2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

SAM-Event-Adapter: Adapting Segment Anything Model for Event-RGB Semantic Segmentation

Bowen Yao[1], Yongjian Deng[1*], Yuhan Liu[1], Hao Chen[2], Youfu Li[3], *Fellow, IEEE,* and Zhen Yang[1]

***Abstract*— Semantic segmentation, a fundamental visual task ubiquitously employed in sectors ranging from transportation and robotics to healthcare, has always captivated the research community. In the wake of rapid advancements in large model research, the foundation model for semantic segmentation tasks, termed the Segment Anything Model (SAM), has been introduced. This model substantially addresses the dilemma of poor generalizability of previous segmentation models and the disadvantage in requiring to retrain the whole model on variant datasets. Nonetheless, segmentation models developed on SAM remain constrained by the inherent limitations of RGB sensors, particularly in scenarios characterized by complex lighting conditions and high-speed motion. Motivated by these observations, a natural recourse is to adapt SAM to additional visual modalities without compromising its robust generalizability. To achieve this, we introduce a lightweight SAM-Event-Adapter (SE-Adapter) module, which incorporates event camera data into a cross-modal learning architecture based on SAM, with only limited tunable parameters incremental. Capitalizing on the high dynamic range and temporal resolution afforded by event cameras, our proposed multi-modal Event-RGB learning architecture effectively augments the performance of semantic segmentation tasks. In addition, we propose a novel paradigm for representing event data in a patch format compatible with transformer-based models, employing multi-spatiotemporal scale encoding to efficiently extract motion and semantic correlations from event representations. Exhaustive empirical evaluations conducted on the DSEC-Semantic and DDD17 datasets provide validation of the effectiveness and rationality of our proposed approach.**

I. INTRODUCTION

Semantic segmentation, as a fundamental computer vision task, has significant practical applications in areas such as autonomous driving [1], robot control [2], medical image analysis [3], *etc*. Starting from [4], most semantic segmentation works are based on traditional RGB sensors [5], [6]. However, constrained by the limitations of traditional RGB cameras, these works exhibit limited robustness [7]–[9] under extreme conditions such as high contrast (*i.e.*, over- or under-exposure) and rapid motion as shown in Fig. 1. To address these issues, approaches of fusing other modalities to enhance the segmentation performance has begun to receive widespread attention.

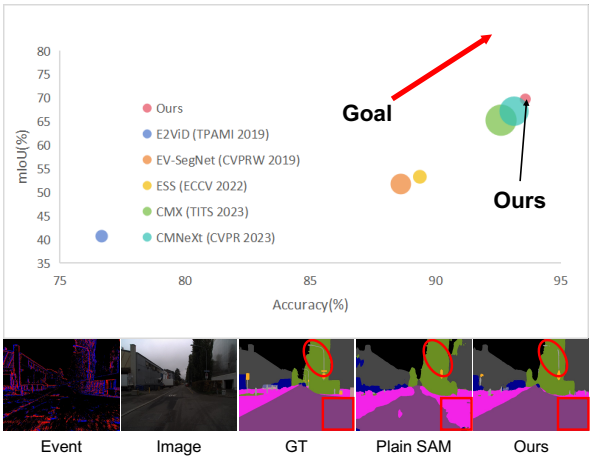

Fig. 1. **Top**: Segmentation accuracy vs mIoU on the DSEC-Semantic dataset. The circle areas are proportional to tunable parameters of each model. **Bottom**: Visual comparison of the SAM that only contains RGB modal and includes Event-RGB modalities powered by our design.

The Dynamic Vision Sensor (DVS) [10], also known as the event camera, is a novel sensor that captures scenes by detecting pixel-level changes in brightness and only generates events when there is a change in the pixels. Their advantages in high dynamic range, high temporal resolution, and accurate motion encoding make them the ideal complementary modality to RGB images for semantic segmentation. As a result, recent research tends to fuse these two vision modalities.

Initially, researches in [11]–[13] simply fuse the event and RGB features by addition or concatenation for sub-stream tasks. Methods proposed in [14]–[16] try to fuse cross-modal features via attention operations for better exploiting their complementary. However, these studies commonly neglect two key problems that exist in Event-RGB fusion. (1) Training the entire model repeatedly for each dataset introduce a tremendous waste of computing resources. (2) Due to the lack of the Event-RGB datasets, performing whole-model fine-tuning often leads to undermining of the high generalization advantage of pre-trained model weights, resulting in performance drops.

Actually, the recently proposed foundation modal SAM (Segment Anything Model) [17] has provided answers to these questions. Benefiting from its pre-trained highly gen-

[1]Bowen Yao, Yongjian Deng, Yuhan Liu and Zhen Yang are with the College of Computer Science, Beijing University of Technology, Beijing, China. {ybw0818@emails., yjdeng@, liuyuhan@emails., yangzhen@}bjut.edu.cn

[2]Hao Chen is with Key Lab of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China. haochen303@seu.edu.cn

[3]Youfu Li is with Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR. meyfli@cityu.edu.hk

This work is partially supported by National Key R&D Program of China (No. 2022YFB3103100), the National Natural Science Foundation of China (62203024, 92167102, 61873220, 62102083, 62173286, 61875068, 62177018), 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 (CityU 11213420).

*: Corresponding author

eralizable weights, many works [18], [19] achieve leading performance in specific tasks by adapting [20]–[22] the foundation model SAM with frozen weights. However, rare methods discuss how to adapt the frozen SAM for multi-modal learning, which is more common in real world. SAD [23] has achieved notable performance in open vocabulary segmentation using RGBD information. However, it places more emphasis on static geometric information. Its performance in scenarios requiring motion information, such as autonomous driving, still needs further consideration. In this work, we introduce a lightweight SAM-Event-Adapter (SE-Adapter), successfully adapting the frozen SAM to our Event-RGB multi-modal segmentation task. SE-Adapter is capable of enhancing the segmentation performance of SAM by complementing the shortage of RGB information utilizing Event data as additional features. In each SE-Adapter (Fig. 2), we perform a image-guided feature projection by treating image features from the SAM as queries, treating event features as keys and values, then processing them with attentive operations. This way, we aim to let event features learn adaptively according to their corresponding image features for achieving better fusion effects. Then, a gate block is employed to re-weight projected event features for balancing the impacts from event and image features in the SAM learning process. Finally, we feed the weighted event features back to the SAM for further consideration. Thanks to the proposed lightweight SE-Adapter (only 0.8M parameters for each), we achieve state-of-the-art segmentation results by adapting the foundation model with frozen weights.

Furthermore, considering the transformer-based architecture of the SAM model, the efficient encoding of event representations into patch form represents another pivotal challenge. To address this issue, existing works [24]–[30] typically first convert event signals into a voxel-grid representation [31], wherein each channel in the voxel-grid encapsulates event features from distinct temporal ranges. Following the processing conventions in [32], information from multiple channels within the same spatial region is integrated into a single patch to serve as the model input. This method undoubtedly hinders the precision of motion scene encoding. To alleviate this, we propose a novel paradigm termed Multi-Spatiotemporal-Scale Spiking Patch Embedding (MSP), which provides a more efficacious mechanism for embedding events into patches compatible with transformer-based models. Specifically, we aggregate events across various temporal scales into a series of voxel-grids and process them utilizing spiking-convolution blocks to achieve spatiotemporal encoding with different receptive fields. By learning event features with varying receptive fields across the spatial and temporal dimensions, the MSP enables the resultant event patches to encapsulate comprehensive spatiotemporal messages, thereby furnishing subsequent learning models with highly efficient initial event features.

To summarize, our main contributions are as followed:

- To the best of our knowledge, we are the first to adapt the foundation model SAM to Event-RGB multi-modal segmentation task. With the help of the proposed lightweight SAM-Event-Adapter (SE-Adapter), our approach achieves leading performance with limited tunable parameters.
- We provide a new paradigm, named Multi-Spatiotemporal-Scale Spiking Patch Embedding (MSP), to effectively embed event data into event patches that are compatible with transformer-based models through multi-scale spatiotemporal encoding.
- Comprehensive empirical analyses performed on both the DSEC-Semantic [33] and DDD17 [34] datasets demonstrate the effectiveness of our proposed method.

II. RELATED WORKS

A. Semantic Segmentation with Event Cameras

Compared to frame-based RGB sensors, event cameras offer higher temporal resolution, dynamic range, and lower power consumption, which make them suitable for deployment on platforms such as vehicles and drones. Event-based semantic segmentation is an emerging field, and it is experiencing rapid growth in interest. EV-SegNet [35] is the first event-based semantic segmentation work, where they follow the approaches used in RGB segmentation fields for event representation learning. Work in [36] provides several Spiking Neural Network structures to further explore the abundant temporal information behind event data. Methods mentioned in [37]–[40] improve performance significantly in this task with the help of the effectiveness of the attention module for semantic encoding. However, these studies are all based on learning architectures that are based on event data individually, which are inevitably constrained by the limitations of event cameras in capturing color and low-contrast scenes, resulting in unsatisfactory segmentation performance.

Observing the aforementioned issues, recent researchers focus on the fusion of event and RGB modalities for segmentation tasks. For instance, ESS [41] fuses knowledge from both modalities through the reconstruction process from event data to RGB images. Inspired by ESS, Zhang *et al.* [12], [13] introduce the sparse-to-dense and dense-to-sparse translations to fuse the image and event flow. Methods mentioned in [15], [16] provide RGB-X semantic segmentation models based on transformer-based networks, with their feature fusion component suitable for multiple modalities including event data. While these Event-RGB fusion studies have made progress in performance gain, they are hampered by the lack of available datasets, making it challenging to attain weights with high generalizability. This necessitates the repetitive training of the entire model $w.r.t$ various datasets, thereby leading to considerable computational overhead. Inspired by advancements in large model research, our work adapts the SAM for multi-modal segmentation based on Event-RGB data by leveraging the proposed lightweight SE-Adapter module. This adaptation enables the effective utilization of SAM's intrinsically high generalizability, obviating the need for redundant retraining.

B. Adaptation in Segment Anything Model

Recently, pre-trained large models [42]–[45] show signif-

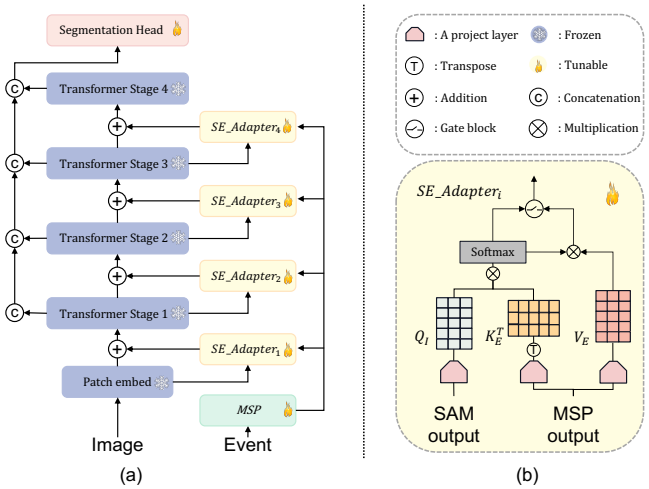

Fig. 2. (a) The overall architecture of our model. During training, we freeze the weights of the SAM backbone (blue blocks) and optimize adapter structures (SE-Adapter), the event-based patch embedding module (MSP), and the segmentation head using task-specific guidance. (b) The design of our SE-Adapter, which is composed of a cross-modal attention mechanism and a gate block. Best viewed in color.

icant potential in the areas of natural language processing and computer vision. However, despite the fact that their generalizable weights allow them to extract excellent semantic information from inputs, fine-tuning these models for downstream tasks can incur significant training costs due to their large parameter count. Adaptation methods have proven effective in addressing this issue.

The adaptation method is first applied in natural language processing [46]. In the field of computer vision, Adaptformer [47] adds a bottleneck connection in the MLP layer of Vit to adapt to different downstream tasks. Vit-adapter [20] applies cross-attention to inject information into the VIT backbone. With the advent of SAM, its outstanding performance has garnered increasing attention. As a result, adaptation methods based on SAM have started to show up. SAM-Adapter [18] is the first application to introduce adapters in SAM and achieves promising results on several tasks. Medical SAM Adapter [19] extends SAM's application to medical image segmentation tasks. However, due to the complexity of dense prediction downstream tasks and the differences between modalities, there is currently no SAM adaptation method specifically designed for Event-RGB semantic segmentation tasks. Existing methods are difficult to apply directly on it due to the lack of modality-specific design, and we bridge this gap in this paper.

III. Preliminary

A. Event Representation

The i-th event e_i in an event stream can be represented as (x_i, y_i, p_i, t_i), where x_i and y_i denote the spatial coordinates, p_i denotes the polarity of the event, and t_i denotes the timestamp of the event. Due to the sparse, noisy, and unstructured nature of the input event stream, the common way [31] to represent event data is to discretize the time dimension into B consecutive temporal bins and then integrate events into a 3D spatiotemporal voxel-grid ($E \in \mathbb{R}^{B\times H\times W}$, $\{H, W\}$ represent the resolution of the input) linearly. The integration of a specific temporal bin can be formulated as Eq. (1).

$$E(m) = \sum_i p_i \max\left(0, 1 - \left|m - \frac{t_i - t_0}{t_{N_e} - t_0}(B-1)\right|\right), \quad (1)$$

where t_0 and t_{N_e} respectively denote the start and end time of the integrated event stream, and N_e represents the number of event data. The range of m is in $[0, B-1]$.

In our approach, we first obtain a series of voxel-grids $\{E_1, E_2, \ldots, E_k\}$ that are integrated from multi-scale temporal ranges, all ended at the timestamp where we perform segmentation and starting from defined timestamps. Then converting $E_k \in \mathbb{R}^{B\times H\times W}$ into MSP-compatible format $\mathcal{E}_k \in \mathbb{R}^{T\times 1\times H\times W}$, where T represents the dimension of the time step and is equal to B in our case.

B. Spiking Neuron Network

Spiking Neuron Networks (SNNs) can handle event representations recurrently according to the trigger sequence of event data while maintaining energy-efficient computation, resulting in the preservation of motion cues of event data in both accurate and efficient considerations. Inspired by [48], we adopt a type of SNN structure, Leaky Integrate-and-Fire (LIF) [49], as basic units in the MSP module. The principle of the LIF can be formulated as:

$$H(t) = V(t-1) + \frac{1}{\tau}(X(t) - (V(t-1) - V_{reset}), \quad (2)$$

$$S(t) = Heaviside(H(t) - V_{th}), \quad (3)$$

$$V(t) = H(t)(1 - S(t)) + V_{reset}S(t), \quad (4)$$

where $X(t)$ is the input at time step t, $H(t)$ and $V(t)$ represent the membrane potential after leak and charge, respectively. V_{th} denote the fire threshold, $Heaviside()$ is the Heaviside step function, V_{reset} denotes the reset potential, τ is the membrane time constant. As the name suggests, LIF accumulates membrane potential when receiving spikes, and the membrane potential decays when there are no pulses. When the membrane potential exceeds a certain threshold, it will fire a spike. Such a learning unit helps extract temporal information from event representations and filters out noise.

C. Problem Formulation

We take images $I \in \mathbb{R}^{3\times H\times W}$ and event representations $\mathcal{E}_k \in \mathbb{R}^{T\times 1\times H\times W}$ as inputs for the SAM and the proposed MSP respectively. We obtain the event patches from the MSP module and input them along with the image features from the SAM into the introduced SE-Adapter for cross-modal adaptation. Subsequently, interactions take place within the adapter and SAM, and the segmentation results are finally generated through the segmentation head.

IV. Method

A. Overall Architecture

As shown in Fig. 2, our network is composed of: (i) A frozen SAM [45] backbone used for extracting highly generalized features from multi-modal input; (ii) A Multi-Spatiotemporal-Scale Spiking Patch Embedding module (MSP) for obtaining event-based patches with comprehensive

motion and semantic cues; (iii) SAM-Event-Adapters (SE-Adapter) for adapting the SAM in multi-modal segmentation by introducing event features; and (iv) a SegFormer decoder [50] used for outputting segmentation results. In the following, we will detail the design of core components of our method such as the SE-Adapter and MSP module.

B. SAM-Event-Adapter (SE-Adapter)

The detailed architecture of the proposed SE-Adapter is illustrated in Fig. 2.(b). As shown in the figure, we divide SAM into four stages, where each stage contains eight transformer blocks, and employ the SE-Adapter before each stage in order to achieve targets of both the cross-modality fusion and the sub-stream task adaptation.

Unlike previous adaptation methods designed for SAM, our work includes an additional visual modality, *i.e.,* event data. This means that besides adapting the SAM to specific downstream tasks, our model needs to leverage the complementary between event and image data, and contribute these two modalities to the segmentation task collaboratively through the frozen SAM. To accomplish this, we introduce an image-guided attention mechanism, as illustrated in Fig. 2(b), to construct the SE-Adapter. Specifically, we treat image features from the SAM as queries ($Q_I \in \mathbb{R}^{N^I \times C_p}$) and event features from the MSP as keys ($K_E \in \mathbb{R}^{N^E \times C_p}$) and values ($V_E \in \mathbb{R}^{N^E \times C_p}$), where N^I and N^E are numbers of patches determined by input resolution. Our motivation is to use the highly generalizable features learned from the SAM to guide the event feature encoding procedure. We aim to let the network know which event-based information can be exploited as supplementary cues for enhancing segmentation performance. In this way, high-quality weights of the SAM that are pre-trained using RGB images can be fully utilized. Consequently, we treat image features as queries to determine the way that event features should be extracted and added back to the SAM backbone. The attention process can be formulated as:

$$V_E^* = \mathcal{M}V_E = Softmax(\frac{Q_I \times K_E^T}{\sqrt{C_p}})V_E. \quad (5)$$

Noticeably, projection layers are employed before the attention operation for cross-modal feature dimension unification. $V_E^* \in \mathbb{R}^{N^I \times C_p}$ is the extracted event features corresponding to image patches.

Moreover, we propose an adaptive gate block, which is composed of a re-weighting matrix and a projection layer, to control the impact of messages from different modalities.

$$\Gamma = Sigmoid(ReLU(norm(linear(\mathcal{M})))), \quad (6)$$

where $\Gamma \in \mathbb{R}^{N^I \times C_p}$ represents the re-weighting matrix, $Sigmoid$ is used to normalize the Γ to the range of 0 to 1. We then element-wise multiply Γ with the obtained V_E^*. Afterward, we pass it through a projection layer and add it back to the SAM.

TABLE I

RESULTS ON THE DSEC-SEMANTIC. †: E AND I REPRESENT EVENT DATA AND RGB IMAGES RESPECTIVELY.

Methods	Inference Modality†	Accuracy [%]	mIoU [%]	Tunable Params(M)
E2ViD [51]	E	76.67	40.70	<u>10.71</u>
EV-SegNet [35]	E	88.61	51.76	29.09
ESS [41]	E	89.37	53.29	12.91
EDCNet-S2D [13]	E & I	92.39	55.75	61.99
CMX [15]	E & I	92.61	65.29	66.5
CMNeXt [16]	E & I	<u>93.13</u>	<u>67.2</u>	58.69
Ours	E & I	**93.58**	**69.77**	**8.2**

C. Muti-spatiotemporal-scale Spiking Patch Embedding

In this section, we propose a new paradigm for processing event data. Previous works mostly treat the time steps of event data as individual channels. Some approaches embed multiple temporal channels into a single patch, akin to frame-based methods, which leads to a loss of the inherent temporal relationships of event data. Other methods generate a patch for each temporal channel, but this introduces a high computational burden due to the large number of patches. As we mentioned in Sec. III-B, SNN can accumulate event signals in the time domain while maintaining lightweight model size. This design allows SNN to effectively preserve the temporal information of event data without compressing channels or introducing a significant computational burden. Based on these reasons, we design the Multi-Scale Spiking Patch Embed Block (MSP) developed from LIF neurons and multi-scale embedding strategy, as shown in Fig. 3. We utilize the method mentioned in Sec. III-A to aggregate event data into voxel-grids with various temporal scales ($\{\mathcal{E}_1 ... \mathcal{E}_k\} \in \mathbb{R}^{T \times 1 \times H \times W}$, where T is the fixed time step), obtaining coarse features at multiple temporal scales. Subsequently, we feed $\mathcal{E}_k$ into ConvLIF blocks for spatiotemporal feature learning. Each ConvLIF module consists of multiple convolutions, batch normalization, and LIF neurons. Particularly, ConvLIF$^{4\times}$ also includes a max-pooling layer. Specifically, the event representation with the maximum time range $\mathcal{E}_1$ pass through only one ConvLIF$^{4\times}$ module, while the voxel-grid with the minimum time range $\mathcal{E}_k$ pass through one Conv-LIF$^{4\times}$ module and $k-1$ ConvLIF$^{2\times}$ modules. This processing approach retains the temporal information of the events and connects the temporal and spatial aspects of event information effectively.

After being processed by ConvLIF blocks, the event representations have a shape of $\mathbb{R}^{T \times kC \times H \times W}$, where C is defined intermediate feature dimension. We calculate the average along the temporal dimension and flatten them along the spatial dimension. Next, we convert the obtained features into patches with Avgpooling operation on T dimension and follow it with a linear layer to project the feature dimension from C_p. Finally, we concatenate patches derived from different integration spans and attain the output of the MSP with the shape of $\mathbb{R}^{N^E \times C_p}$.

V. EXPERIMENTS

In this section, we compare our method with state-of-the-art (SOTA) semantic segmentation approaches in two

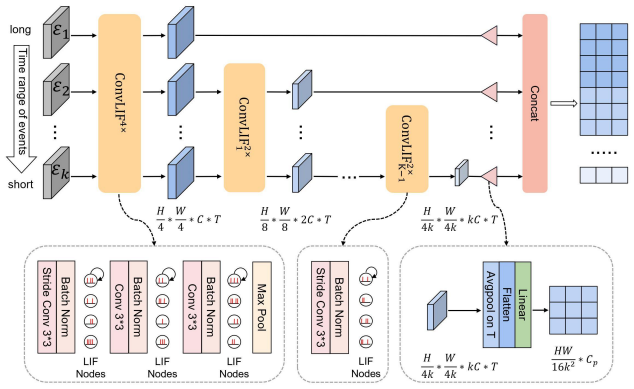

Fig. 3. The structure of the MSP. $\mathcal{E}_k$ represents different temporal ranges of event representations. The ConvLIF$^{4\times}$ and ConvLIF$^{2\times}$ represent convolution-LIF blocks with $4\times$ and $2\times$ downsampling separately.

TABLE II

RESULTS ON THE DDD17.

Methods	Inference Modality	Accuracy [%]	mIoU [%]	Tunable Params(M)
E2ViD [51]	E	83.24	44.77	10.71
ViD2E [52]	E	90.19	56.01	29.09
EV-SegNet [35]	E	89.76	54.81	29.09
ESS [41]	E	90.37	60.43	12.91
EDCNet-S2D [13]	E & I	93.71	60.23	61.99
CMX [15]	E & I	94.20	67.47	66.5
CMNeXt [16]	E & I	93.82	66.99	58.69
Ours	E & I	**95.32**	**69.06**	**8.2**

categories. (1) Event-based segmentation, *e.g.*, EV-SegNet [35], ESS [41], E2ViD [51] *etc.* (2) Event-RGB segmentation, *e.g.*, EDCNet-S2D [13], CMX [15] and CMNeXt [16]. For approaches that have not been evaluated on our adopted datasets, we re-train them with the identical training settings as ours.

A. Experiment on DSEC-Semantic

1) ***Dataset:*** The DSEC-Semantic dataset [41] contains 53 driving sequences with 8082 training samples and 2809 testing samples from DSEC [33], which data was collected in various urban and rural environments, using automotive-grade standard cameras and high-resolution event cameras. DSEC-Semantic includes RGB and event data, with semantic segmentation labels provided by a model introduced in [53].

2) ***Training Settings:*** We construct four $\mathcal{E}$ as inputs of the MSP ($k = 4$) and set the integrating span of these four voxel-grids as $\{30ms, 50ms, 70ms, 90ms\}$. For each $\mathcal{E} \in \mathbb{R}^{T\times 1\times H\times W}$, we define their time step T as 20. We set the feature dimension C_p of each patch derived from the MSP as 256, and we set C in MSP as 32. For data augmentation, we crop the input frames and their paired event representations to a size of 256×256 and apply rotation and flipping randomly. The initial learning rate was set to (2×10^{-4}), with a minimum learning rate of 10^{-7}, adjusted by Cosine Annealing over 20 epochs. To facilitate fair comparisons, we employed identical hyper-parameters as our model for the corresponding models. Specifically, we set the integrating span of the voxel-grid used by the corresponding model to 90ms, as they only support event inputs in a single temporal scale.

3) ***Quantitative Results:*** We use accuracy and mean intersection over union (mIoU) to evaluate the effectiveness of our model. As shown in Tab. I, when compared to event-based segmentation methods like E2ViD [51], EV-SegNet [35], and ESS [41], our model achieves significantly higher results (at least 16.48% higher) with fewer tunable parameters (at least 2.51M lower). In comparison to the SOTA Event-RGB segmentation models such as CMX [54] and CMNeXt [55], our method still holds leading performance with lower tunable parameters. We attribute the observed phenomena to two factors: (1) Constrained by the limited scale of Event-RGB segmentation datasets, prior models in the domain of Event-RGB segmentation fail to achieve sufficient optimization, thereby lacking robust predictive capabilities across diverse scenarios. In contrast, our methodology, anchored by the high generalizability of frozen SAM, yields high-quality, transferable features across the Event-RGB modalities, consequently manifesting superior performance during the testing phase. (2) Our proposed Multi-Spatiotemporal-Scale Spiking Patch Embedding (MSP) module is adept at efficiently encoding semantic information and motion cues within event data, thereby furnishing subsequent neural network training with a well-initialized features.

4) ***Qualitative Results:*** As shown in Fig. 4, compared to prior Event-RGB segmentation models, our method obtains more detailed segmentation results, as highlighted in red boxes. Since the images in DSEC-Semantic are captured by a moving vehicle, small-scale objects such as streetlights and traffic signs are highly susceptible to motion blur, making them challenging to segment. The compared Event-RGB models might not handle this situation well due to their frame-like method on embedding event data. Instead, the powerful spatiotemporal information encoding capability of the MSP enables our model to handle the motion cues efficiently, resulting in more accurate prediction.

B. Experiment on DDD17

1) ***Dataset:*** Similar to DSEC-Semantic, DDD17 [34] includes 40 different driving sequences along with synchronized grayscale images and event streams. The grayscale images in DDD17 suffer from issues of low quality and low resolution. Moreover, there are artifacts present in the semantic segmentation labels. As a result, this dataset poses a greater challenge to assess the SE-Adapter's ability to handle the event modality and optimize image features within the SAM backbone.

2) ***Training Settings:*** The Training settings on DDD17 share many similarities with those on DSEC-Semantic. The difference is that we construct four $\mathcal{E}$ as inputs of the MSP ($k = 3$) and set the integrating duration of these four voxel-grids as $\{10ms, 50ms, 250ms\}$. The settings of the corrsponding models are samilar with the settings in DSEC-Semantic.

3) ***Results:*** The performance of our model on DDD17 is shown in Tab. II. Similar to DSEC-Semantic, we com-

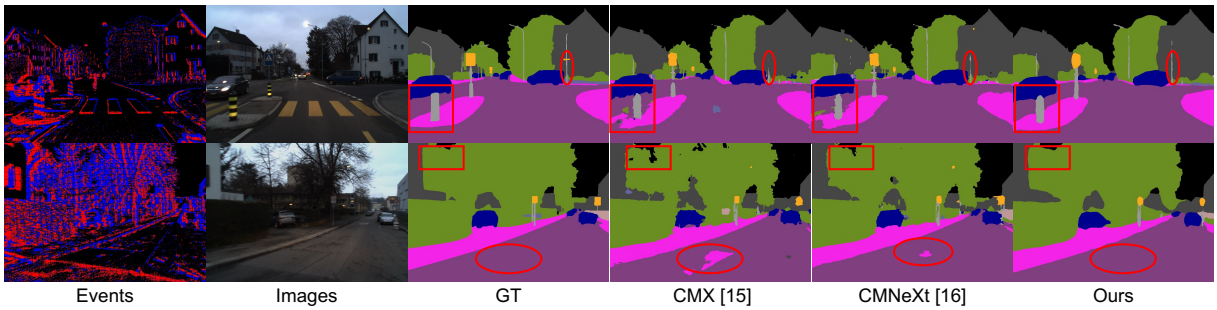

Fig. 4. Visual comparison results on DSEC-Semantic of different methods. The highlighted regions for comparison are within bounding boxes.

TABLE III

THE ABLATION STUDY FOR THE SE-ADAPTER. ♮: FINE-TUNING SEGMENTATION HEAD SOLELY. §: FINE-TUNING THE WHOLE MODEL INCLUDING THE SAM.

Variants	Methods	Modalities	Accuracy [%]	mIoU [%]
A	SAM♮	I	91.21	59.7
B	SAM w/ SE-Adapter§	E & I	87.98	47.3
C	**SAM w/ SE-Adapter (ours)**	E & I	**93.58**	**69.77**

TABLE IV

THE ABLATION STUDY FOR THE MSP.

Variants	ConvLIF	Muti-scale Embedding	mIoU [%]
A			63.27
B	✓		67.66
C		✓	67.64
D	✓	✓	**69.77**

pared our approach with state-of-the-art (SOTA) methods and found that the performance of our method is still outperforms, maintaining consistency across different datasets. This indicates that even on dataset like DDD17 with low image quality, which may limit the effectiveness of the SAM backbone, our model remains effective, demonstrating the generalizability of our model and the efficacy of our proposed SE-Adapter in RGB feature enhancement.

C. Ablation studies

In order to validate the effectiveness of the model design and optimizing strategy, we conduct several ablation experiments on DSEC-Semantic, and the results in Tab. III & Tab. IV confirm that each of our designs is rational and indispensable.

*1) **Efficacy of the SE-Adapter**:* Although fine-tuning SAM consumes substantial computational resources, it seems justified to allocate additional resources when computational capacity is abundant, in pursuit of higher accuracy. However, the results (B & C) in Tab. III refute this conjecture and demonstrate the necessity of the adapter structure. By comparing the outcomes of the models mentioned in the table, it is evident that fine-tuning the entire model (including the SAM) actually reduces model accuracy; this is because the limited scale of Event-RGB segmentation datasets makes it challenging for the parameters of the large model SAM to be well optimized. Instead, only fine-tuning the lightweight SE-Adapter and segmentation head eases the training pressure largely and thereby results in better prediction. Also, the Tab. III (A & C) shows the proposed SE-Adapter enhances the final performance effectively through fusing additional event-based modal.

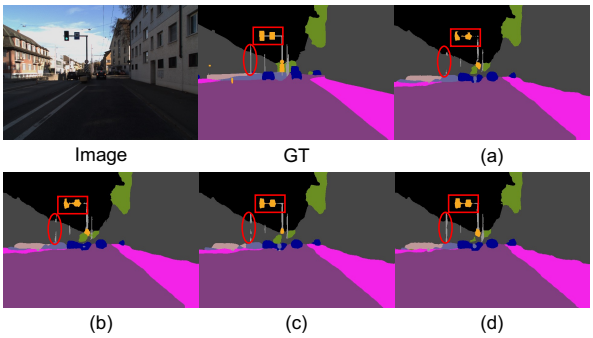

Fig. 5. Visual comparison on different designs, where (a), (b), (c), (d) represent the segmenting results of variants A, B, C, D in the table.

*2) **Designs in the MSP**:* To determine whether MSP can indeed be a new paradigm for processing event data, we evaluate each part of the MSP separately and place the results in Tab. IV. In the setting A, we replace the LIF neurons in ConvLIF with ReLU activation functions and alter the multi-scale representations to a single voxel-grid with an integrating span of $90ms$. Then, we add the LIF neuron and the multi-scale design to the model individually to achieve settings B and C. Finally, we place the performance of the full-power MSP module as variant D. From the table, it can be observed that each part of the MSP contributes positively to the model, and the complete MSP achieves the highest mIoU. The visualization in Fig. 5 corresponding to settings A, B, C, D also validates the efficacy of each component in the MSP.

VI. CONCLUSIONS

In this work, we introduce the SAM-Event-Adapter (SE-Adapter) and Multi-Spatiotemporal-Scale Spiking Patch Embedding (MSP) modules, addressing critical challenges in model adaptability concerning multi-modal fusion and spatiotemporal encoding of event data. Extensive experiments show the effectiveness of our design, and validate the success of our model in adapting the SAM with frozen weights to our Event-RGB multi-modal segmentation tasks. We believe that the SE-Adapter will not only offer novel avenues for research in the Event-RGB domain but also find applicability in a broader field of cross-modal fusion. Concurrently, our MSP module can be seamlessly integrated as a foundational component into any research endeavors that leverage event-based data.

REFERENCES

[1] X. Li, G. Zhang, H. Pan, and Z. Wang, "Cpgnet: Cascade point-grid fusion network for real-time lidar semantic segmentation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 117–11 123.

[2] V. Holomjova, A. J. Starkey, and P. Meißner, "Gsmr-cnn: An end-to-end trainable architecture for grasping target objects from multi-object scenes," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3808–3814.

[3] Y. Lu, Y. Shen, X. Xing, and M. Q.-H. Meng, "Multiple consistency supervision based semi-supervised oct segmentation using very limited annotations," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8483–8489.

[4] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1629541

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:3429309

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3638670

[7] W. Zhen and S. A. Scherer, "Estimating the localizability in tunnel-like environments using lidar and uwb," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4903–4908, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198910004

[8] Z. Wu, S. Gobichettipalayam, B. Tamadazte, G. Allibert, D. P. Paudel, and C. Demonceaux, "Robust rgb-d fusion for saliency detection," *2022 International Conference on 3D Vision (3DV)*, pp. 403–413, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251280370

[9] J. Lin and F. Zhang, "R3live: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10 672–10 678, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237532664

[10] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 128 120 db 15 s latency asynchronous temporal contrast vision sensor," 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:2497402

[11] A. Tomy, A. K. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, "Fusing event-based and rgb camera for robust object detection in adverse conditions," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 933–939, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250512606

[12] J. Zhang, K. Yang, and R. Stiefelhagen, "Issafe: Improving semantic segmentation in accidents by fusing event-based data," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1132–1139, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221186954

[13] J. Zhang, K. Yang, X. Hu, and R. Stiefelhagen, "Exploring event-driven dynamic context for accident scene segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 2606–2622, 2021.

[14] L. Sun, C. Sakaridis, J. Liang, P. Sun, J. Cao, K. Zhang, Q. Jiang, K. Wang, and L. V. Gool, "Event-based frame interpolation with ad-hoc deblurring," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 043–18 052, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:255749379

[15] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[16] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1136–1147.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[18] T. Chen, L. Zhu, C. Ding, R. Cao, Y. Wang, Z. Li, L. Sun, P. Mao, and Y.-D. Zang, "Sam fails to segment anything? – sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258437348

[19] J. Wu, R. Fu, H. Fang, Y. Liu, Z.-Y. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *ArXiv*, vol. abs/2304.12620, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258309597

[20] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *ArXiv*, vol. abs/2205.08534, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248834106

[21] Y. Li, H. Mao, R. B. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," *ArXiv*, vol. abs/2203.16527, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247793203

[22] W. Liu, X. Shen, C.-M. Pun, and X. Cun, "Explicit visual prompting for low-level structure segmentations," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 434–19 445, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257631722

[23] J. Cen, Y. Wu, K. Wang, X. Li, J. Yang, Y. Pei, L. Kong, Z. Liu, and Q. Chen, "Sad: Segment any rgbd," 2023.

[24] Y. Deng, H. Chen, H. Liu, and Y. Li, "A voxel graph cnn for object classification with event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1172–1181.

[25] Y. Deng, H. Chen, and Y. Li, "Mvf-net: A multi-view fusion network for event-based object classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8275–8284, 2022.

[26] A. Sabater, L. Montesano, and A. C. Murillo, "Event transformer+. a multi-purpose solution for efficient event data processing," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253760920

[27] J. Zhao, S. Zhang, and T. Huang, "Transformer-based domain adaptation for event data classification," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4673–4677, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249436566

[28] Y. Deng, H. Chen, B. Xie, H. Liu, and Y. Li, "A dynamic graph cnn with cross-representation distillation for event-based recognition," *arXiv preprint arXiv:2302.04177*, 2023.

[29] Z. Zhu, J. Hou, and D. O. Wu, "Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers," *ArXiv*, vol. abs/2307.04129, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259501878

[30] Y. Deng, H. Chen, H. Chen, and Y. Li, "Learning from images: A distillation learning framework for event cameras," *IEEE Transactions on Image Processing*, vol. 30, pp. 4919–4931, 2021.

[31] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 989–997.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 000–16 009.

[33] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, pp. 4947–4954, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232170230

[34] J. Binas, D. Neil, S.-C. Liu, and T. Delbrück, "Ddd17: End-to-end davis driving dataset," *ArXiv*, vol. abs/1711.01458, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:396580

[35] I. Alonso and A. C. Murillo, "Ev-segnet: Semantic segmentation for event-based cameras," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1624–1633, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54063435

[36] Y. Kim, J. Chough, and P. Panda, "Beyond classification: directly training spiking neural networks for semantic segmentation,"

Neuromorphic Computing and Engineering, vol. 2, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:239009877

[37] R. Hamaguchi, Y. Furukawa, M. Onishi, and K. Sakurada, "Hierarchical neural memory network for low latency event processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 867–22 876.

[38] Z. Jia, K. You, W. He, Y. Tian, Y. Feng, Y. Wang, X. Jia, Y. Lou, J. Zhang, G. Li, and Z. Zhang, "Event-based semantic segmentation with posterior attention," *IEEE Transactions on Image Processing*, vol. 32, pp. 1829–1842, 2023.

[39] J. Tsai, C.-C. Chang, and T. Li, "Autonomous driving control based on the technique of semantic segmentation," *Sensors*, vol. 23, no. 2, p. 895, 2023.

[40] S. Zhang, L. Sun, and K. Wang, "A multi-scale recurrent framework for motion segmentation with event camera," *IEEE Access*, vol. 11, pp. 80 105–80 114, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260298093

[41] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza, "Ess: Learning event-based semantic segmentation from still images," pp. 341–357, 2022.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[43] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *ArXiv*, vol. abs/2106.08254, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235436185

[44] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233444273

[45] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, "Segment anything," *ArXiv*, vol. abs/2304.02643, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257952310

[46] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:59599816

[47] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *ArXiv*, vol. abs/2205.13535, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249097890

[48] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, and Y. Tian, "Deep residual learning in spiking neural networks," in *Neural Information Processing Systems*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235359262

[49] A. Delorme, J. Gautrais, R. van Rullen, and S. J. Thorpe, "Spikenet: A simulator for modeling large networks of integrate and fire neurons," *Neurocomputing*, vol. 26-27, pp. 989–996, 1999.

[50] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[51] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 1964–1980, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:189998802

[52] D. Gehrig, M. Gehrig, J. Hidalgo-Carri'o, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3583–3592, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:214743146

[53] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *ArXiv*, vol. abs/2005.10821, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218763375

[54] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[55] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136–1147, 2023.