB-event segmentation dataset covering diverse scenarios. Experiments on the MVSEC and ComScene datasets further validate the strong generalization ability of IC^2-ESeg.

Acknowledgements

This work was supported by STI 2030-Major Projects (2021ZD0200300), and the State Key Laboratory of Precision Measurement Technology and Instruments (2025PMTI03).

References

1. Alonso, I., Murillo, A.C.: Ev-segnet: Semantic segmentation for event-based cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
2. Benosman, R., Clercq, C., Lagorce, X., Ieng, S.H., Bartolozzi, C.: Event-based visual flow. IEEE transactions on neural networks and learning systems **25**(2), 407–417 (2013)
3. Bentley, J.L.: K-d trees for semidynamic point sets. In: Proceedings of the sixth annual symposium on Computational geometry. pp. 187–197 (1990)
4. Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. IEEE Transactions on Image Processing **30**, 2488–2500 (2021)
5. Cao, R., Galor, D., Kohli, A., Yates, J.L., Waller, L.: Noise2image: noise-enabled static scene recovery for event cameras. Optica **12**(1), 46–55 (2025)
6. Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. arXiv preprint arXiv:2304.09148 (2023)
7. Chen, Z., Zhu, Z., Zhang, Y., Hou, J., Shi, G., Wu, J.: Segment any event streams via weighted adaptation of pivotal tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3890–3900 (2024)
8. Delbruck, T., et al.: Frame-free dynamic digital vision. In: Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society. vol. 1, pp. 21–26. Citeseer (2008)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Deng, Y., Chen, H., Chen, H., Li, Y.: Learning from images: A distillation learning framework for event cameras. IEEE Transactions on Image Processing **30**, 4919–4931 (2021)
11. Dille, S., Blondal, A., Paris, S., Aksoy, Y.: A bottom-up approach to class-agnostic image segmentation. In: European Conference on Computer Vision. pp. 345–362. Springer (2024)
12. Ding, Y., Yao, B., Liu, Y., Chen, H., Ding, D., Yang, Z., Li, Y., Deng, Y.: Evsam: Segment anything model with event-based assistance. ACM Transactions on Multimedia Computing, Communications and Applications (2026)
13. 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)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
16. Hu, Y., Delbruck, T., Liu, S.C.: Learning to exploit multiple vision modalities by using grafted networks. In: European Conference on Computer Vision. pp. 85–101. Springer (2020)

17. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the fourth Eurographics symposium on Geometry processing. vol. 7 (2006)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
19. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. IEEE journal of solid-state circuits **43**(2), 566–576 (2008)
20. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
22. Maqueda, A.I., Loquercio, A., Gallego, G., García, N., Scaramuzza, D.: Event-based vision meets deep learning on steering prediction for self-driving cars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5419–5427 (2018)
23. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
26. Qi, L., Kuen, J., Shen, T., Gu, J., Li, W., Guo, W., Jia, J., Lin, Z., Yang, M.H.: High quality entity segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4047–4056 (2023)
27. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Torr, P., Lin, Z., Jia, J.: Open world entity segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8743–8756 (2022)
28. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE transactions on pattern analysis and machine intelligence **43**(6), 1964–1980 (2019)
29. Reza, M.A., Manley, E., Chen, S., Chaudhary, S., Elafros, J.: Segbuilder: A semi-automatic annotation tool for segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8494–8503. IEEE (2025)
30. Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D.: Fast image reconstruction with an event camera. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 156–163 (2020)
31. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: Ess: Learning event-based semantic segmentation from still images. In: European Conference on Computer Vision. pp. 341–357. Springer (2022)
32. Szeliski, R.: Computer vision: algorithms and applications. Springer Nature (2022)
33. Tang, C., Wang, X., Huang, J., Jiang, B., Zhu, L., Zhang, J., Wang, Y., Tian, Y.: Revisiting color-event based tracking: A unified network, dataset, and metric. arXiv preprint arXiv:2211.11010 (2022)

34. 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 international conference on computer vision. pp. 2135–2145 (2021)
35. 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)
36. Wang, X., Li, J., Zhu, L., Zhang, Z., Chen, Z., Li, X., Wang, Y., Tian, Y., Wu, F.: Visevent: Reliable object tracking via collaboration of frame and event flows. IEEE Transactions on Cybernetics **54**(3), 1997–2010 (2023)
37. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2563–2572 (2021)
38. Yang, M., Liu, S.C., Delbruck, T.: A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding. IEEE Journal of Solid-State Circuits **50**(9), 2149–2160 (2015)
39. Yao, B., Deng, Y., Liu, Y., Chen, H., Li, Y., Yang, Z.: Sam-event-adapter: Adapting segment anything model for event-rgb semantic segmentation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 9093–9100. IEEE (2024)
40. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023)
41. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5217–5226 (2019)
42. Zhang, H., Li, F., Qi, L., Yang, M.H., Ahuja, N.: Csl: Class-agnostic structure-constrained learning for segmentation including the unseen. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7078–7086 (2024)
43. Zhao, J., Zhang, Z., Qi, F.: An intelligent image segmentation annotation method based on segment anything model. In: BenchCouncil International Symposium on Intelligent Computers, Algorithms, and Applications. pp. 237–243. Springer (2023)
44. Zhao, Y., Lyu, G., Li, K., Wang, Z., Chen, H., Yang, Z., Deng, Y.: Eseg: event-based segmentation boosted by explicit edge-semantic guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10510–10518 (2025)
45. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)
46. 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 conference on computer vision and pattern recognition. pp. 989–997 (2019)