Segment Any Event Streams via Weighted Adaptation of Pivotal Tokens

Zhiwen Chen[1], Zhiyu Zhu[2] *, Yifan Zhang[2], Junhui Hou[2], Guangming Shi[1,3], and Jinjian Wu[1,3]
Xidian University[1], City University of Hong Kong[2], Pazhou Lab[3]
zhiwen.chen@stu.xidian.edu.cn, zhiyuzhu2-c@my.cityu.edu.hk

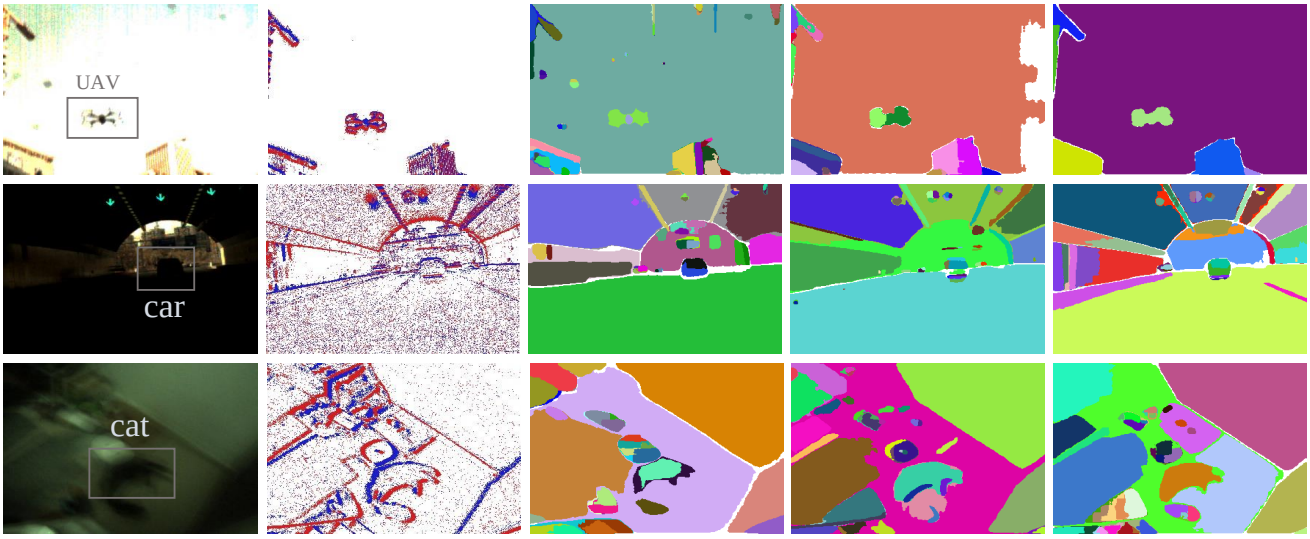

Figure 1. Illustration of object segmentation with different Segment Anything Models (SAMs)-based methods, where (a) RGB images, (b) event data (polarity map for visualization), (c) segmentation results by the original SAM w/ RGB images, (d) the original SAM w/ event data, and (e) our fine-tuned SAM w/ event data. **Please refer to the supplementary for more demos and applications.**

Abstract

In this paper, we delve into the nuanced challenge of tailoring the Segment Anything Models (SAMs) for integration with event data, with the overarching objective of attaining robust and universal object segmentation within the event-centric domain. One pivotal issue at the heart of this endeavor is the precise alignment and calibration of embeddings derived from event-centric data such that they harmoniously coincide with those originating from RGB imagery. Capitalizing on the vast repositories of datasets with paired events and RGB images, our proposition is to harness and extrapolate the profound knowledge encapsulated within the pre-trained SAM framework. As a cornerstone to achieving this, we introduce a multi-scale feature distillation methodology. This methodology rigorously optimizes the alignment of token embeddings originating from event data with their RGB image counterparts, thereby preserving and enhancing the robustness of the overall architecture. Considering the distinct significance that token embeddings from intermediate layers hold for higher-level embeddings, our strategy is centered on accurately calibrating the pivotal token embeddings. This targeted calibration is aimed at effectively managing the discrepancies in high-level embeddings originating from both the event and image domains. Extensive experiments on different datasets demonstrate the effectiveness of the proposed distillation method. Code in https://github.com/happychenpipi/EventSAM.

1. Introduction

As the vanguard of sensor technology, event cameras offer a suite of compelling advantages that compensate traditional image sensors. These advantages include, but are not limited to, unparalleled temporal resolution [48, 52], expansive dynamic range [18, 40, 53], markedly reduced latency [13], and commendable energy efficiency [47]. Current research in event cameras has showcased their vast potential through a multitude of applications, including object classification

* The first two authors contributed to this paper equally. Corresponding author: Zhiyu Zhu.

[11], precise detection [1, 35, 65], tracking [41, 73], and segmentation [2, 60]. Despite these advances, event-based research always faces with the challenge of the limited scale of annotated datasets, which are often constrained in both scale and diversity. This bottleneck severely hinders the ability of algorithms to generalize beyond controlled environments and to process the variable and unpredictable nature of real-world scenes effectively.

In the contemporary landscape of visual perception, there has been a gradual shift, primarily fueled by the advanced capabilities of deep learning architecture techniques, such as transformers. These innovations have yielded a new era of performance in image-based tasks, achieving unprecedented success in object recognition [57], detection [22], and segmentation [60]. Yet, this progress has predominantly occurred under nearly ideal conditions, which include well-controlled lighting and minimal object movement, creating an observable disconnect from the complex and often less-than-ideal conditions of real-world applications. Traditional image processing techniques frequently decline when faced with challenging environments, such as poorly illuminated areas [74] or scenes containing high-speed objects [5]. To bridge this divide, there is a pressing need to combine the unique sensing characteristics of event-based cameras with the extensive repositories of knowledge embedded within large-scale image datasets and the advanced learning capabilities of pre-trained models. Doing so could catalyze the development of more robust and universal visual perception systems that can confidently navigate and interpret complex, dynamic environments, shown as Fig. 1.

Due to the huge distribution gap between the image and event domains, it's not easy for us to directly adapt pre-trained models onto the event data. Fortunately, a considerable number of datasets with paired image and event data have been proposed recently [62, 73, 86]. To take advantage of such abundant cross-modal pairs, we propose to transfer the rich knowledge of pre-trained SAM to event domain. Given that the dominant information and weights are in encoder part of SAM, i.e., the ViT backbone [80], we enforce a calibration of high-level token embeddings from event domain to the image embeddings. Specifically, to avoid the loss of knowledge and well adapt the distribution of image and event data, we mix event and a minimal subset of image tokens together into student-ViT backbone and minimize the gap between the hierarchical embeddings from event and image domains. However, owing to the inherent distinction between the event and image domains, fully eliminating the discrepancies between different modalities is not only intractable but also deemed impossible. Consequently, we put forth a proposition to prioritize network optimization on the intermediate pivotal token embeddings, thereby effectively alleviating the gap that exists within the deeper layers.

The contributions of the proposed method mainly lie in following three-fold:

- we make the first attempt to adapt the SAMs for event data, which results in event-based universal object segmentation models;
- we propose to weight the regularization of the intermediate token embeddings by a approximated significance, thereby facilitating the refinement of the embeddings in the terminal layer; and
- we carry extensive experiments to evaluate the effectiveness of the proposed method.

2. Related Work

2.1. Cross-modal Knowledge Distillation and Transfer

Knowledge distillation and transfer of classical neural networks. Knowledge distillation or transfer is a common tool for researchers to achieve efficient visual recognition and adaptation of different domains or modalities [23, 24, 36]. Thoker *et al.* [63] utilized the mutual learning techniques with multiple students to distill knowledge from an image pre-trained recognition network onto human skeleton data for action recognition. Hu *et al.* [29] proposed an unsupervised knowledge distillation framework to calibrate the embedding distribution from achieving cross-modal hashing. Gupta *et al.* [25] made use of the knowledge from the teacher model to train the student model on a new unlabeled input modality, e.g., the paired depth image and optical flow. Garcia *et al.* [19] achieved cross-modality distillation with an additional modality of depth image to reconstruct a hallucination stream. Tian *et al.* [64] applied a contrastive loss to distillate pair-wise relationships across different modalities. Roheda *et al.* [49] proposed to compensate for the missing modalities and achieve knowledge distillation on available modalities using GANs. Do *et al.* [14] utilized a cross-modal knowledge distillation method for a visual question answering method.

Knowledge distillation of large pre-trained models. Recently, with the emergence of large pre-trained models, e.g., GPTs [8], llamas [66], SAMs [33] and stable diffusion models [50], many works attempt to adapt the pre-trained models with specific tasks on other modalities to make use of their capacities. Fathullah *et al.* [16] proposed to adapt LLMs to speech recognition. Mohit *et al.* [55] introduced a large pre-trained model for robotic manipulation. Chu *et al.* [10] introduced pre-trained LLMs for recommender systems. [61] adapted language models onto aerial robots. He *et al.* [27] proposed an adapter-based tuning method for the adaptation of large pre-trained LLMs on specific tasks. Mondal *et al.* [43] proposed to adapt pre-trained SAMs with different input transformations for robust object recogni-

tion.

In summary, many delicate and impressive methods have been proposed for knowledge distillation. However, most of them are built on human priors/thoughs, e.g., naturally matching the feature maps from different views or modalities, but with less consideration of the inherent characteristics of neural networks. At the age of Transformer, the network is quite easy to interpret by its inherent attention matrix pattern. Thus, we would further improve those methods via analysis of the inherent demands of neural networks.

2.2. Event-based Vision

Owing to its innate characteristics of high temporal resolution and dynamic range, event-based vision has been a progressively prevalent subject for research in recent years, e.g., frame reconstruction [37, 46, 75, 78, 84], interpolation [31, 59, 67, 68], flow estimation [17, 18, 56, 69, 70, 86], object detection [22, 35, 41, 51] and tracking [62, 73, 81, 82, 88, 89]. Among them, object segmentation using event data is also one of the most popular research topics for robust object recognition [58, 60]. Specifically, Stoffregen *et al.* [58] proposed to utilize the motion information exhibited in event data to achieve motion segmentation. Mitrokhin [42] proposed to utilize event surfaces to achieve visual motion segmentation. Alonso *et al.* [2] utilized a gray-scale Cityscapes dataset to train a network to label event dataset for training an event-based semantic segmentation network. Zhou *et al.* [85] proposed a graph-cut based approach to achieve object motion segmentation on an event dataset. Sun *et al.* [60] introduced a pre-trained E2VID encoder for acquiring rich frame prior for object segmentation. Yang *et al.* [77] proposed an event-based instance segmentation benchmark. Wang *et al.* [71, 72] proposed to apply knowledge distillation from image semantic segmentation and E2VID processes for semantic segmentation.

In summary, although different methods have been proposed to explore and utilize the characteristics of event data, most current research of event cameras is limited to a specific field or dataset, which is quite contradictory with current trends in a universal and unified paradigm for visual recognition with large pre-trained models. However, it's quite effort-consuming to construct large-scale annotated dataset and train the event-based visual perception model from scratch. Thus, utilizing the pre-trained model the image domain to boost performance with event-based vision is a promising approach for building a universal event-based visual perceptual model.

2.3. Object Segmentation & SAMs

In the realm of computer vision, object segmentation is one of cornerstone tasks, focusing on generating pixel-level masks that precisely delineate targeted objects within images [3, 20, 21]. This intricate process is traditionally segmented into several sub-disciplines based on the nature of mask annotations. Semantic segmentation, as outlined in foundational works [39, 44, 79], concentrates on classifying each pixel into a fixed set of categories without differentiating between individual object instances. Conversely, instance segmentation [6, 34, 38], not only categorizes pixels but also separates different instances of the same category. Panoptic segmentation, a term coined and explored in studies [9, 32, 76], amalgamates the principles of semantic and instance segmentation to provide a holistic view of scene parsing, distinguishing countable objects.

Recently, a breakthrough in this domain is the development of SAMs [33]. SAMs stand out for their universal object segmentation capability, which is underpinned by a formidable zero-shot generalization ability. This means SAMs can effectively segment objects in images they have never encountered during training. Moreover, SAMs are designed to adapt at interpreting a variety of input prompts, including points, bounding boxes, or free-form texts.

3. Proposed Method

The cornerstone of leveraging the capabilities of pre-trained expansive vision models, e.g., SAMs, lies in the effective transfer and distillation of knowledge through aligning high-level embeddings across both event and RGB domains. This endeavor initiates an in-depth analysis of the underlying neural network architecture, which is fundamental to the process of object segmentation. SAMs comprise several components: a feature embedding network, or encoder, designed to extract and process complex visual information, coupled with a streamlined object information regression network, or decoder, aimed at precise pixel-level object classification. To delve into the intricacies of this challenge, it is imperative first to establish a formulation for the feature embedding process as

$$\mathcal{F}(\mathbf{W}_d, X_d^{in}) = \mathbf{X}_d^{out}, \tag{1}$$

where $d \in \{M, E\}$ indicates the modality of image (M) or event data (E), $\mathcal{F}(\mathbf{W}_M, \cdot)$ (resp, $\mathcal{F}(\mathbf{W}_E, \cdot)$) indicates the network with the weights $\mathbf{W}_M$ (resp, $\mathbf{W}_M$) for feature embedding on RGB images (resp, events), $\mathbf{X}_d^{out} = \{\mathbf{X}_d^{(s_1)}, \cdots, \mathbf{X}_d^{(s_n)}\}$ represents a set of multi-layer token embeddings. Thus, this knowledge distillation task is to minimize the distribution gap between $\mathbf{X}_M^{out}$ and $\mathbf{X}_E^{out}$. Moreover, due to the fact that the embedding backbone, i.e., ViT, is a plain architecture and only the last layer feature map is fed for object segmentation, the similarity between embedding $\mathbf{X}_M^{(s_n)}$ and $\mathbf{X}_E^{(s_n)}$ is most pivotal issue.

To adapt the embedding in the event domain and minimize the discrepancy between the RGB and event, we propose the following two regularization strategies.

1) We first feed mixed tokens from both modalities to facilitate network training and then extract multi-layer feature

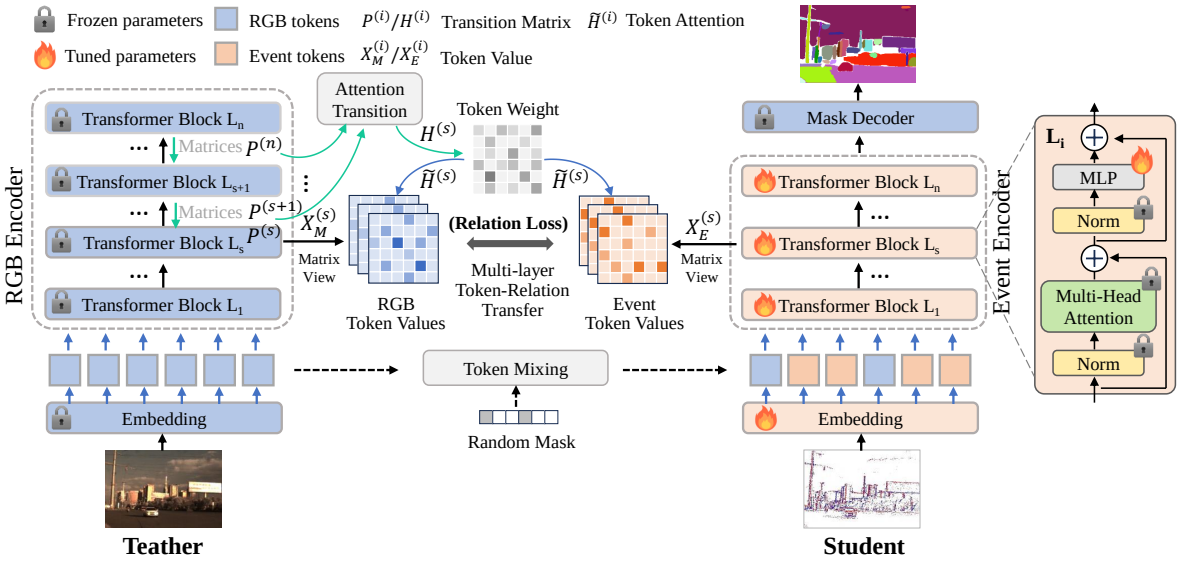

Figure 2. Training workflow for knowledge distillation from pre-trained SAMs to the event domain. We employ the original SAMs as a teacher network to derive meaningful semantic features from RGB tokens. Subsequently, we align multi-layer event token embeddings with image embeddings to enable knowledge distillation. Here, we adopt the 3-channel time-voxel image as input, the polarity image only for visualization. The training is facilitated by simultaneously inputting mixed event and RGB tokens into the student-SAMs (Only event data in the testing phase). To preserve knowledge integrity, we frozen the most of original neural network weights, focusing fine-tuning efforts on the final MLPs in each block. Lastly, we introduce a novel attention-aware embedding weighting strategy to efficiently regularize intermediate-layer token embeddings and significantly improve distillation effectiveness.

maps to fine-tune the model to compensate for the large distribution gaps between the RGB and event images.

2) Due to the distinctive sensing patterns of event and RGB cameras, there are inherent differences in the information captured by image and event data. **While we aim to minimize the distribution gaps between these modalities, it is important to acknowledge that completely eliminating such discrepancies is fundamentally impossible.** Therefore, it is paramount to ascertain the relative importance of individual token embeddings and apply regularizations based on their significance. Previous methods [54, 83] have demonstrated that network gradients can serve as indicators of the network's focus. However, calculating the gradient on teacher network would bring an inevitable huge computational burden. Considering the large size of model, it would make the situation even worse. Fortunately, given the inherent self-attention mechanism of transformers, the focus of network is explicitly encoded as the magnitude of attention values. Thus, we further quantify such a correlation as the self-attention matrix. Then, a weighted distillation algorithm is proposed to focus the training on the pivotal token embeddings.

3.1. Cross-modal Distillation with Mixed Inputs

As shown in Fig. 2, during training, we adopt following two simple yet effective ways to minimize the distribution gap between $\mathbf{X}_M^{out}$ and $\mathbf{X}_E^{out}$ and facilitate network training.

Modification of Inputs. Inspired by current success of mask modeling [4, 26], we randomly replace a small number of event embeddings with images. We expect that it would help retain the original knowledge and facilitate network training.

Trainable Weights. Moreover, motivated by the fact that the strong generalization and zero-shot ability of SAMs may stem from their large-scale training dataset (which contains $11M$ images and 1 $billion$ masks), we do not change the model structure to reuse the pre-trained model. Furthermore, instead of retraining the whole model, we only fine-tune the several MLPs in network to avoid over-fitting and preserve the intrinsically learned patterns.

Based on the aforementioned techniques, a naïve implementation of event-centric SAM could be trained. In the next section, we will discuss a precise embedding regularization via weighting by intermediate correlations.

3.2. Correlation-aware Weighted Token Distillation

Initiating from an elemental Vision Transformer (ViT) layer, we quantify the salience of each embedding by treating it as the fundamental unit of message propagation and focusing on the most important inter-embedding information flow. As depicted in the referenced Fig. 3, we detailedly illustrate the self-attention operation and residuals aggregation in a ViT layer, which facilitate information transition at the inter-embedding level.

We denote $\mathbf{P}^{(i)} \in \mathbb{R}^{k_i \times k_{i+1}}$ as a transition matrix, averaging of multi-head attention on the head dimension, from the i^{th} layer embeddings $\mathbf{X}^{(i)} \in \mathbb{R}^{k_i \times c}$ to its subsequent layer embeddings $\mathbf{X}^{(i+1)} \in \mathbb{R}^{k_{i+1} \times c}$, alongside an iden-

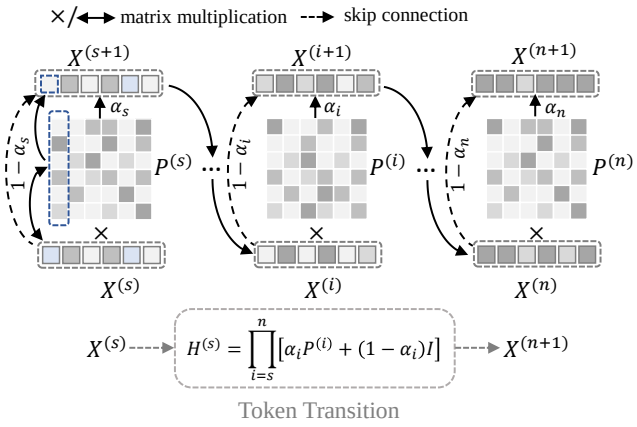

Figure 3. The flow of information from specific layer token embeddings $\mathbf{X}^{(s)}$ to higher level token embeddings $\mathbf{X}^{(n+1)}$ in self-attention layers, with gray shades of attention ($\mathbf{P}^{(\mathbf{i})}$) representing higher values and more intensified information transfer. The illustration reveals that across multiple self-attention layers, information originating from a minimal yet crucial set of token embeddings (indicated by the two gray ones on the left inputs $\mathbf{X}^{(s)}$) can significantly influence the embeddings in deeper layers $\mathbf{X}^{(n+1)}$. Consequently, by regulating the features in the initial layers, our approach aims to direct the network's focus towards reducing the disparity between these pivotal embeddings to narrow the gap of embeddings from different modalities.

tity matrix $\mathbf{I} \in \mathbb{R}^{k_i \times k_{i+1}}$ representing residual aggregation ($k_i \equiv k_{i+1}$ for a plain ViT). The multi-layer inter-embedding transition is then mathematically formulated as

$$\mathbf{H}^{(s)} = \prod_{i=s}^{n} [\alpha_i \mathbf{P}^{(i)} + (1-\alpha_i)\mathbf{I}], \tag{2}$$

where $\mathbf{H}^{(s)} \in \mathbb{R}^{k_s \times k_{n+1}}$ symbolizes a comprehensive information transition matrix from a specific s^{th} layer to the terminal n^{th} layer. Moreover, the scalar α_i signifies the scaling influence of the normalization layers and MLPs. This results in an alteration of the information ratio coming from the attention mechanism and the residual connections. Due to the fact that it's quite time-consuming and even intractable to properly estimate all the α_i, we propose to simplify it by the assumption that the embeddings transition process of multi-layer self-attention mechanism is actually a Markov Chain with a stationary distribution [45]. Thus, we could approximate $\mathbf{H}^{(s)} \approx \widehat{\mathbf{H}}^{(s)} = \beta \prod_{i=s}^{n} \mathbf{P}^{(i)} + (1-\beta)\mathbf{I}$ (Please refer to the supplementary for more analysis). The token-specific transformer's interests/attention from the s^{th} layer, denoted as $\widetilde{\mathbf{H}}^{(s)} \in \mathbb{R}^{k_s}$, is further gauged by multiplying the transition matrix $\mathbf{H}^{(s)}$ with the significance of the final layer's output $\mathbf{e}^T$, formulated as

$$\widetilde{\mathbf{H}}^{(s)} = \widehat{\mathbf{H}}^{(s)} \times \mathbf{e}^T, \tag{3}$$

where we empirically set $\mathbf{e}$ as a vector of ones, underpinning the premise that the embeddings in the last layer uniformly contribute towards the regression of pixel-level object categories. *We calculate the $\widetilde{\mathbf{H}}^{(s)}$ by the attention matrices from the teacher. Please refer to the 2^{nd} ablation study in Sec. 4.2 and Table 4.*

Statement of novelty. Note that some significant advancements have been made in the field of event-based segmentation [60, 71, 72]. They usually take advantage of image knowledge from an E2VID encoder for object segmentation. In the era of large-scale models, leveraging the original architecture and weights is optimal to maintain the robust zero-shot generalization capabilities inherent in pre-trained backbones [28]. Affinity graph distillation [72] presents a compelling approach to guide student networks in acquiring structurally analogous knowledge from their teacher counterparts. The proposed approach, however, emphasizes intensifying the regulation of pivotal proceeding tokens to minimize the distribution gap of feature maps at the terminal layers, implying that the network might permit and manage minor discrepancies in feature maps in less critical regions. Moreover, we avoid the back-propagation of gradients in teacher network by utilizing the off-the-shelf attention matrix in the forward process to approximate the cross-layer embedding correlations, thereby streamlining the process without sacrificing learning efficacy.

3.3. Training Objectives

Combining the aforementioned multi-layer training strategy with the correlation-aware weighting scheme of pivotal token embeddings, we train the network with the following objective function as

$$\mathcal{L} = \sum_{s_i \in S} \gamma_i \mathcal{L}_{s_i}, \quad \mathcal{L}_{s_i} = \|\widetilde{\mathbf{H}}^{(s_i)} \times (\mathbf{X}_M^{(s_i)} - \mathbf{X}_E^{(s_i)})\|_1, \tag{4}$$

where γ_i denotes a specific weight scalar for s_i^{th} layer embeddings, we broadcast $\widetilde{\mathbf{H}}^{(s_i)}$ to fit the channel dimension of embeddings. Empirically, we regularize four layers of embeddings with a corresponding depth of $\{s_i\} = \{0, 3, 6, 9, 12\}$, where 0 corresponds to the resulting token embeddings of the patch embedding layer. We do not weight the regularization on the patch embedding layer to learn a robust embedding process. In our experiment, we empirically set β as 0.5 and γ_i as $\{0.1, 0.4, 0.7, 1.0\}$ for these four layers. Note that the segment results of SAM may be dependent upon the distinct prompts employed within the decoder module. Thus, inspired by [80], we only close the discrepancies between the embeddings to enable an efficient adaptation process and preserve the inherent generalization ability of the decoder.

4. Experiment

Datasets. We collected a large-scale RGB-Event dataset, from current available pixel-level aligned datasets, *i.e.*, VisEvent [73] and COESOT [62]. Both of them were captured by the DAVIS346 camera [7]. The DAVIS346 equips

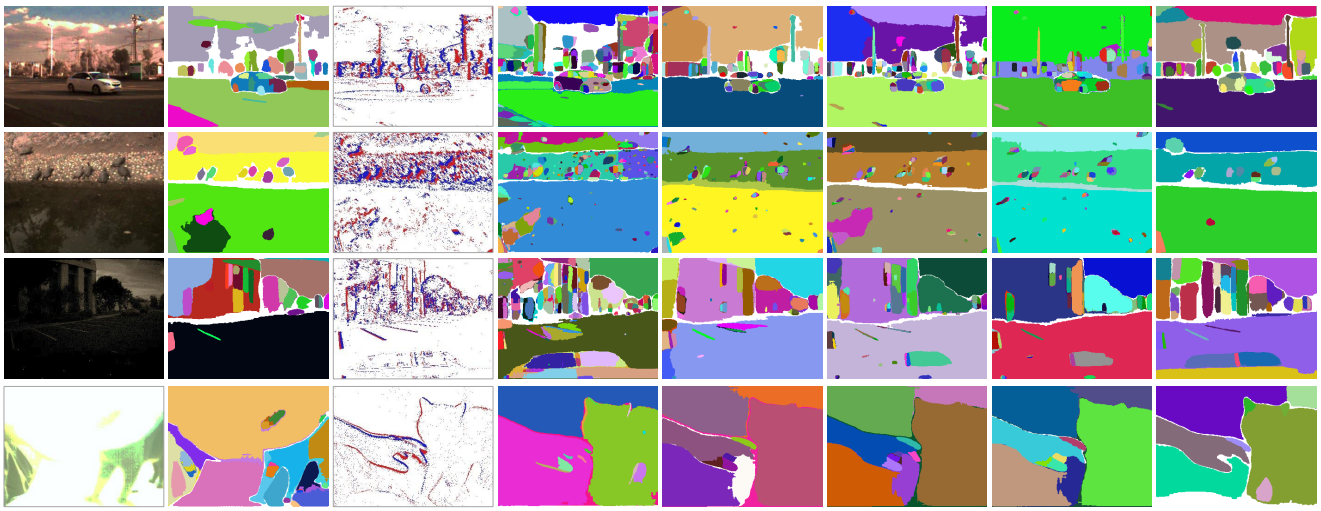

Figure 4. Visual comparison of different methods on RGBE-SEG (the first two rows), MVSEC (the third row) and the degradation scene (the last row) . (a) RGB image, (b) SAM with image, (c) Event data, (d) E2VID [48] with event data, (e) NGA [30] with event data, (f) DTL [71] with event data, (g) ESS [60] with event data, and (h) Our method with event data.

Table 1. Comparison of the segmentation performance between our method and other representative approaches based on RGBE-SEG dataset. For all metrics, the higher, the better. The best and second-best results under each metric are highlighted in **bold** and underlined, respectively. Note that the mIOU is the most essential metric and mP (mR) cannot comprehensively evaluate the performance.

Method	Easy				Medium				Hard				All			
	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU
SAM [33]	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 [48]	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 [75]	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 [30]	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 [12]	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 [71]	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 [72]	**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 [60]	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
Ours	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**

a 346×260 pixels active sensor and a dynamic vision sensor, which can acquire aligned images and events. These samples consisted of diverse indoor and outdoor scenarios. For effective knowledge transfer, we only collected image-event pairs without degradation (e.g., no low dynamic range and motion blur) and removed the duplicate samples. After integrating and cleaning, we extended the two datasets to a segmentation dataset, **RGBE-SEG**. The RGBE-SEG included 65,957 image-event pairs, 64,957 for training and 1,000 for testing. The test set contained 38,760 masks, and we artificially divided it into easy, medium, and hard subsets based on the complexity of scenarios. All ground truth masks were generated by images and the well-trained SAM [33]. To further explore the zero-shot performance of our method, we showed more evaluation results on **MVSEC** dataset (500 image-event pairs each in "indoor_flying1" and "outdoor_day2" sequences) [87], containing 54,600 masks.

Evaluation Metrics. For each instance-level ground truth mask, we define the predicted mask with the maximum intersection over union as its matched mask. This matching splits the pair masks into three sets: true positives (TP), false positives (FP), and false negatives (FN), representing matched segments, unmatched predicted segments, and unmatched ground truth segments, respectively. Based on the split segments, we calculate the precision (P), recall (R) and intersection over union (IoU) to evaluate the segmentation quality. We calculate these metrics for each instance independently and average over instances, obtaining mP, mR, and mIoU. In addition, we introduce the area-weighted intersection-over-union (aIoU). mIoU focuses on the details of scenarios and gives equal treatment to each instance. Instead, aIoU cares about the overall mask quality. These four metrics collaborate to describe the segmentation quality.

Implementation details. We set the input image and event

Table 2. Comparison of the segmentation performance between our method and other representative approaches based on MVSEC dataset.

Method	Indoor				Outdoor			
	mP	mR	mIoU	aIoU	mP	mR	mIoU	aIoU
SAM [33]	0.38	0.66	0.24	0.41	0.41	0.71	0.28	0.51
E2VID [48]	0.48	0.68	0.32	0.46	0.64	0.59	0.38	0.52
ETNet [75]	0.54	0.60	0.36	0.43	**0.71**	0.52	0.38	0.46
NGA [30]	0.46	0.72	0.32	0.49	0.41	**0.73**	0.31	0.50
LFI [12]	0.44	0.74	0.31	0.49	0.48	0.70	0.32	0.54
DTL [71]	0.44	0.74	0.32	0.49	0.52	0.68	0.34	**0.55**
EVDistill [72]	**0.61**	0.47	0.29	0.34	0.64	0.47	0.29	0.38
ESS [60]	0.42	**0.76**	0.31	0.49	0.46	0.72	0.32	0.53
Ours	0.53	0.70	**0.38**	**0.49**	0.64	0.67	**0.42**	**0.55**

frame sizes to 512×512. We directly adopted the base vision transformer (ViT-B) [15] with pre-trained weights as our dual-modal encoders. During training, we used the Adam optimizer to train the event encoder for 5 epochs, 13,500 iterations, with a batch size of 24. We set the initial learning rate as 2e-4. The learning rate is adjusted by a decay scheduler, which is scaled by 0.9 on the 4^{th} epoch. All networks are implemented in PyTorch and run on a computer equipped with 4 GPUs (GeForce RTX 3090).

Comparing methods. To fully investigate the capacity of the proposed method, we investigate the performance of different SOTA event-based segmentation methods. Note that the SAM could predict much more precise masks than other segmentation models. Thus, it's unfair to directly adopt those pre-trained models for comparison. Instead, we retrain different distillation and adaptation methods to adapt SAMs. Generally, they could be divided into three categories. The first directly applied frame reconstruction algorithm, e.g., E2VID [48] and ETNet [75], on event data to get a pseudo-frame, then directly fed such a frame into SAM without pre-training. The second utilized pre-trained E2VID as encoder to close the gap between event and RGB domain and then feed those features into the segmentation network, e.g., EVDistill [72], DTL [71] and ESS [60]. Finally, there are knowledge distillation-based methods which do not explicitly utilize the pre-trained decoder by trying to transfer knowledge from image backbone onto event data, e.g., NGA [30], LFI [12].

4.1. Experimental Results

RGBE-SEG. In the comparative analysis presented in Table 1, our method demonstrates a notable enhancement in performance, yielding a 9% improvement in both mIOU and aIoU over existing methodologies. While certain approaches like EvDistill exhibit higher mP, they fall short in terms of mR compared to our method. Intriguingly, the original SAM, without any fine-tuning, attains remarkably competitive results, particularly in simpler scenarios. This observation underscores the potential detriment to performance when large pre-trained models are not appropriately adapted. Among the various methods evaluated, E2VID emerges as the most effective, underscoring the efficacy of transitioning from event to video data in enhancing event-based object segmentation. Contrarily, methods employing the E2VID encoder experience a significant performance decline, likely due to modifications in the original SAM architecture which adversely affect model generalization. A visual comparison of these methods, illustrated in the first row of Fig. 4, further validates our findings. Our method distinctly differentiates between sky and cloud, and offers clearer segmentation of ground and vehicles compared to others. Notably, it even seems to surpass the SAM with image data in certain aspects, reinforcing the viability of using event data for object segmentation.

Table 3. Ablation study results of the fine-tuning layers based on RGBE-SEG dataset. All models are evaluated with event frame input. "w/o Fine-tuning" indicates the original SAM without any retraining, "Embed" denotes the embedding layers, "Four MLPs" represents the last MLPs from $\{3, 6, 9, 12\}$ blocks , "Four Blocks" also indicates those blocks, "All MLPs" represents the last MLPs from all the transformer layers, "All Blocks" denotes all the blocks contained in ViT layers. Underline for the adopt strategy (**our baseline**). Note that all methods are w/o weighted regularization.

Fine-tuning Layers	Trainable #Param	mP	mR	mIoU	aIoU
w/o Fine-tuning	-	0.39	0.73	0.26	0.43
Embed	0.6M (0.7%)	0.46	0.73	0.32	0.47
Embed + Four MLPs	19.5M (22.3%)	0.52	0.73	0.37	0.53
Embed + Four Blocks	29.0M (33.2%)	0.51	0.68	0.32	0.47
Embed + All MLPs	57.3M (65.7%)	0.48	0.69	0.30	0.45
Embed + All Blocks	87.3M (100%)	0.51	0.63	0.29	0.41

MVSEC. The segmentation results of MVSEC are presented in Table 2. Our method continues to outperform competing techniques, evidencing robust zero-shot capabilities. It is observed that other event-based segmentation methods, particularly those employing frame reconstruction, exhibited quite improvements. This enhancement may be attributed to the high similarity of the scenes in the dataset, which simplifies the reconstruction process. Furthermore, as depicted in the third row of Fig. 4, the proposed method effectively distinguishes the lane line and car from background. This is further validated by the consistent results obtained when SAM is applied to the image data, thereby demonstrating the successful adaptation of SAM on the event-based object segmentation task.

4.2. Ablation Study

Multi-layer Regularization. As aforementioned, to manage the strong zero-shot ability of SAM in such a cross-modal distillation task, maintaining the original architecture and weights is one of the keys. To verify this point, we conduct ablation studies, showing network performance with re-training different numbers of parameters. As shown in Table 3, we gradually increase the number of fine-tuning

Table 4. Ablation study results of the mixing tokens and weighted adaptation schemes based on RGBE-SEG dataset, where the first row is a baseline finetuned with only multi-layer feature alignment as 3^{rd} row of Table 3, "Token Mixing" means to mix RGB image tokens with event tokens for facilitating adaptation, "$\widetilde{\mathbf{H}}_E^{(s)}$" indicates to calculate embedding weights via aggregating event (student) attention matrices, "$\widetilde{\mathbf{H}}_M^{(s)}$" indicates to calculate embedding weights via aggregating image (teacher) attention matrices, "single $\widetilde{\mathbf{H}}_M^{(s)}$" represents to only utilize one layer attention matrix to calculate regularization weights.

No.	Token Mixing	$\widetilde{\mathbf{H}}_E^{(s)}$	$\widetilde{\mathbf{H}}_M^{(s)}$	single $\widetilde{\mathbf{H}}_M^{(s)}$	mP	mR	mIoU	aIoU
a.	×	×	×	×	0.52	0.73	0.37	0.53
b.	✓	×	×	×	0.53	0.74	0.38	0.54
c.	×	×	✓	×	0.58	0.70	0.40	0.55
d.	✓	✓	×	×	0.41	0.77	0.30	0.49
e.	✓	×	×	✓	0.52	0.74	0.37	0.54
f.	✓	×	✓	×	0.59	0.71	0.41	0.55

layers in SAM. We could observe a incremental performance improvement at the beginning stage, which indicates that fine-tuning some layers could indeed improve the domain adaptation ability. However, with the retraining parameters continuing to increase, the generalization ability of adapted SAM tends to gradually decline.

Attention-aware Weighted Adaptation Scheme. We approximate the significance of each intermediate embedding by aggregating the attention matrix. To evaluate the effectiveness of such a regularization scheme, we further conduct extensive experiments, as shown in Table 4. Through comparing **a.** and **c.** (or **b.** and **f.**), we could figure out that applying weighted distillation could significantly improve the mIOU with more than 3% and mP of 6%, indicating the effectiveness of the proposed regularization term. The visual comparison of Fig. 5 further underscores the effectiveness of the proposed strategy. However, such improvement only appears when regularization is derived from the attention matrix of the teacher-transformer, denoted as $\widetilde{\mathbf{H}}_M^{(s)}$. Conversely, weighted regularization stemming from $\widetilde{\mathbf{H}}_E^{(s)}$ may lead to a deterioration in network performance, evident in **d)** and **b)** 7% of mIOU, resulting in a 7% reduction in mIOU. It indicates that weighted cross-modal distillation could indeed play a significant role in the model training. Furthermore, our observations indicate that attention from the student network inadequately captures token embedding significance, thereby resulting in suboptimal performance within this modality.

Token Mixing. As shown in Table 4 **a.** and **b.** (**c.** and **f.**), through mixing image and event tokens, proposed method achieves a moderate improvement of 1% consistently on different metrics. Although such improvement is not significant compared with the main contribution of the weighted adaptation scheme, it's still important to help the proposed method manage a leading performance in some datasets, e.g., Table 2 MVSEC Indoor.

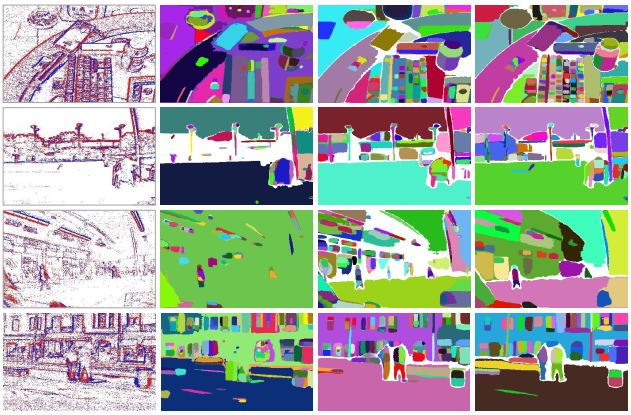

Figure 5. Visual ablation comparison of our designs odawdsda-dataset. Each row represents that (a) Event data, (b) Original SAM, (c) Our baseline, and (d) Our method with token mixing and weighted adaptation scheme.

5. Conclusions and Discussions

In this paper, we have presented a cross-modal adaptation method of SAM for event-centric vision, characterized by an innovative weighted distillation approach. This methodology is inspired by the inherent self-attention mechanism, strategically formulated to reduce discrepancies among key token embeddings, thereby enhancing the adaptation of embeddings from higher-level. Extensive experimental results qualitatively and quantitatively validate the effectiveness of the proposed method. Notwithstanding these significant strides, the study identifies critical areas for future enhancement. To further advance segmentation accuracy, we may need a more voluminous and diverse dataset to refine the adaptation process of the event-based SAM. In addition, exploring the synergy of this specialized SAM with Large Language Models (LLMs) could pioneer new frontiers in achieving seamless integration of vision and language in the context of event-based vision systems. Moreover, the potential application of this adapted SAM extends to other complex tasks such as object tracking and detection. It is imperative to acknowledge that the substantial size of SAMs presents a challenge, highlighting the need for optimization and streamlining of these networks to ensure more efficient adaptation and deployment in diverse scenarios.

Limitations. Due to the sensing characteristics of event camera, failure cases may occur in the texture-less region, e.g., the sky. Meanwhile, according to sensor development, there are some limitations, e.g., auto-focus, resolution, and noise of the sensor.

Acknowledgements. This work is supported by Fundamental Research Funds for the Central Universities (QTZX23038), CityU Strategic Research Grant (7005990) and Innovation and Technology Fund (MHP/117/21).

References

[1] Saeed Afshar, Andrew Peter Nicholson, Andre Van Schaik, and Gregory Cohen. Event-based object detection and tracking for space situational awareness. *IEEE Sensors Journal*, 20(24):15117–15132, 2020. 2

[2] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *CVPRW*, pages 0–0, 2019. 2, 3

[3] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2021. 3

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, pages 348–367. Springer, 2022. 4

[5] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *AVSS*, pages 1–6. IEEE, 2017. 2

[6] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, pages 9157–9166, 2019. 3

[7] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 5

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NIPS*, 33:1877–1901, 2020. 2

[9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, pages 12475–12485, 2020. 3

[10] Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, et al. Leveraging large language models for pre-trained recommender systems. *arXiv preprint arXiv:2308.10837*, 2023. 2

[11] Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IRAL*, 5(3):4596–4603, 2020. 2

[12] Yongjian Deng, Hao Chen, Huiying Chen, and Youfu Li. Learning from images: A distillation learning framework for event cameras. *IEEE TIP*, 30:4919–4931, 2021. 6, 7

[13] Rika Sugimoto Dimitrova, Mathias Gehrig, Dario Brescianini, and Davide Scaramuzza. Towards low-latency high-bandwidth control of quadrotors using event cameras. In *ICRA*, pages 4294–4300. IEEE, 2020. 1

[14] Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D Tran. Compact trilinear interaction for visual question answering. In *ICCV*, pages 392–401, 2019. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[16] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. Prompting large language models with speech recognition abilities. *arXiv preprint arXiv:2307.11795*, 2023. 2

[17] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In *CVPR*, pages 12280–12289, 2019. 3

[18] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE TPAMI*, 44(1):154–180, 2020. 1, 3

[19] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *ECCV*, pages 103–118, 2018. 2

[20] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017. 3

[21] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, 70:41–65, 2018. 3

[22] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *CVPR*, pages 13884–13893, 2023. 2, 3

[23] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129: 1789–1819, 2021. 2

[24] Jianping Gou, Liyuan Sun, Baosheng Yu, Shaohua Wan, and Dacheng Tao. Hierarchical multi-attention transfer for knowledge distillation. *ACM TMC*, 20(2):1–20, 2023. 2

[25] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836, 2016. 2

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 4

[27] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*, 2021. 2

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[29] Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *CVPR*, pages 3123–3132, 2020. 2

[30] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. Learning to exploit multiple vision modalities by using grafted networks. In *ECCV*, pages 85–101. Springer, 2020. 6, 7

[31] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *CVPR*, pages 18032–18042, 2023. 3

[32] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 3

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 6, 7

[34] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, pages 13906–13915, 2020. 3

[35] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE TIP*, 31:2975–2987, 2022. 2, 3

[36] Lei Li, Alexander Liniger, Mario Millhaeusler, Vagia Tsiminaki, Yuanyou Li, and Dengxin Dai. Object-centric cross-modal feature distillation for event-based object detection. *arXiv preprint arXiv:2311.05494*, 2023. 2

[37] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *ICCV*, pages 10615–10625, 2023. 3

[38] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018. 3

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 3

[40] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. Multi-bracket high dynamic range imaging with event cameras. In *CVPR*, pages 547–557, 2022. 1

[41] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *IROS*, pages 1–9. IEEE, 2018. 2, 3

[42] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In *CVPR*, pages 14414–14423, 2020. 3

[43] Arnab Kumar Mondal, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Sai Rajeswar, and Siamak Ravanbakhsh. Equivariant adaptation of large pre-trained models. *arXiv preprint arXiv:2310.01647*, 2023. 2

[44] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015. 3

[45] CC Paige, George PH Styan, and Peter G Wachter. Computation of the stationary distribution of a markov chain. *Journal of Statistical Computation and Simulation*, 4(3):173–186, 1975. 5

[46] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE TPAMI*, 44(5): 2519–2533, 2020. 3

[47] Bharath Ramesh, Andrés Ussa, Luca Della Vedova, Hong Yang, and Garrick Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in neuroscience*, 14:135, 2020. 1

[48] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 43(6):1964–1980, 2019. 1, 6, 7

[49] Siddharth Roheda, Benjamin S Riggan, Hamid Krim, and Liyi Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In *ICASSP*, pages 2926–2930. IEEE, 2018. 2

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[51] Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *CVPR*, pages 12371–12381, 2022. 3

[52] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *WACV*, pages 156–163, 2020. 1

[53] Ionut Schiopu and Radu Ciprian Bilcu. Entropy coding-based lossless compression of asynchronous event sequences. In *CVPR*, pages 3922–3929, 2023. 1

[54] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 4

[55] Mohit Sharma, Claudio Fantacci, Yuxiang Zhou, Skanda Koppula, Nicolas Heess, Jon Scholz, and Yusuf Aytar. Lossless adaptation of pretrained vision models for robotic manipulation. *arXiv preprint arXiv:2304.06600*, 2023. 2

[56] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *ECCV*, pages 628–645. Springer, 2022. 3

[57] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *CVPR*, pages 1731–1740, 2018. 2

[58] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *ICCV*, pages 7244–7253, 2019. 3

[59] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *CVPR*, pages 18043–18052, 2023. 3

[60] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *ECCV*, pages 341–357. Springer, 2022. 2, 3, 5, 6, 7

[61] Andrea Tagliabue, Kota Kondo, Tong Zhao, Mason Peterson, Claudius T Tewari, and JONATHAN P HOW. Real: Resilience and adaptation using large language models on

autonomous aerial robots. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023. 2

[62] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*, 2022. 2, 3, 5

[63] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *ICIP*, pages 6–10. IEEE, 2019. 2

[64] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 2

[65] Abhishek Tomy, Anshul Paigwar, Khushdeep S Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *ICRA*, pages 933–939. IEEE, 2022. 2

[66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[67] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *CVPR*, pages 16155–16164, 2021. 3

[68] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *CVPR*, pages 17755–17764, 2022. 3

[69] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE TIP*, 31:7237–7251, 2022. 3

[70] Zhexiong Wan, Yuxin Mao, Jing Zhang, and Yuchao Dai. Rpeflow: Multimodal fusion of rgb-pointcloud-event for joint optical flow and scene flow estimation. In *ICCV*, pages 10030–10040, 2023. 3

[71] Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *ICCV*, pages 2135–2145, 2021. 3, 5, 6, 7

[72] Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *CVPR*, pages 608–619, 2021. 3, 5, 6, 7

[73] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE TCYB*, 2023. 2, 3, 5

[74] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2

[75] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *ICCV*, pages 2563–2572, 2021. 3, 6, 7

[76] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, pages 8818–8826, 2019. 3

[77] Sifan Yang, Qi Zheng, Xiaowei Hu, and Guijin Wang. Vess: Variable event stream structure for event-based instance segmentation benchmark. In *ICDSP*, pages 112–116, 2020. 3

[78] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *CVPR*, pages 13924–13934, 2023. 3

[79] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, pages 325–341, 2018. 3

[80] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 2, 5

[81] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *ICCV*, pages 13043–13052, 2021. 3

[82] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *CVPR*, pages 8801–8810, 2022. 3

[83] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 4

[84] Chu Zhou, Minggui Teng, Jin Han, Jinxiu Liang, Chao Xu, Gang Cao, and Boxin Shi. Deblurring low-light images with events. *IJCV*, 131(5):1284–1298, 2023. 3

[85] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE TNNLS*, 2021. 3

[86] Alex Zihao Zhu and Liangzhe Yuan. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems*, 2018. 2, 3

[87] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IRAL*, 3(3):2032–2039, 2018. 6

[88] Zhiyu Zhu, Junhui Hou, and Xianqiang Lyu. Learning graph-embedded key-event back-tracing for object tracking in event clouds. *NIPS*, 35:7462–7476, 2022. 3

[89] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In *ICCV*, pages 22045–22055, 2023. 3