

Adaptive Event Stream Slicing for Open-Vocabulary Event-Based Object Detection via Vision-Language Knowledge Distillation

Jinchang Zhang, Zijun Li, Jiakai Lin, Guoyu Lu

Abstract—Event camera offers advantages in object detection tasks for its properties such as high-speed response, low latency, and robustness to motion blur. However, event cameras inherently lack texture and color information, making open-vocabulary detection particularly challenging. Current event-based detection methods are typically trained on predefined target categories, limiting their ability to generalize to novel objects, where encountering previously unseen objects is common. Vision-language models (VLMs) have enabled open-vocabulary object detection in RGB images. However, the modality gap between images and event streams makes it ineffective to directly transfer CLIP to event data, as CLIP was not designed for event streams. To bridge this gap, we propose an event-image knowledge distillation framework, leveraging CLIP’s semantic understanding to achieve open-vocabulary object detection on event data. Instead of training CLIP directly on event streams, we use image frames as teacher model inputs, guiding the event-based student model to learn CLIP’s rich visual representations. Through spatial attention-based distillation, the student network learns meaningful visual features directly from raw event inputs, while inheriting CLIP’s broad visual knowledge. Furthermore, to prevent information loss due to event data segmentation, we design a hybrid Spiking Neural Network (SNN) and Convolutional Neural Network (CNN) framework. Unlike fixed-group event segmentation methods, which often discard crucial temporal information, our SNN adaptively determines the optimal event segmentation moments, ensuring that key temporal features are extracted. The extracted event features are then processed by CNNs for object detection.

I. INTRODUCTION

Event cameras are bio-inspired vision sensors that fundamentally differ from traditional frame-based cameras. They capture event streams asynchronously and sparsely, gaining attention for their superior characteristics, including high temporal resolution, high dynamic range, low latency, and low power consumption. In recent years, leveraging these inherent advantages, event-based vision perception has advanced across various domains, including object tracking, depth estimation, object detection.

As one of the core tasks in event-based perception, event-based object detection has gained significant attention but remains constrained to closed-set settings. Due to the unique imaging modality of event cameras and the lack of large datasets, existing models struggle to generalize to unseen categories in real-world scenarios. Consequently, open vocabulary object detection for event cameras has become a critical challenge. In frame-based vision tasks, pretrained vision-language models (VLMs) such as CLIP [1] have achieved

open vocabulary detection by learning aligned image-text representations from large datasets. However, CLIP or other VLM [2] was designed for regular visible images, instead of event camera streams. This issue primarily stems from the modality gap between event data and conventional 2D images, making CLIP inapplicable to raw event streams. Moreover, the absence of large-scale event-text datasets makes it impractical to train a vision-language model for event cameras from scratch. Another challenge of event cameras lies in slicing of the raw event stream. When using event streams as input, two key steps are required: (1) segmenting the raw event stream into multiple sub-event groups and (2) converting these sub-event groups into different event representations. Current research focuses on optimizing event representations [3], while overlooking the crucial segmentation step. Common segmentation methods [4] employ fixed grouping strategies, such as slicing event streams based on a fixed number of events or a fixed time interval. However, these approaches suffer from information imbalance—potentially leading to information loss in low-speed motion scenarios and excessive redundancy in high-speed motion conditions. Although some recent studies [5] have proposed adaptive sampling methods, with [6] requiring explicit searches over multiple time periods, they still do not directly learn an adaptive segmentation process.

To address the event stream slicing problem, we propose the Adaptive Event Slicing module, which leverages spiking neural networks to dynamically determine the optimal event segmentation timing, overcoming the limitations of traditional fixed slicing strategies. Additionally, to bridge the modality gap between event streams and image frames, which prevents the direct application of pretrained VLMs, we introduce Event-Based Vision-Language Knowledge Distillation. This framework employs the image encoder from VLMs as the teacher network and an event-based object detection model as the student network, enabling knowledge transfer from VLMs to event data. Specifically, during SNN-based event triggering, we adaptively determine the optimal event segmentation timing and extract event stream features for subsequent object detection. To enhance the stability of event segmentation, we introduce the Linear Incremental Constraint Loss, preventing premature spike activations. Additionally, we design the Self-Supervised Feedback Loss (SSF-Loss), which adjusts the membrane potential based on object detection results, guiding the SNN to fire at the optimal time step and adaptively refine the event segmentation strategy. In the object detection module, we employ a category-agnostic proposal module to improve the model’s ability to generalize

Jinchang Zhang, Zijun Li, Jiakai Lin and Guoyu Lu are with the Intelligent Vision and Sensing (IVS) Lab at SUNY Binghamton, USA. guoyulu62@gmail.com

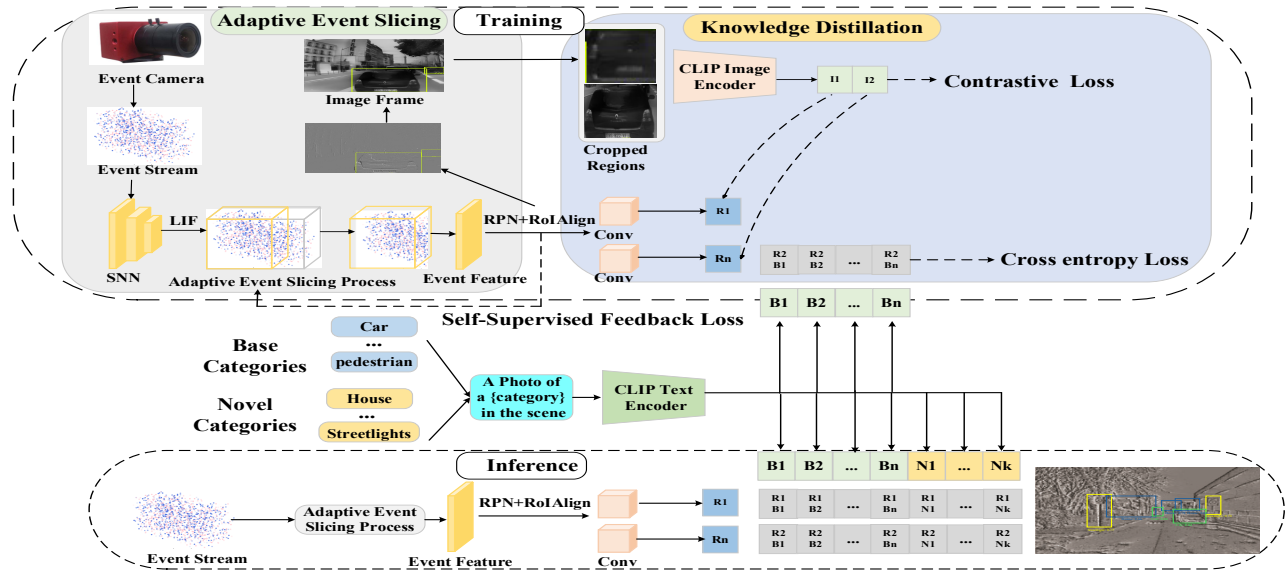


Fig. 1. The overview of our framework. The event stream is first fed into a spiking neural network, where Self-Supervised Feedback Loss is utilized to dynamically adjust the membrane potential based on object detection results, enabling adaptive event segmentation and feature extraction. We transfer image knowledge from CLIP to event data, using the CLIP image encoder as a teacher model. Through knowledge distillation, the student detector trained on event data learns the rich visual representations from CLIP. Additionally, category text is input into the frozen CLIP text encoder to generate text embeddings, and the cosine similarity between each region embedding and all category text embeddings is computed for object classification. During the inference phase, the model performs open-vocabulary object detection using only event stream data, without relying on image frames.

to unseen object categories. To align event-based features with image representations, we extract image features from a frozen CLIP image encoder and perform feature alignment with event stream embeddings from the same ROI region. We adopt CLIP’s image encoder as the teacher model and incorporate knowledge distillation based on a spatial self-attention mechanism, enabling the student detector trained on event data to learn rich visual representations from CLIP. Furthermore, category text embeddings are generated by a frozen CLIP text encoder and integrated with region embeddings for object classification. During the inference phase, we perform open-vocabulary object detection using only event stream data, without relying on image frames.

In summary, our main contributions include the following aspects: 1. To the best of our knowledge, we are the first to introduce an event-based open vocabulary object detection framework, enabling object detection directly based on textual descriptions. 2. We design a knowledge distillation method leveraging CLIP to enhance event-based object detection, effectively transferring rich semantic knowledge to event data. 3. We propose a self-supervised spiking neural network slicing feedback scheme, which dynamically adjusts the membrane potential according to object detection results, enabling adaptive event segmentation and feature extraction. Figure 1 shows our framework.

II. RELATED WORK

A. Event-based Object Detection

In event-based object detection, mainstream approaches can be broadly grouped into two families: CNN-based models and energy-efficient, biologically inspired SNN-based methods. CNN-based methods typically convert event streams into frame-like representations—such as event histograms, time surfaces, and event volumes—so they can be

processed by existing deep learning detectors [7], [8], [9], [10], [11], [12], [13]. These studies also extend detection beyond RGB image object detection, including [14], [15], [16], [17], [18]. However, these frame-based encodings often discard the intrinsic spatiotemporal information of event data, thereby limiting detection performance. To mitigate this, recent work such as SNN-based methods focus on exploiting the sparsity and computational efficiency of event data. Traditional SNN detectors largely rely on ANN-to-SNN conversion, as in Spiking-YOLO [19]. Yet such approaches typically require many timesteps to match ANN performance, which limits real-time applicability [20]. To address this, directly trained SNN detection frameworks have been proposed, including EMS-YOLO and SpikeYOLO [21]. EMS-YOLO introduces an all-spiking residual block (EMS-ResNet) [22], while SpikeYOLO integrates integer-valued LIF (I-LIF) neurons to reduce quantization error [23]. Furthermore, recent advances such as SFOD and CREST further optimize SNN-based detection on event-camera data: SFOD leverages multi-scale feature fusion, and CREST introduces a spiking spatiotemporal IoU loss [23], [24].

B. Open-Vocabulary Object Detection

Open-Vocabulary Object Detection has evolved from zero-shot object detection based on visual attributes to open-vocabulary detection leveraging VLMs [25], progressively enhancing the model’s generalization ability to unseen categories. Early ZSD methods [26] relied on visual attributes for unseen category inference. However, these methods suffer from limited scalability and generalization, making them less effective for large-scale open-vocabulary detection. With the rise of VLMs researchers have explored replacing detector classifiers with text embeddings and employing cross-modal feature alignment to enhance open-vocabulary detection [27].

For instance, ViLD leverages knowledge distillation to inject open-vocabulary knowledge into two-stage detectors [28]. PromptDet [29] enhances vision-text embedding alignment using learnable prompts. Recent research has focused on cross-modal contrastive learning and spatiotemporal feature fusion to adapt event data for VLMs. EventCLIP [30] converts event data into 2D grids and employs an adapter to align event features with CLIP knowledge. E-CLIP [31] adopts Hierarchical Triple Contrastive Alignment, jointly optimizing event, image, and text embeddings. However, open-vocabulary detection models trained on images perform poorly when directly applied to event cameras due to the modality gap. Therefore, for heterogeneous modalities, it is crucial to design and select a suitable distillation scheme for knowledge transfer [32]. To address this issue, we propose a knowledge transfer method to adapt the knowledge learned from image-based models to event-based detection models.

III. ADAPTIVE EVENT STREAM SLICING PROCESS

An event stream is an asynchronous data representation, defined as a set: $E = \{[x_i, y_i, t_i, p_i]\}_{i=1}^N$ with a temporal span of T , i.e., $t_i \in [t_0, t_0 + T]$. To convert the event stream into a format suitable for Spiking Neural Networks (SNNs), we adopt the voxel grid representation method [33]. Given that SNNs naturally align with event stream data, we utilize them as event stream slicers to enable a dynamic event slicing process and enhance object detection performance. For an event stream, slicing is determined by the state of spiking neurons (excited/resting). As dynamic event triggers, spiking neurons execute event slicing upon generating a spike at the current time step n_c , where $S_{out} = 1$ indicates a slicing event. This operation captures the time interval from the last spike to the current one, forming an event group within this interval, which can then be used for object detection tasks.

Membrane Potential Driven Loss: Since the event segmentation position depends on whether SNN generates a spike at a given time step, it is essential to guide the network to precisely trigger a spike at the optimal time step n^* . Specifically, if the segmentation is expected to occur at time n^* , the neuron’s membrane potential should reach the threshold V_{th} at this moment to generate a spike. However, since the membrane potential resets to its resting state immediately after a spike, this process may introduce inaccuracies in guiding subsequent time steps [34]. Therefore, we impose supervision on the non-reset membrane potential $U[n]$ at n^* , ensuring that $U[n^*] \geq V_{th}$. The Membrane Potential Driven Loss (Mem-Loss) as follows:

$$\mathcal{L}_{Mem} = \|U[n^*] - (1 + \alpha)V_{th}\|_2^2, \quad (1)$$

where $\alpha \geq 0$ is a hyperparameter that controls the extent to which the expected membrane potential surpasses the threshold. This loss function effectively guides the spiking neuron to fire at the target time step during training, thereby enabling precise segmentation of the event stream.

Linear Incremental Constraint Loss: Due to the hll effect [34], if the membrane potential at a later time step is guided above the threshold, an earlier time step may trigger a spike first, causing the neuron to enter a resting state and thereby affecting the accuracy of event segmentation. Even

with the introduction of the membrane potential-driven loss (Mem-Loss), premature triggering of segmentation time steps may still occur. To ensure stable spike triggering, we expect the membrane potential at later time steps to monotonically increase, reducing the impact of premature activation and improving the robustness of segmentation time steps. To this end, we propose a **Linear Incremental Constraint Loss** to ensure that the membrane potential at the target time step n^* precisely reaches the threshold. We establish the following linear growth assumption:

$$U[n_c] \geq U[n^*] \cdot \left(\frac{n_c}{n^*}\right)^\beta, \quad (2)$$

where n^* represents the target spike time step, n_c denotes the current spike time step, the exponential factor β controls the rate of membrane potential growth. This constraint ensures that the membrane potential maintains an increasing trend across the entire time domain, effectively preventing excessive early activation from suppressing spikes at later time steps, thereby enhancing the temporal consistency of segmentation time steps. Based on the above constraint, we define the **Linear Incremental Constraint Loss** as follows:

$$\mathcal{L}_{LA} = \begin{cases} \|U[n_c] - V_{th} \cdot \left(\frac{n_c}{n^*}\right)^\beta\|^2, & \text{if } U[n_c] \geq U[n^*] \cdot \left(\frac{n_c}{n^*}\right)^\beta; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

This loss is designed to ensure that the membrane potential remains monotonically increasing at time step n , preventing premature spike triggering and enhancing temporal consistency. Additionally, it precisely controls the membrane potential at the target time step n^* to reach the threshold V_{th} , optimizing spike triggering accuracy. The introduction of the exponential factor β allows for adjustable membrane potential growth rates, enabling the model to adapt to different event distributions.

Self-Supervised Feedback Loss: To enhance the event slicing performance of SNN in object detection tasks, we propose a Self-Supervised Feedback Loss (SSF-Loss). This loss enables SNN’s event slicing points to adaptively optimize the downstream object detection performance, rather than relying solely on a fixed loss function for optimization. Specifically, we introduce the object detection task loss $L_M(n)$ as a feedback signal to directly guide the slicing strategy of the SNN. The loss function is defined as follows:

$$L_{SSF} = \sum_n (L_M(n) \cdot |U[n] - V_{th}|) \quad (4)$$

where $U[n]$ represents the membrane potential of the SNN at time step n , determining whether a spike is triggered at that step; V_{th} is the threshold for spike firing; and $L_M(n)$ represents the loss of the downstream object detection task, reflecting whether the current event slicing aids detection. If a spike at time step n increases object detection loss, the optimization adjusts the membrane potential $U[n]$ to lower slicing probability at that step, shifting it toward a more optimal timing. Conversely, if $L_M(n)$ is low, indicating beneficial slicing, the SNN favors spiking at that step for better event utilization. This self-supervised mechanism enables the SNN to refine event slicing over time, extracting key events more effectively and enhancing detection performance in dynamic scenes.

IV. OPEN-VOCABULARY EVENT OBJECT DETECTION

A. Localization For Novel Categories

The core challenge of open-vocabulary object detection lies in accurately localizing unseen object categories. To address this, we enhance the standard two-stage object detector (Mask R-CNN [35]) by introducing a self-supervised slicing SNN feature extraction network to replace the traditional feature extraction module, making it more suitable for handling the sparsity and asynchronicity of event streams. This design effectively captures the dynamic characteristics of event data, improving the detector’s performance on event-based inputs. To further enhance open-vocabulary object detection capabilities, we adopt a category-agnostic design, where object detection is performed based on visual features and region proposals rather than predefined category labels. Within this framework, we optimize bounding box regression and mask prediction, replacing the category-specific bounding box regression and mask prediction layers with a category-agnostic proposal module. This module predicts a single generic bounding box and mask for each region of interest (RoI) rather than generating separate predictions for each category. By transitioning from category-specific predictions to category-agnostic predictions, the model significantly improves its generalization ability to unseen object categories, making it more suitable for open-vocabulary detection tasks.

Category-Agnostic Bounding Box Regression: In category-agnostic bounding box regression, we remove category-specific bounding box prediction, allowing all objects to share a single set of bounding box regression parameters, formulated as:

$$\mathcal{L}_{box}^{CA} = \sum_i 1_{\{y_i > 0\}} \sum_{j \in \{x, y, w, h\}} \text{SmoothL1}(b_{ij} - b_{ij}^*), \quad (5)$$

where i denotes the index of the detected object; $j \in \{x, y, w, h\}$ represents the four bounding box parameters (center coordinates x, y , width w , and height h); $1_{\{y_i > 0\}}$ is an indicator function ensuring that regression is applied only to foreground objects (excluding background); b_{ij} represents the predicted bounding box parameters; b_{ij}^* represents the ground-truth bounding box parameters; $\text{SmoothL1}(\cdot)$ denotes the Smooth L1 loss function. This optimization removes dependency on class labels, focusing solely on object localization, which improves generalization to unseen categories.

Category-Agnostic Mask Prediction Similarly, in category-agnostic mask prediction, we eliminate category-specific mask prediction, ensuring that all objects share a single mask prediction mechanism:

$$\mathcal{L}_{mask}^{CA} = - \sum_i M_i^* \log M_i + (1 - M_i^*) \log(1 - M_i), \quad (6)$$

where: i denotes the object index; M_i represents the predicted object mask, with values in the range $[0, 1]$; M_i^* is the ground-truth binary mask, where 1 represents foreground pixels and 0 represents background pixels; This loss function employs binary cross-entropy loss (BCE Loss) to measure the similarity between the predicted and ground-truth masks. By eliminating class dependencies in mask prediction, the

model becomes more adaptable to unseen object instances, significantly improving generalization capability in open-vocabulary object detection.

B. Image-to-Event Contrastive Distillation

Once candidate proposals are generated, we leverage a pretrained vision-language model (CLIP) to classify each region, thereby enabling open vocabulary object detection. However, while CLIP excels in frame-based vision tasks, event cameras perceive dynamic scenes in an asynchronous and sparse manner, resulting in a significant modality gap in data distribution and feature representation compared to conventional images. As a result, directly applying CLIP to zero-shot event-based detection leads to high errors. To bridge this modality gap, we propose cross-modal knowledge distillation, transferring the image-based knowledge from CLIP’s pretrained model to an event-based object detector. Specifically, we use the CLIP image encoder as a teacher model and apply knowledge distillation to enable the event-based student detector to learn CLIP’s rich visual representations, thereby improving its generalization in open vocabulary event-based detection.

We divide the categories in the detection dataset into a base category subset and a novel category subset, denoted as C_B and C_N , respectively. The model is trained only using annotations from C_B . In the pre-trained CLIP model, the text encoder and image encoder are represented as $\mathcal{T}(\cdot)$ and $\mathcal{V}(\cdot)$. We train a proposal generation network on the base categories C_B and extract region proposals $r_e \in P_e$ from the event stream. Subsequently, we reconstruct the corresponding image frames from the event stream and map the proposals r_e onto the image frames P_i . These candidate regions r_i are then cropped and resized before being fed into the frozen CLIP image encoder \mathcal{V} to compute the image embedding: $\mathcal{V}(\text{crop}(P_i, r_i))$. To transfer the image knowledge from the CLIP pre-trained model to an event-based object detection model, we align the event-region embedding detected from the event stream, $\mathcal{R}(\phi(P_e), r_e)$, with the image embedding extracted by the CLIP image encoder from the proposal regions in the image frames, $\mathcal{V}(\text{crop}(P_i, r_i))$. This process aims to bridge the modality gap between event data and image frames, enabling the effective utilization of CLIP’s pre-trained visual representations. To further enhance cross-modal feature alignment, we apply a contrastive loss between event camera region embeddings and image frame embeddings to minimize their representation discrepancy. Additionally, we introduce trainable projection layers, which map event features $f_{r_e}^{evt}$ and image frame features $\mathcal{V}(\text{crop}(P_i, r_i))$ into the same feature space, ensuring efficient cross-modal knowledge transfer. In the process of knowledge distillation for event features, we take into full consideration the sparsity of event data and its prominent edge characteristics. To better extract and align event features with image features, we introduce an enhanced spatial attention mechanism into the distillation process. Specifically, we first utilize high-level semantic information from the teacher network to generate a spatial attention map, which highlights key information

regions in the event data while suppressing redundant noise. The spatial attention map is constructed as follows:

$$\mathcal{P}(F)_{i,j} = \frac{1}{C} \sum_{c=1}^C |F_{c,i,j}|, \mathcal{N}(F) = \text{softmax} \left(\frac{\mathcal{P}(F)}{\tau} \right), \quad (7)$$

where $\mathcal{P}(F)$ represents the average pooling result along the channel dimension, and $\mathcal{N}(F)$ is the normalized attention map obtained through softmax, which adjusts the feature distribution to focus on important spatial locations. Next, we fuse the attention maps of event features $\mathcal{N}(F^{evt})$ and image features $\mathcal{N}(F^{img})$ to obtain the final spatially enhanced representation: $A = \frac{\mathcal{N}(F^{evt}) + \mathcal{N}(F^{img})}{2}$. We then apply a transformation to the event features as $F^* = \mathcal{G}(F^{evt})$, where \mathcal{G} is a mapping module that adjusts the feature scale. On this basis, our event-to-image knowledge distillation loss function \mathcal{L}_{F2E} is modified as follows:

$$\mathcal{L}_{F2E}(\theta_e, \omega_e, \omega_f) = - \sum_i \log \frac{e^{(A_i \cdot f_i^{evt}, \mathcal{V}(\text{crop}(P_i, r_i))) / \tau_1}}{\sum_{j \neq i} e^{(A_i \cdot f_i^{evt}, f_j^{img}) / \tau_1}}, \quad (8)$$

where the spatial attention weight A_i is applied to the event feature f_i^{evt} , ensuring that the event features focus more on key regions, thereby enhancing the distillation effect. Ultimately, this strategy improves the semantic consistency of event features, enabling more effective alignment with image features during the knowledge distillation process and enhancing the model’s generalization ability in object detection tasks. Through this method, we achieve VLM knowledge transfer to event cameras without requiring large-scale event-text dataset training, enabling open vocabulary object detection while mitigating generalization issues caused by the modality gap.

C. Classification For Novel Categories

For open vocabulary object detection, another critical challenge is how to classify novel category samples. We address this issue by replacing the traditional classifier with text embeddings extracted from CLIP. Specifically, we embed category names into a prompt template (e.g., "A photo of a class in the scene.") and pass them through the text encoder \mathcal{T} to generate category text embeddings. Our objective is to enable the knowledge-distilled event region features to be classified using text embeddings. During training, we use only $\mathcal{T}(C_B)$, i.e., the text embeddings of base categories C_B . For proposal regions that do not match any ground-truth annotations in C_B , we classify them as background categories. Since the textual representation "background" may not adequately describe these unmatched proposals, we allow the background category to learn its own embedding e_{bg} , ensuring it acquires an independent representation in the semantic space. To achieve this, we compute the cosine similarity between each region embedding $\mathcal{R}(\phi(P_e), r_e)$ and all category embeddings, including both $\mathcal{T}(C_B)$ and e_{bg} . We apply softmax normalization with a temperature parameter τ to compute the \mathcal{L}_{CE} cross-entropy loss, optimizing the classification distribution. Meanwhile, to train the first-stage region proposal network in the two-stage object detector, we

extract region proposals $r_e \in P_e$ and train the detector, with the loss function defined as: $e_r = \mathcal{R}(\phi(P_e), r_e)$

$$\mathbf{z}(r) = [\text{sim}(e_r, e_{bg}), \text{sim}(e_r, t_1), \dots, \text{sim}(e_r, t_{|C_B|})], \quad (9)$$

$$\mathcal{L}_{\text{text}} = \frac{1}{N} \sum_{r_e \in P_e} \mathcal{L}_{CE}(\text{softmax}(\mathbf{z}(r_e)/\tau), y_r), \quad (10)$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b} / (\|\mathbf{a}\| \|\mathbf{b}\|)$, t_i denotes elements in $\mathcal{T}(C_B)$, y_r denotes the class label of region r_e , N is the number of proposals per event ($|P_e|$).

Inference: During inference, we rely solely on event stream data for open vocabulary object detection, in contrast to the training phase, where image frames from the event stream were additionally used. Meanwhile, we introduce novel categories C_N to extend the model’s open vocabulary detection capability. Our goal is that the knowledge distilled from CLIP’s image representations can enhance generalization of the event-based object detection model to C_N .

V. EXPERIMENT

A. Datasets and Implementation

NCAR Dataset: The NCAR dataset [36] is a binary classification dataset consisting of 12,336 car samples and 11,693 background samples. Each sample spans a duration of 100 ms and exhibits varying spatial dimensions. **GEN1 Automotive Detection Dataset:** The first-generation automotive detection dataset [37] consists of 39 hours of event camera recordings with a resolution of 304×240. Overall, the dataset contains 228,000 car bounding boxes and 28,000 pedestrian bounding boxes. **The DSEC dataset [38]** is a high-resolution, large-scale event-frame dataset designed for real-world driving scenarios, captured with a 640×480 resolution event camera alongside RGB image frames. The original dataset lacked object detection annotations, so we utilize the labels introduced in [39], which include three object categories: cars, pedestrians, and large vehicles.

Implementation ViLD with Enhanced Teacher Models. For experiments with stronger teachers (CLIP ViT-L/14, ALIGN), we adopt EfficientNet-B7 as the backbone and the ViLD-ensemble architecture. RoI features are extracted only from FPN level P3, and the image jittering range is reduced to [0.5, 2.0]. For CLIP ViT-L/14 (768-d embeddings), the fully connected layers in Faster R-CNN heads are expanded to 1,024 dimensions, and the FPN feature dimension is set to 512. For ALIGN, which combines an EfficientNet-L2 image encoder with a BERT-large text encoder, we modify Mask R-CNN to better distill teacher knowledge. The ViLD-image head is enhanced with EfficientNet MBConvBlocks, followed by global average pooling to produce embeddings consistent with ALIGN. The ViLD-text head keeps the original Faster R-CNN design. Since ALIGN outputs 1,376-d embeddings (2.7× CLIP), the fully connected layers in the ViLD-text head are expanded to 2,048 units, with FPN features increased to 1,024 dimensions.

B. Comparative Study

Novel Category: For the DSEC dataset [38], the primary results are shown in Table II. All the compared methods are trained using both event streams and image frames.

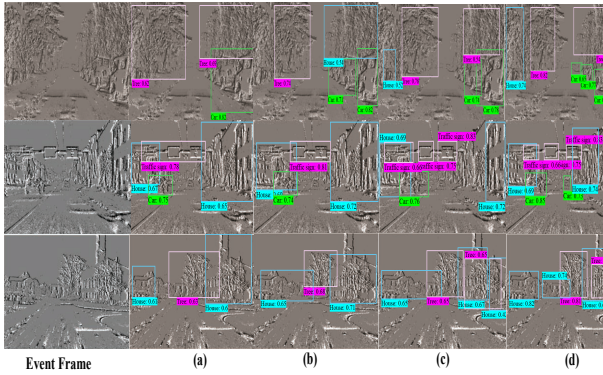


Fig. 2. **Open Vocabulary Object Detection results on DSEC dataset[38]:** From left to right; the models are Event frame; ViLD [28]; RegionCLIP [40]; YOLO-World [41]; Ours.

Method	Architecture	ACC
HATS [36]	N/A	0.902
HybridSNN [42]	SNN-CNN	0.906
EvS-S [43]	GNN	0.931
HybridSNN [42]	SNN	0.770
Gabor-SNN [36]	SNN	0.789
SqueezeNet [44]	SNNs	0.846
MobileNet-64 [44]	SNN	0.917
DenseNet169-16[44]	SNN	0.904
VGG-11 [44]	SNN	0.924
SFOD[45]	SNN	0.937
CREST [46]	SNN	0.949
Ours	SNN+CNN	0.957

TABLE I

COMPARISON WITH EXISTING METHODS ON NCAR DATASET [36].

Our model is trained solely on Cars and Pedestrians as base categories, while Large Vehicle is treated as a novel category to evaluate the model’s open-vocabulary detection capability. On the base category test set, our model achieves an accuracy of 54.5% on Car and 31.2% on Pedestrian. Notably, despite not being trained on Large Vehicle, our model still achieves 42.4% accuracy, surpassing models that include Large Vehicle in their training data. This result demonstrates the strong generalization ability of our model in open-vocabulary object detection.

Zero-shot Object Recognition: Our model, trained on the DSEC dataset[38], is directly applied to the NCAR dataset [36] for zero-shot object recognition. Table I compares our model with other SOTA methods trained on the DSEC dataset. The results demonstrate that our model not only outperforms other SNN-based models in terms of accuracy but also surpasses non-SNN-based models, further validating its effectiveness in zero-shot object recognition tasks.

Zero-shot Object Detection: We train our model on the DSEC dataset [38] and directly apply it to the Gen1 dataset [37] for zero-shot open-vocabulary object detection, with the key results shown in Table III. Notably, even compared to models trained on Gen1, our model demonstrates superior performance in open-vocabulary detection, highlighting its strong cross-dataset generalization capability. Specifically, our model achieves a mean Average Precision (mAP_{50}) of 65.7% at an IoU threshold of 0.5, and $mAP_{50:95}$ of 38.3% over the IoU range of 0.5 to 0.95. These results validate that our method effectively transfers the knowledge learned from

Method	Car mAP(%)	Pedestrian mAP(%)	Large vehicle mAP
RAMNet [47]	0.244	0.108	0.176
SENet [48]	0.384	0.149	0.260
ECANet [49]	0.367	0.128	0.275
CBAM [50]	0.377	0.135	0.270
SAGate [51]	0.325	0.104	0.160
DCF [52]	0.363	0.127	0.280
SPNet [53]	0.392	0.178	0.262
FPN-Fusion [39]	0.375	0.109	0.249
DRFuser [54]	0.386	0.151	0.306
CMX [55]	0.416	0.164	0.294
FAGC [56]	0.398	0.144	0.336
RENet [57]	0.405	0.172	0.306
EFNet [58]	0.411	0.158	0.326
CAFR [59]	0.499	0.258	0.382
Ours	0.545 (Basic)	0.312 (Basic)	0.408 (Novel)

TABLE II

COMPARISON OF THE STATE-OF-THE-ART METHODS ON THE DSEC DATASET [38]. BASE CATEGORIES: CLASSES USED FOR TRAINING. NOVEL CATEGORY: UNSEEN CLASS EVALUATED WITHOUT TRAINING.

DSEC and enables open-vocabulary detection on unseen event datasets, achieving robust detection performance even for novel categories.

Comparison with Open-vocabulary Object Detection:

To assess our model’s open-vocabulary performance on event data, gauge its advantage over image-based methods, and evaluate the SNN-driven Adaptive Event Slicing module, we compared it with SOTA open-vocabulary detectors. We first fed the grayscale event frames provided by the dataset into the image detectors and followed the open-vocabulary protocol: Car and Pedestrian served as base categories for training, while Large Vehicle was kept as a zero-shot novel category. As Table IV shows, detection accuracy on base classes is markedly higher than on the novel class, confirming that image-trained methods struggle with zero-shot detection on event cameras due to the modality gap between events and images. Next, we inserted Adaptive Event Slicing as the feature extractor so that features were drawn directly from raw event streams. This raised overall accuracy, demonstrating its effectiveness, but still fell short of our model. Overall, the results indicate that our approach successfully transfers CLIP’s visual knowledge to event data and improves open-vocabulary object detection on event cameras.

Figure 2 compares our method with existing open-vocabulary detectors on event data. For a fair evaluation, all methods are tested on the grayscale event frames provided by the dataset. The results indicate that detectors trained on conventional images perform well only when object contours are crisp (e.g., houses, trees) but suffer from mis-detections under occlusion or blur. By contrast, our approach accurately recognises targets even in overlapping or blurred regions, exhibiting greater robustness and generalisation, and confirming its advantage on event-camera inputs.

C. Ablation Study

Adaptive Event Slicing: We conducted an ablation study on the two key loss functions in the Adaptive Event Slicing module: Linear Incremental Constraint Loss and Self-Supervised Feedback Loss. As demonstrated in Table V, the adaptive event segmentation strategy significantly improves detection accuracy. By comparing the first and second rows, we observed that Linear Incremental Constraint Loss

Method	Architecture	mAP ₅₀	mAP _{50:95}
S-Center [60]	ANN	-	0.278
EGO-12 [61]	ANN	-	0.504
Spiking-Yolo [62]	ANN2SNN	-	0.257
Spike Calib [63]	ANN2SNN	0.454	-
Spike Transformer v2 [64]	ANN2SNN	0.512	-
VC-Dense [44]	SNN	-	0.189
S-Center [60]	SNN	-	0.229
TR-YOLO [65]	SNN	0.451	-
EMS-YOLO [66]	SNN	0.501	0.301
SFOD [45]	SNN	0.5093	0.321
CREST [46]	SNN	0.632	0.360
Ours	SNN + ANN	0.657	0.383

TABLE III

COMPARISON OF EXISTING METHODS ON THE GEN1 DATASET [37].

Method	Car mAP(%)	Pedestrian mAP(%)	Large vehicle mAP
	Base	Base	Novel
VILD [28]	0.343	0.229	0.080
RegionCLIP [40]	0.357	0.231	0.085
FVLM [68]	0.361	0.235	0.088
YOLO-World [41]	0.364	0.237	0.092
Adaptive Event Slicing (SNN) +VILD [28]	0.375	0.254	0.092
Adaptive Event Slicing (SNN) +RegionCLIP [40]	0.381	0.258	0.099
Adaptive Event Slicing (SNN) + FVLM [68]	0.392	0.261	0.105
Adaptive Event Slicing (SNN) + YOLO-World [41]	0.398	0.264	0.113
Ours	0.545	0.312	0.408

TABLE IV

COMPARISON OF THE STATE-OF-THE-ART OPEN-VOCABULARY OBJECT DETECTION METHODS ON THE DSEC DATASET [38]. BASE CATEGORIES: CLASSES USED FOR TRAINING. NOVEL CATEGORY: UNSEEN CLASS EVALUATED WITHOUT TRAINING.

effectively enhances the model’s object detection accuracy. Furthermore, comparing the second and third rows, we found that Self-Supervised Feedback Loss dynamically adjusts the membrane potential based on object detection results, adaptively optimizing the event segmentation strategy and extracting more discriminative event features, leading to further performance improvement. **Knowledge Distillation:** We investigated the impact of removing the knowledge distillation between image and event data, instead performing object detection using event features extracted by the SNN. As shown in Table V (third and fourth rows), the detection performance drops significantly. This result highlights the critical role of knowledge distillation in feature transfer between event data and images, demonstrating its effectiveness in improving the detector’s generalization ability. Building on this, spatial attention was further incorporated. As shown in the fourth and fifth rows, the results indicate that it further improves the effectiveness of knowledge transfer. **Frame Branch:** We also examined the scenario where the model relies solely on event camera data (without frame camera input). During knowledge distillation, we replaced the frame input of the teacher branch with event reconstruction results [67]. As shown in Table V (fifth and sixth rows), since event-reconstructed images provide only limited visual cues, this substitution generally reduces representation learning quality. Performance degrades compared to knowledge transfer based on real frame data. However, our model remains applicable in cases where no image frames are available.

LIC	SSF	KD	Frame	mAP ₅₀	mAP _{50:95}
				0.459	0.227
✓				0.470	0.232
✓	✓			0.486	0.257
✓	✓	✓		0.621	0.352
✓	✓	✓	✓	0.657	0.383

TABLE V

ABLATION STUDY OF OUR METHODS ON GEN1: LIC: LINEAR INCREMENTAL CONSTRAINT LOSS. SSF: SELF-SUPERVISED FEEDBACK LOSS. KD: KNOWLEDGE DISTILLATION.

VI. CONCLUSION

We present the first open-vocabulary object detection framework for event data, which transfers visual knowledge from CLIP into an event-based detector. To address event stream segmentation, we introduce an SNN-based module that adaptively selects slicing points and extracts discriminative features, forming a collaborative paradigm between SNNs and ANNs. Experiments demonstrate that our model consistently outperforms dataset-specific baselines. Moreover, even without access to image frames, frame reconstructions from event streams can be used for knowledge distillation, enabling the model to retain both strong learning capacity and robust generalization on event data.

Acknowledgment: This work is supported by NSF Grants NO. 2340882, 2334624, 2334246, and 2334690.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*. PmLR, 2021, pp. 8748–8763.
- [2] M. Xu, H. Liu, N. Peng, Q. Chen, and C. Xiao, “Affordance-first decomposition for continual learning in video-language understanding,” *arXiv preprint arXiv:2512.00694*, 2025.
- [3] L. Wang, Y.-S. Ho, K.-J. Yoon *et al.*, “Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks,” in *CVPR*, 2019.
- [4] Y. Li, Y. Li, K. Zhang, F. Zhang, C. Yang, Z. Guo, W. Ding, and T. Huang, “Achieving fair medical image segmentation in foundation models with adversarial visual prompt tuning,” *Information Sciences*, p. 122501, 2025.
- [5] Y. Peng, Y. Zhang, P. Xiao, X. Sun, and F. Wu, “Better and faster: Adaptive event conversion for event-based object detection,” in *AAAI*, 2023.
- [6] J. Zhang, Y. Wang, W. Liu, M. Li, J. Bai, B. Yin, and X. Yang, “Frame-event alignment and fusion network for high frame rate tracking,” in *CVPR*, 2023.
- [7] Z. Wang, B. Li, C. Wang, and S. Scherer, “Airshot: Efficient few-shot detection for autonomous exploration,” in *IROS*. IEEE, 2024.
- [8] H. Zheng, Y. Lin, Q. He, Y. Zou, and H. Wang, “Blockchain payment fraud detection with a hybrid cnn-gnn-lstm model,” in *ICCECE*. IEEE, 2026.
- [9] F. Li, Q. Xu, S. Bao, Z. Yang, R. Cong, X. Cao, and Q. Huang, “Size-invariance matters: Rethinking metrics and losses for imbalanced multi-object salient object detection,” *arXiv preprint arXiv:2405.09782*, 2024.
- [10] S. Bao, Q. Xu, F. Li, B. Han, Z. Yang, X. Cao, and Q. Huang, “Towards size-invariant salient object detection: A generic evaluation and optimization approach,” *TPAMI*, 2025.
- [11] Z. Zhao, “Balf: Simple and efficient blur aware local feature detector,” in *WACV*, 2024.
- [12] J. Edstedt, G. Bökman, and Z. Zhao, “Dedode v2: Analyzing and improving the dedode keypoint detector,” in *CVPR*, 2024.
- [13] C. Ni, W. Jiang, C. Cai, Q. Zhu, and Y. Fang, “Feature adaptive yolo for remote sensing detection in adverse weather conditions,” in *VCIP*. IEEE, 2023.
- [14] J. Zhang and G. Lu, “Underground mapping and localization based on ground-penetrating radar,” in *ACCV*, 2024.

- [15] J. Lin, J. Zhang, G. Jin, W. Song, T. Liu, and G. Lu, “3d plant root skeleton detection and extraction,” in *IROS*. IEEE, 2025.
- [16] J. Lin, J. Zhang, and G. Lu, “Keypoint detection and description for raw bayer images,” in *ICRA*. IEEE, 2025.
- [17] Q. Ning, J. Zhang, Y. Xie, K. Liu, K. Gao, B. Chen, G. Chen, Q. Fan, H. Liu, and R. Du, “Multi-resolution context augmentation and dual channel attention for 3d lane detection,” *ITJ*, 2025.
- [18] P. Pan, L. Chen, Q. He, K. Yuan, H. Wang, and W. Zhang, “Finscra: An llm-powered multi-chain reasoning framework for interpretable node classification on text-attributed graphs,” 2026.
- [19] Y. Hu, Q. Zheng, X. Jiang, and G. Pan, “Fast-snn: Fast spiking neural network by converting quantized ann,” *TPAMI*, 2023.
- [20] J. Qu, Z. Gao, T. Zhang, Y. Lu, H. Tang, and H. Qiao, “Spiking neural network for ultralow-latency and high-accurate object detection,” *TNNLS*, 2024.
- [21] D. Przewlocka-Rus, T. Kryjak, and M. Gorgon, “Poweryolo: Mixed precision model for hardware efficient object detection with event data,” in *DSD*. IEEE, 2024, pp. 210–217.
- [22] B. Zhou and J. Jiang, “Deep event-based object detection in autonomous driving: A survey,” *arXiv preprint arXiv:2405.03995*, 2024.
- [23] X. Wang, Y. Jin, W. Wu, W. Zhang, L. Zhu, B. Jiang, and Y. Tian, “Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset,” *arXiv:2412.06647*, 2024.
- [24] N. Xu, Z. Ma, Y. Xia, Y. Dong, J. Zi, D. Xu, F. Xu, X. Su, H. Zhang, and F. Chen, “A serial multi-scale feature fusion and enhancement network for amur tiger re-identification,” *Animals*, 2024.
- [25] K. Chen, X. Jiang, H. Wang, C. Yan, Y. Gao, X. Tang, Y. Hu, and W. Xie, “Ov-dar: Open-vocabulary object detection and attributes recognition,” *IJCV*, 2024.
- [26] C. Zhu and L. Chen, “A survey on open-vocabulary detection and segmentation: Past, present, and future,” *TPAMI*, 2024.
- [27] P. Du, Y. Wang, Y. Sun, L. Wang, Y. Liao, G. Zhang, E. Ding, Y. Wang, J. Wang, and S. Liu, “Lami-det: Open-vocabulary detection with language model instruction,” in *ECCV*. Springer, 2024, pp. 312–328.
- [28] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, “Open-vocabulary object detection via vision and language knowledge distillation,” *arXiv preprint arXiv:2104.13921*, 2021.
- [29] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, “Promptdet: Towards open-vocabulary detection using uncurated images,” in *ECCV*. Springer, 2022, pp. 701–717.
- [30] Z. Wu, X. Liu, and I. Gilitschenski, “Eventclip: Adapting clip for event-based object recognition,” *arXiv:2306.06354*, 2023.
- [31] W. Shin, J. Park, T. Woo, Y. Cho, K. Oh, and H. Song, “e-clip: Large-scale vision-language representation learning in e-commerce,” in *CIKM*, 2022, pp. 3484–3494.
- [32] Y. Li, C. Yang, H. Zeng, Z. Dong, Z. An, Y. Xu, Y. Tian, and H. Wu, “Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting,” in *ICCV*, 2025.
- [33] P. Bardow, A. J. Davison, and S. Leutenegger, “Simultaneous optical flow and intensity estimation from an event camera,” in *CVPR*, 2016.
- [34] J. Cao, M. Sun, Z. Wang, H. Cheng, Q. Zhang, R. Xu *et al.*, “Spiking neural network as adaptive event stream slicer,” in *NeurIPS*, 2024.
- [35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *ICCV*, 2017.
- [36] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, “Hats: Histograms of averaged time surfaces for robust event-based object classification,” in *CVPR*, 2018.
- [37] P. De Tournemire, D. Nitti, E. Perot, D. Migliore, and A. Sironi, “A large scale event-based detection dataset for automotive,” *arXiv preprint arXiv:2001.08499*, 2020.
- [38] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, “Dsec: A stereo event camera dataset for driving scenarios,” *RAL*, 2021.
- [39] A. Tomy, A. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, “Fusing event-based and rgb camera for robust object detection in adverse conditions,” in *ICRA*. IEEE, 2022, pp. 933–939.
- [40] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, “Regionclip: Region-based language-image pretraining,” in *CVPR*, 2022, pp. 16 793–16 803.
- [41] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, “Yolo-world: Real-time open-vocabulary object detection,” in *CVPR*, 2024.
- [42] A. Kugele, T. Pfeil, M. Pfeiffer, and E. Chicca, “Hybrid snn-ann: Energy-efficient classification and object detection for event-based vision,” in *GCPR*. Springer, 2021, pp. 297–312.
- [43] Y. Li, H. Zhou, B. Yang, Y. Zhang, Z. Cui, H. Bao, and G. Zhang, “Graph-based asynchronous event processing for rapid object recognition,” in *ICCV*, 2021.
- [44] L. Cordone, B. Miramond, and P. Thierion, “Object detection with spiking neural networks on automotive event data,” in *IJCNN*. IEEE, 2022.
- [45] Y. Fan, W. Zhang, C. Liu, M. Li, and W. Lu, “Sfod: Spiking fusion object detector,” in *CVPR*, 2024.
- [46] R. Mao, A. Shen, L. Tang, and J. Zhou, “Crest: An efficient conjointly-trained spike-driven framework for event-based object detection exploiting spatiotemporal dynamics,” *AAAI*, 2025.
- [47] D. Gehrig, M. Rügge, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, “Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction,” *RAL*, vol. 6, no. 2, pp. 2822–2829, 2021.
- [48] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *CVPR*, 2018.
- [49] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in *CVPR*, 2020.
- [50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *ECCV*, 2018.
- [51] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, “Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation,” in *ECCV*. Springer, 2020, pp. 561–577.
- [52] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, “Calibrated rgb-d salient object detection,” in *CVPR*, 2021.
- [53] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, “Specificity-preserving rgb-d saliency detection,” in *ICCV*, 2021.
- [54] F. Munir, S. Azam, K.-C. Yow, B.-G. Lee, and M. Jeon, “Multimodal fusion for sensorimotor control in steering angle prediction,” *EAAI*, vol. 126, p. 107087, 2023.
- [55] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelwagen, “Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers,” *IEEE TITS*, vol. 24, no. 12, pp. 14 679–14 694, 2023.
- [56] H. Cao, G. Chen, J. Xia, G. Zhuang, and A. Knoll, “Fusion-based feature attention gate component for vehicle detection based on event camera,” *IEEE Sensors Journal*, 2021.
- [57] Z. Zhou, Z. Wu, R. Boutteau, F. Yang, C. Demonceaux, and D. Ginhac, “Rgb-event fusion for moving object detection in autonomous driving,” in *ICRA*. IEEE, 2023, pp. 7808–7815.
- [58] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, “Event-based fusion for motion deblurring with cross-modal attention,” in *ECCV*. Springer, 2022, pp. 412–428.
- [59] H. Cao, Z. Zhang, Y. Xia, X. Li, J. Xia, G. Chen, and A. Knoll, “Embracing events and frames with hierarchical feature refinement network for object detection,” in *ECCV*. Springer, 2024.
- [60] L. Bodden, D. B. Ha, F. Schwaiger, L. Kreuzberg, and S. Behnke, “Spiking centernet: A distillation-boosted spiking neural network for object detection,” in *IJCNN*. IEEE, 2024, pp. 1–9.
- [61] N. Zubić, D. Gehrig, M. Gehrig, and D. Scaramuzza, “From chaos comes order: Ordering event representations for object recognition and detection,” in *ICCV*, 2023.
- [62] S. Kim, S. Park, B. Na, and S. Yoon, “Spiking-yolo: spiking neural network for energy-efficient object detection,” in *AAAI*, 2020.
- [63] Y. Li, X. He, Y. Dong, Q. Kong, and Y. Zeng, “Spike calibration: Fast and accurate conversion of spiking neural network for object detection and segmentation,” *arXiv preprint arXiv:2207.02702*, 2022.
- [64] M. Yao, J. Hu, T. Hu, Y. Xu, Z. Zhou, Y. Tian, B. Xu, and G. Li, “Spike-driven transformer v2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips,” *arXiv preprint arXiv:2404.03663*, 2024.
- [65] M. Yuan, C. Zhang, Z. Wang, H. Liu, G. Pan, and H. Tang, “Trainable spiking-yolo for low-latency and high-performance object detection,” *NN*, vol. 172, p. 106092, 2024.
- [66] Q. Su, Y. Chou, Y. Hu, J. Li, S. Mei, Z. Zhang, and G. Li, “Deep directly-trained spiking neural networks for object detection,” in *ICCV*, 2023.
- [67] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” *TPAMI*, 2019.
- [68] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, “F-vlm: Open-vocabulary object detection upon frozen vision and language models,” *arXiv preprint arXiv:2209.15639*, 2022.