

Object Detection using Event Camera: A MoE Heat Conduction based Detector and A New Benchmark Dataset

Xiao Wang¹, Yu Jin¹, Wentao Wu², Wei Zhang³, Lin Zhu⁴, Bo Jiang^{1*}, Yonghong Tian^{3,5,6}

¹School of Computer Science and Technology, Anhui University, Hefei, China

²School of Artificial Intelligence, Anhui University, Hefei, China

³Peng Cheng Laboratory, Shenzhen, China

⁴Beijing Institute of Technology, Beijing, China

⁵National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University, China

⁶School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China

{xiaowang, jiangbo}@ahu.edu.cn, {jy0x4f, wuwentao0708}@163.com,
zhangwei1213052@126.com, {linzhu, yhtian}@pku.edu.cn

Abstract

Object detection in event streams has emerged as a cutting-edge research area, demonstrating superior performance in low-light conditions, scenarios with motion blur, and rapid movements. Current detectors leverage spiking neural networks, Transformers, or convolutional neural networks as their core architectures, each with its own set of limitations including restricted performance, high computational overhead, or limited local receptive fields. This paper introduces a novel MoE (Mixture of Experts) heat conduction-based object detection algorithm that strikingly balances accuracy and computational efficiency. Initially, we employ a stem network for event data embedding, followed by processing through our innovative MoE-HCO blocks. Each block integrates various expert modules to mimic heat conduction within event streams. Subsequently, an IoU-based query selection module is utilized for efficient token extraction, which is then channeled into a detection head for the final object detection process. Furthermore, we are pleased to introduce EvDET200K, a novel benchmark dataset for event-based object detection. Captured with a high-definition Prophesee EVK4-HD event camera, this dataset encompasses 10 distinct categories, 200,000 bounding boxes, and 10,054 samples, each spanning 2 to 5 seconds. We also provide comprehensive results from over 15 state-of-the-art detectors, offering a solid foundation for future research and comparison. The source code has been released on: <https://github.com/Event-AHU/OpenEvDET>

1. Introduction

Object detection aims to identify predefined target objects by delineating them with bounding boxes and assigning category labels. It stands as a cornerstone problem in computer vision and finds extensive application across fields such as intelligent video surveillance, autonomous vehicles, and industrial automation. With the advent of deep learning, a plethora of cutting-edge deep object detectors have emerged, demonstrating remarkable performance with RGB cameras. Notable examples include the RCNN variants [11, 12, 31], YOLO-based models [1, 16, 24, 33, 35, 41], and DETR-inspired detectors [3, 21, 46, 48, 52, 53]. Nonetheless, frame-based object detectors continue to struggle in demanding conditions such as low light, complex backgrounds, and rapid motion. The constraints in image quality inherent in traditional frame cameras, which capture images at a fixed frame rate (e.g., 30 FPS) and employ a uniform exposure setting, are largely to blame for the prevalence of missed and erroneous detections across various detection algorithms.

To overcome the limitations of traditional sensors, researchers have turned to innovative technologies for object detection. Among these, bio-inspired event cameras, also known as Dynamic Vision Sensors (DVS), have garnered significant interest within the computer vision sphere. Event cameras outshine conventional RGB frame-based cameras in several aspects: *high dynamic range, high temporal resolution, low energy consumption*, and nearly no motion blur. These sensors have been applied to a range of tasks, from high-level applications such as event-based tracking [38], recognition [39], and captioning [37, 47]; to

* Corresponding Author: Bo Jiang

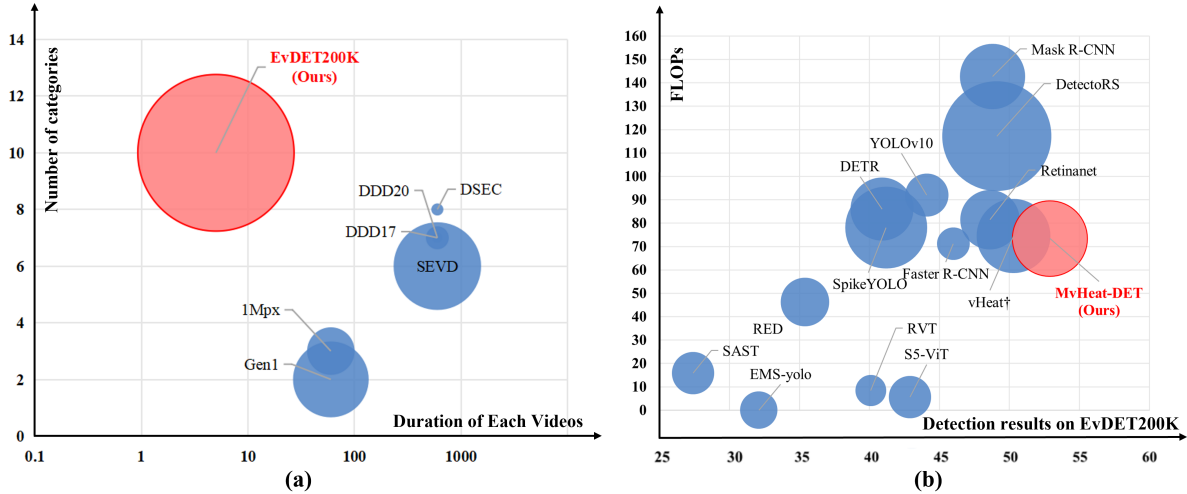


Figure 1. (a). Comparison of existing datasets and our proposed EvDET200K dataset for event stream based detection, the bubble size represents the scale of the dataset; (b). Comparison of our proposed MvHeat-DET and existing SOTA detectors on the EvDET200K dataset, the bubble size represents parameters.

low-level operations including image enhancement and reconstruction [51]. In the realm of event-based object detection, DAGr [9], introduced by Daniel et al., integrates RGB frames with event streams to strike an optimal balance between latency and bandwidth, all while maintaining detection precision. SpikeYOLO [24] has made strides by incorporating the I-LIF spiking neuron, along with integer training and spike-driven inference, to minimize quantization errors in Spiking Neural Networks (SNNs).

Despite the notable advancements made, current event-based object detectors still face the following issues: **1).** Most event-based detectors rely on CNNs (Convolutional Neural Networks) or Transformers as their backbone architectures. However, CNNs are constrained by their local receptive fields, which hinders their ability to capture long-range and intricate dependencies. In contrast, Transformer-based vision models, such as ViT [6], grapple with high computational complexity, scaling as $\mathcal{O}(N^2)$, and lack interpretability. **2).** Some researchers resort to bio-inspired SNNs (Spiking Neural Networks) [24, 33] for encoding event streams, reaping benefits in terms of energy efficiency. Nonetheless, their overall performance lags significantly behind that of ANN (Artificial Neural Network)-based detectors. Consequently, the quest for an effective, efficient, and interpretable event-based object detection algorithm remains a formidable challenge.

Recently, Wang et al. proposed a physics-inspired vision backbone model, vHeat [42], which is grounded in the principles of heat conduction. The core module of this model is the Heat Conduction Operator (HCO), which envisions image patches as heat sources and conceptualizes the determination of their correlations as the process of ther-

mal energy diffusion. By employing 2D Discrete Cosine Transform (DCT) and Inverse Discrete Cosine Transform (IDCT) operations to approximate the HCO, they achieve a lower computational complexity of $\mathcal{O}(N^{1.5})$. When applied to object detection, the HCO outperforms both the Swin-Transformer [22] and ConvNeXt [23]. However, we argue that the 2D DCT and IDCT operators may not be the most suitable for simulating heat conduction in the context of vision models. As these operators are designed for general signal processing, they may not fully capture the spatial and temporal dynamics inherent in visual data, which are critical for accurate object detection. This motivates us to explore alternative approaches that can more accurately and efficiently model the heat conduction process for event-based detection.

In this paper, we propose a novel backbone network for event-based detection that consists of Mixture-of-Expert (MoE) based heat conduction operators, termed MvHeat-DET. We split the input event streams into multiple clips and get their feature embeddings using a stem network. Then, we pass the event embeddings into an MHCO layer, as shown in Fig. 2, which first selects a transform branch using the policy network. Then, randomly initialized features FEs are fed into a linear layer to predict the thermal diffusivity and multiplied by the transformed representations. After that, an inverse operator is adopted for complete signal transformation. The obtained tokens will be fed into an IoU-based query selection module and detection head for object localization and recognition by following [48].

Although several event stream object detection datasets have been proposed, e.g., Gen1 [5] and 1Mpx [29], they are still relatively scarce compared to RGB frame based detec-

tion datasets. Therefore, this paper introduces a new benchmark dataset to fill this gap, named EvDET200K. It involves ten categories of target objects, including *people, cars, bicycles, electric bicycles, basketball, ping pong, goose, cats, birds,* and *UAVs*. The dataset was captured using the Prophesee EVK4-HD event camera and comprises 10,054 samples, each ranging from 2 to 5 seconds in duration. An extensive annotation effort has yielded 200,000 high-quality bounding boxes, and we provide over 15 state-of-the-art detectors for future research to benchmark against. We are confident that the introduction of EvDET200K, along with the associated benchmark algorithms, will mark a significant stride forward in the realm of event camera-based object detection.

To sum up, we draw the contributions of this paper as the following three aspects:

1). We introduce a novel Mixture-of-Experts (MoE)-based heat conduction framework, named MvHeat-DET, designed for event stream object detection. This framework strikingly balances performance, efficiency, and interpretability, offering an improved trade-off in these critical areas.

2). We present EvDET200K, a new high-definition benchmark dataset for event stream object detection. Comprising 10,054 samples captured with the Prophesee EVK4-HD camera, each sample spans 2 to 5 seconds and includes 200,000 bounding boxes spanning 10 distinct object categories.

3). We have re-trained and evaluated over 15 state-of-the-art (SOTA) object detectors, including models from the YOLO, RCNN, and DETR families, on the newly introduced EvDET200K dataset. This provides a comprehensive baseline for future research to compare and build upon.

2. Related Works

• **RGB Frame based Detection.** Object detection using RGB images has progressed significantly in recent years, primarily due to advancements in deep learning [20, 44, 49]. These detectors can be divided into three main streams, i.e., RCNN-based detectors [11, 12, 15, 31], YOLO series, and DETR-based detection algorithms. The DETR [3] model simplifies the object detection process, effectively eliminating the need for many manually designed components, such as non-maximum suppression (NMS) or anchor generation. Based on DETR, methods like Deformable DETR [52], Adaptive Clustering Transformer [50], PnP-DETR [36], and Sparse DETR [32] reduce computational resource consumption by applying sparse processing to the transformer’s attention mechanism, achieving a favorable balance between efficiency and accuracy. Furthermore, models like Conditional DETR [25], Anchor DETR [40], Efficient DETR [45], and Dynamic DETR [4] leverage spatial prior knowledge to better focus on regions of interest

(ROIs), significantly reducing learning difficulty. Notably, [34] proposes integrating R-CNN with Transformer to promote DETR convergence and yield better results. Different from these Transformer-based detectors, in this paper, we propose a novel MoE Heat Conduction based backbone network for event stream based detection algorithm.

• **Event Stream based Detection.** Event-based object detection has recently gained traction, particularly with the development of neuromorphic sensors that operate in an asynchronous, event-driven manner. Due to the specific features of event streams, current researchers usually adopt SNN (Spiking Neural Networks), Graph Neural Networks (GNN), LSTM (Long-short Term Memory) for event-based object detection. To be specific, in methods based on SNNs, SpikingYOLO [13], SpikeYOLO [24], and EMS-YOLO [33] combine YOLO with SNNs. SpikingYOLO is the first object detection model implemented in deep SNNs. SpikeYOLO introduces the I-LIF spiking neuron to reduce quantization errors in SNNs. SFOD [7] achieves multi-scale feature map fusion in SNNs for the first time, improving the model’s ability to detect objects of various sizes. DAG-r [8] employs an efficient asynchronous graph neural network to handle event data, achieving a trade-off between bandwidth and latency. Additionally, AED [19] is a lightweight detector with fast detection speed, better suited to the high temporal resolution of event cameras. GET [27] introduces the EDSA block, which effectively extracts features and enables feature communication in the spatial and time-polarity domains. S5-ViT [54] performs temporal aggregation using a state-space model (SSM), addressing the challenge of RNNs’ limited generalization when handling inputs with varying frequencies.

3. MvHeat-DET

3.1. Preliminaries: Physical Heat Conduction

In an ideal scenario, the temperature at a point with coordinates (x, y) in a two-dimensional object at time t , marked $u(x, y, t)$, is governed by the following heat conduction equation in an isotropic medium:

$$\frac{\partial u(x, y, t)}{\partial t} = k \left(\frac{\partial^2 u(x, y, t)}{\partial x^2} + \frac{\partial^2 u(x, y, t)}{\partial y^2} \right) = k(u_{xx} + u_{yy}) \quad (1)$$

where k is the thermal diffusivity, which measures the efficiency of heat diffusion within the material. To find the general solution of Eq. 1, we apply the Fourier transform to both sides of the equation, rewriting it as:

$$\mathcal{F} \frac{\partial u(x, y, t)}{\partial t} = k \mathcal{F}(u_{xx} + u_{yy}) \quad (2)$$

where \mathcal{F} is the Fourier transform function. Let $\hat{u}(v_x, v_y, t)$ denotes the Fourier transform of $u(x, y, t)$, and v_x and v_y are the frequency variables in Fourier space. This transfor-

mation converts the partial differential equation into an algebraic equation, which is simpler to solve. Therefore, we can rewrite the Eq. 2 as:

$$\frac{\partial \hat{u}(v_x, v_y, t)}{\partial t} = -k(v_x^2 + v_y^2) \hat{u}(v_x, v_y, t) \quad (3)$$

where $t = 0$ represent the initial state of the object, i.e., $u(x, y, t) |_{t=0}$. For short, we use $f(x, y)$ instead, and $\hat{f}(v_x, v_y)$ denotes the FT-transformed $f(x, y)$. By setting the initial state in Eq. 3, we can get the following solution:

$$\hat{u}(v_x, v_y, t) = \hat{f}(v_x, v_y) e^{-k(v_x^2 + v_y^2)t} \quad (4)$$

To obtain a general solution of the heat equation in the spatial domain, we apply the inverse Fourier transform (denoted as \mathcal{F}^{-1}) to Eq. 4 and get the following expression:

$$u(x, y, t) = \mathcal{F}^{-1}(\hat{f}(v_x, v_y) e^{-k(v_x^2 + v_y^2)t}) \quad (5)$$

Inspired by the aforementioned process, Wang et al. propose a new vision backbone vHeat [42] which is built based on Heat Conduction Operator (HCO). They adopt the 2D discrete cosine transformation DCT_{2D} and the 2D inverse discrete cosine transformation $IDCT_{2D}$ to simulate the HCO process in the visual domain. Despite good results that can be obtained, we think it can be further extended as the DCT-IDCT transformation may not be optimal for such a simulation. In the following subsections, we will introduce our MoE heat conduction based backbone network for event-based object detection.

3.2. Overview

As shown in Fig. 2, given the event streams, we first stack into event frames and get the event embeddings using a stem network. Then, we feed the event embeddings into the MoE-HCO blocks which provides multiple transform candidates. In this work, we consider DFT-IDFT, DCT-IDCT, and HT-IHT as three experts for the validation. A policy network with Gumbel Softmax is utilized for expert selection. In addition, the Frequency Embeddings (FEs) are used to predict the thermal diffusivity and multiplied by the transformed frequency representations. More importantly, we adopt an IoU-based query selection module [48] to find the key tokens for final detection.

3.3. MvHeat Backbone Network

Drawing inspiration from the physical concept of thermal conduction, we investigate the spread of image features across spatial domains through the lens of heat diffusion, culminating in the development of the MvHeat backbone network. This network harnesses the MoE (Mixture of Experts) Heat Conduction Operation, a novel framework that adapts thermal conduction principles for the processing of

discrete visual data features. The MvHeat backbone is engineered to deliver specialized processing for unique image features, thereby enhancing their integration and analysis.

• **MoE-HCO Block.** The MvHeat Encoder is structured into four stages, each comprising $L_i, i = \{1, 2, 3, 4\}$ MHCO (MoE Heat Conduction Operation) Layers. As the data progresses through each stage, the spatial resolution is reduced by half via downsampling, followed by processing through multiple MHCO layers. The MHCO architecture is closely similar to Vision Transformer (ViT), with a pivotal distinction: it substitutes ViT’s attention blocks with our innovative method while retaining the rest of the architectural framework. The efficacy of this structure has been validated by prior Transformer-based research, ensuring both scalability and a reduction in the computational burden associated with traditional attention mechanisms. In the MHCO module, input data is initially funneled through a selection mechanism to identify the optimal expert branch for the current feature set. Subsequently, the thermal diffusivity k is learned through Feature Embeddings (FEs) and multiplied by a coefficient matrix $e^{-k(v_x^2 + v_y^2)t}$ to generate an intermediate output. This intermediate result is then multiplied by the input data transformed into the frequency domain before being reconverted to the temporal domain.

• **MHCO: MoE Heat Conduction Operator.** As described in the section on Physical Heat Conduction, we design the MoE Heat Conduction Operator (MHCO) module to extract visual features by fully simulating thermal diffusion. This module enables effective visual feature extraction and facilitates the exchange of visual information between different image patches. Specifically, we first use deep convolution to extend the temperature distribution in the two-dimensional space along the channel dimension, with the resulting multi-channel features denoted as U_0 . Then, to obtain the output U_t after thermal diffusion, we apply Eq 5:

$$U_t = \mathcal{F}^{-1}(\mathcal{F}(U_0) e^{-k(v_x^2 + v_y^2)t}) \quad (6)$$

Following a logical reasoning process, we utilize the Discrete Fourier Transform (DFT) to transform discrete image patch features into the frequency domain and then apply the inverse Fourier Transform (IDFT) to revert them back to the spatial domain. In the physical sciences, when considering that a medium does not occupy the entire space, a unique solution to the equation requires specifying boundary conditions for u as indicated in Eq. 5. Additionally, each image patch can be viewed as a diffusion of features within a bounded space. Given that visual data is spatially constrained and semantic information does not propagate beyond the boundaries, a natural boundary condition arises. Here, we introduce a common Neumann boundary condition D :

$$\frac{\partial u(x, y, t)}{\partial \mathbf{n}} = 0, \forall (x, y) \in D, t > 0 \quad (7)$$

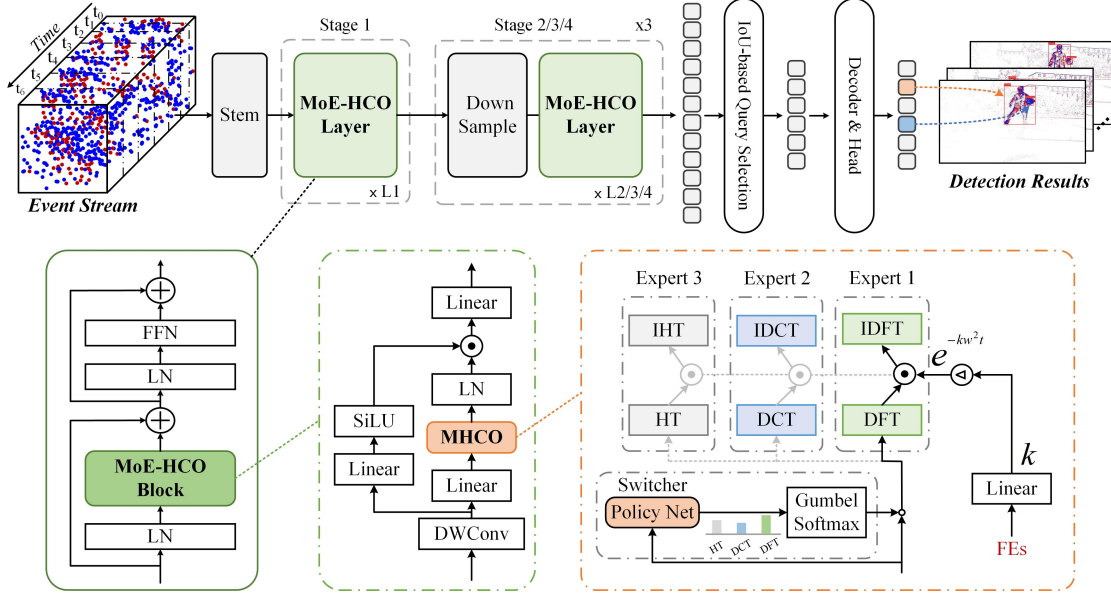


Figure 2. An overview of our proposed event-based object detection framework, termed MvHeat-DET.

where \mathbf{n} denotes the normal to the image boundary D . Additionally, because visual data is typically rectangular, this boundary condition enables us to perform reasonable transformations using the 2D Discrete Cosine Transform (DCT), 2D inverse Discrete Cosine Transform (IDCT), and Haar Transform (HT), inverse Haar Transform (IHT). Thus, Eq. 6 can be also rewritten as:

$$U_t = \mathcal{C}^{-1}(\mathcal{C}(U_0)e^{-k(v_x^2+v_y^2)t}) \quad (8)$$

$$U_t = \mathcal{H}^{-1}(\mathcal{H}(U_0)e^{-k(v_x^2+v_y^2)t}) \quad (9)$$

where \mathcal{C} denotes DCT, \mathcal{C}^{-1} denotes IDCT and \mathcal{H} denotes HT, \mathcal{H}^{-1} denotes IHT. We attempt to construct an expert network utilizing three methods simultaneously. This approach is motivated by the characteristics of transformations: Specifically, the DCT demonstrates superior performance in detecting small targets, while HT proves more effective in complex scenarios. The DFT, on the other hand, offers more generalized applicability. Consequently, the strategic integration of these transformations can lead to enhanced flexibility and superior overall performance in event detection applications.

In physical heat conduction, the thermal diffusivity k indicates how quickly heat spreads within a material. In visual heat conduction, we assume that the most noteworthy content in the image carries more "heat", and thus, visual heat should flow toward these regions. Naturally, the thermal diffusivity parameter k should be learnable and adapt to the image content, enhancing the adaptability of heat diffusion to the learning of visual representations. After applying the DCT, DFT or HT transformation, the input data x is con-

verted into the frequency domain (denoted as \hat{x}). Therefore, the learnable thermal diffusivity k also needs to be derived from the frequency domain information. Inspired by the positional embeddings in ViT, we randomly initialize a Frequency Embeddings (FEs) with the same shape as \hat{x} , which is then fed into a linear layer to predict the thermal diffusivity k . Specifically, we set a fixed value for t , and the FEs is utilized across each stage of the MvHeat network to enhance convergence throughout the training process.

3.4. IQS: IoU-based Query Selection

In the DETR model, object queries are a set of learnable embeddings. To reduce the difficulty of optimizing object queries, some studies have proposed query selection schemes, which typically leverage classification scores to select the top K features from the encoder. However, due to discrepancies between the distributions of classification scores and localization confidence, some predicted boxes with high classification scores may not be close to the ground truth. This results in a bias toward selecting boxes with high classification scores but low IoU scores, while overlooking boxes with lower classification scores but higher IoU scores. Such selection biases undermine the overall performance of the detector. During training, the IoU score is incorporated into the objective function of the classification branch, encouraging the model to associate high classification scores with ground-truth boxes that have high IoU scores. Thus, we can still select the top K based on classification scores to obtain higher-quality queries. The overall loss can be expressed as follows:

$$\mathcal{L}(y, \hat{y}) = \mathcal{L}_{bbox}(b, \hat{b}) + \mathcal{L}_{cls}(IoU, c, \hat{c}) \quad (10)$$

where $y = \{b, c\}$ denotes the ground truth boxes and categories, and $\hat{y} = \{\hat{b}, \hat{c}\}$ represents the predicted boxes and categories.

4. EvDET200K Benchmark Dataset

4.1. Protocols

We aim to provide a good platform for the training and evaluation of event-based object detection. When constructing the EvDET200K benchmark dataset, we follow the following protocols: **1). Large-scale:** With the deep integration of information technology, human production and life, large-scale datasets show an increasingly important position. In our work, we collect more than 10k event sequences, totaling about 200k objects from 10 classes. **2). Diversity:** During the shooting process, we anticipated potential challenges and configured certain factors in advance. These factors may affect the performance of the data captured by the sensor in object detection tasks. More in detail, Multi-view, Multi-illumination, Multi-motion, Dynamic Background, Non-detection Interference are all considered when recording these event streams. **3). Small Object:** We focus on enhancing the detection capability for small objects. Since small objects are often overlooked in detection tasks, we specifically plan to capture data from multiple perspectives to ensure diversity across different scenarios. Finally, the dataset contains 51% small objects, providing a sufficient number of samples for training.

4.2. Statistical Analysis

The EvDET200K dataset comprises 10,054 video streams, annotated with 10 common object categories, totally 202,260 annotations. As shown in the upper left image of Fig. 3, the most annotated class is “people”, with a total of 105,265 annotations. This is followed by “goose” and “car”, which have 30,960 and 25,850 annotations, respectively. Among them, 2,949 videos are taken from dense scenes. Each video has a duration ranging from 2 to 5 seconds. The dataset is randomly divided into training, validation, and test subsets using the ratio 6:1:3, which contains 6,031, 1,002, and 3,021 video streams, respectively.

4.3. Benchmark Baselines

To establish a comprehensive benchmark dataset for event-based object detection, we have selected more than 15 SOTA or representative detectors for evaluation on our proposed dataset, including: **1). CNN-based Detectors:** DetectoRS [30], RED [29], Mask R-CNN [12], RetinaNet [18], and Faster R-CNN [31], all using ResNet-50 as the backbone, as well as YOLO-style detectors like YOLO v10 [35], YOLO v6 [16]. **2). Transformer-based Detectors:** S5-ViT [54], SAST [28], RVT [10], Swin-T [22], DETR [3], and vHeat [42]. **3). SNN-based Detectors:**

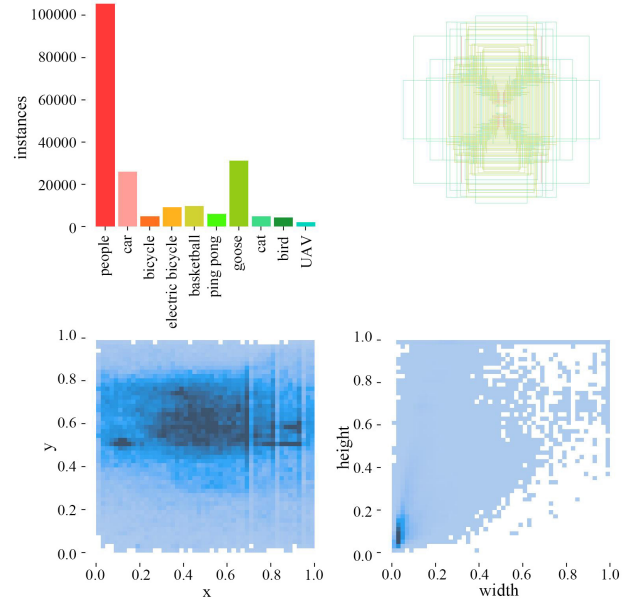


Figure 3. **Visualization of All Annotation Information in the Dataset.** Top Left (Instance Count per Class) shows the number of instances for each class in the whole dataset; Top Right (Bounding Box Size Distribution) illustrates the distribution of bounding box sizes across the dataset; Bottom Left (Object Center Distribution) shows the relative position (x, y coordinates) of object centers within the images. Bottom Right (Aspect Ratio Distribution) displays the distribution of width-to-height ratios of objects in the dataset.

Table 1. Experimental results on the N-Caltech101 dataset.

Methods	Format	mAP
NvS [17]	Event Points	34.6
YOLO [2]	Event Frames	39.8
Jeziorek et al. [14]	Event Frames	53.4
EAS-SNN [43]	Event Points	53.8
Ours	Event Frames	55.7

Spiking neural network (SNN) detectors, such as spikeYOLO [24] and EMS-YOLO [33], are also included for their event-based data processing capability.

5. Experiments

5.1. Dataset and Evaluation Metric

In addition to the newly proposed EvDET200K dataset, we also conducted a comparison with several state-of-the-art detectors on the N-Caltech101 [26] dataset to validate the generalization capability of our method. The N-Caltech dataset contains 101 object categories and approximately 9,000 event streams, which are split into training and test sets in an 8:2 ratio. This dataset features complex and vari-

Table 2. Experimental results on the newly proposed EvDET200K benchmark dataset. $vHeat^\dagger$ means using $vHeat$ as the encoder and Transformer as the decoder.

Index	Algorithm	Publish	Backbone	mAP@50:95	mAP@50	mAP@75	P	R	Params	FLOPs	FPS	Code
01	Faster R-CNN [31]	TPAMI 2016	ResNet50	46.0	73.3	48.6	76.3	88.3	40.9M	71.2G	23	URL
02	S5-ViT [54]	CVPR 2024	Former+SSM	42.9	76.3	44.1	69.7	66.9	18.2M	5.6G	84	URL
03	SAST [28]	CVPR 2024	Transformer	27.4	53.6	25.3	30.1	29.6	18.5M	15.9G	51	URL
04	SpikeYOLO [24]	ECCV 2024	SNN	41.2	74.8	39.8	81.6	68.5	68.8M	78.1G	77	URL
05	YOLOv10-N [35]	arXiv 2024	CNN	42.7	75.1	42.1	75.3	68.9	2.3M	8.2G	116	URL
	YOLOv10-S [35]			43.5	76.2	43.0	75.4	71.2	7.3M	21.6G	83	
	YOLOv10-M [35]			44.0	77.5	42.8	76.6	71.7	15.4M	59.1G	32	
	YOLOv10-B [35]			44.1	77.9	43.1	76.0	73.2	19.1M	92.0G	30	
06	RVT [10]	CVPR 2023	Transformer	40.7	73.1	42.3	70.3	65.9	9.9M	8.4G	88	URL
07	EMS-YOLO [33]	ICCV 2023	SNN	32.1	66.6	27.4	77.5	62.5	14.40M	3.3M	119	URL
08	YOLOv6 [16]	arXiv 2022	RepVGG	41.3	75.7	38.9	50.4	53.8	17.2M	44.2M	70	URL
09	Swin-T [22]	ICCV 2021	Transformer	49.0	78.4	52.7	79.4	88.8	160M	1043G	26	URL
10	DectectRS [30]	CVPR 2021	ResNet50	49.1	78.8	53.5	78.8	85.8	123.2M	117.2G	32	URL
11	RED [29]	NeurIPS 2020	ResNet50	35.4	68.3	35.2	69.0	66.1	24.1M	46.3G	34	URL
12	DETR [3]	ECCV 2020	Transformer	40.9	74.5	39.5	74.5	90.9	41M	86G	29	URL
13	Mask R-CNN [12]	ICCV 2017	ResNet50	48.8	77.6	52.0	77.8	87.1	43.8M	142.7G	28	URL
14	RetinaNet [18]	ICCV 2017	ResNet50	48.6	77.0	50.8	77.8	93.7	36.2M	81.4G	76	URL
15	$vHeat^\dagger$ [42]	arXiv 2024	$vHeat$	50.3	72.2	54.2	56.9	69.4	56.3M	74.5G	50	URL
16	Ours	-	MvHeat	52.9	80.4	55.9	58.9	70.0	47.5M	56.4G	58	-

able backgrounds, which present significant challenges for detection algorithms. For evaluation metrics, we used the mean Average Precision (mAP) at different IoU thresholds, the most commonly used metric in object detection. We also report Precision and Recall to assess the accuracy of predictions and the ability to detect positive instances. Additionally, we measured the number of parameters, FLOPs, and FPS for each detector, providing a more comprehensive and accurate understanding of the models' performance.

5.2. Compare With other Detectors

Results on EvDET200K Dataset. As shown in Tab. 2, our baseline $vHeat^\dagger$ achieves 50.3/72.2/54.2 on mAP/mAP@50/mAP@75, meanwhile, our model MvHeat-DET achieves 52.9/80.4/55.9, which is significantly better than baseline. Obviously, our detector is also better than other SOTA detectors including R-CNN based methods, YOLO-based detectors, SNN-based detectors and Transformer-based methods. This result strongly demonstrates the effectiveness of our method.

Results on N-Caltech Dataset. Tab. 1 presents experimental results on the N-Caltech101 dataset, comparing methods in terms of mean Average Precision (mAP). Among the methods listed, NvS and EAS-SNN (using event points) achieve 34.6/53.8, while YOLE and Jeziorek et al. (both using event frames) achieve mAPs of 39.8 and 53.4, respectively. Our model MvHeat-DET outperforms all others with a mAP of 55.7.

5.3. Component Analysis

We use DETR as the base model for component analysis. IQS denotes adding an IoU-based query selection strategy to the base model, $vEnc.$ indicates replacing the transformer

Table 3. Component Analysis on Our Proposed EvDET200K dataset. IQS means adding an IoU-based Query selection strategy, $vEnc.$ means replace the encoder with $vHeat$ Encoder, MoE means add MoE strategy to encoder.

Index	Baseline	IQS	$vEnc.$	MoE	mAP
1	✓				40.9
2	✓	✓			41.6
3	✓	✓	✓		50.3
4	✓	✓	✓	✓	52.9

encoder with $vHeat$, and MoE represents using a multi-expert strategy. Tab. 3 shows that the base model achieves 40.9 mAP on the EvDET200K dataset. Adding IQS to optimize query selection raises the detection result to 41.6. Next, replacing DETR's encoder with $vHeat$ significantly improves accuracy, reaching 50.3. Finally, the introduction of a multi-expert mechanism further boosts the model to 52.9. With each additional component, the mAP steadily increases, indicating that each component contributes to the model's performance. The notable improvements from IQS and $vHeat$ suggest that these components significantly enhance feature extraction and information processing capabilities. The addition of MoE also further refines the model's performance.

5.4. Ablation Study

Analysis on Number of MHCO in Each Stage. Tab. 4 shows the effect of varying the number of MHCO modules on model performance, computational complexity, and parameter count. We set the input event stream resolution to 640×640 px and varied the number of MHCO modules in the third stage. Specifically, increasing the number of

Table 4. Ablation studies on Number of MHCO in Each Stage.

Number of MHCO	(2,2,6,2)	(2,2,12,2)	(2,2,18,2)	(2,2,24,2)
mAP	52.6	52.9	52.9	53.4
FLOPs	39.3G	56.4G	73.4G	90.5G
Param	36.2M	47.5M	65.1M	70.1MG

Table 5. Ablation studies on the number of experts.

Number of Expert	mAP
1 (DCT)	50.3
2 (DCT+DFT)	52.7
3 (DCT+DFT+HT)	52.9

MHCO modules had a positive impact on the experimental results (6 layers: 52.6, 12 layers: 52.9, 18 layers: 52.9, 24 layers: 53.4). While increasing the number of MHCO modules can improve model accuracy, both FLOPs and the number of parameters grow significantly, leading to a substantial rise in computational and storage demands. We choose the (2, 2, 12, 2) configuration for building the experimental model to achieve a balance between model accuracy and computational efficiency while managing resource consumption.

Analysis on Number of Expert. In this section, we investigate the impact of varying the number of experts on the experimental results. As shown in Tab. 5, we select three transformations (DCT, DFT, and HT) as the experts. As the number of experts increases, the mAP gradually improves, rising from 50.3 with one expert (DCT) to 52.7 with two experts (DCT + DFT), and further to 52.9 with three experts (DCT + DFT + HT). This indicates that increasing the number of experts helps enhance model performance, primarily because different types of experts provide diverse feature extraction capabilities. It also demonstrates the effectiveness of the proposed MHCO module.

Analysis on Different Experts Combinations. Tab. 6 presents an ablation study on the performance of different frequency transformation combinations. When used individually, DFT, DCT, and HT achieve mAP scores of 48.8, 50.3 and 50.6 respectively. Notably, combining two transformations yields consistent improvements, reaching up to 52.7 mAP. The full combination of three experts achieves optimal performance with 52.9 mAP, 80.4 mAP@50, and 55.9 mAP@75, demonstrating the complementary nature of these experts.

Analysis on thermal diffusivity k . The Tab. 7 presents experimental results on different settings for the thermal diffusivity k , evaluating its impact on model performance (mAP). The experiment is conducted using a model with layer configurations of (2, 2, 12, 2). When k is fixed, the model achieves the mAP of 48.2. When k is treated as a learnable parameter, the mAP increases to 49.1, indicating that allowing the model to automatically learn k im-

Table 6. Ablation studies on different expert combinations.

NO.	DFT	DCT	HT	mAP	mAP@50	mAP@75
1	✓			48.8	71.5	47.6
2		✓		50.3	71.9	48.8
3			✓	50.6	78.8	52.8
4	✓	✓		52.7	80.1	55.8
5	✓		✓	52.4	80.0	55.1
6		✓	✓	51.9	79.6	54.9
7	✓	✓	✓	52.9	80.4	55.9

Table 7. Ablation studies on thermal diffusivity k .

Settings	mAP
Fixed $k = 5$	48.2
k as learnable parameter	49.1
Predicting k using FEs	49.7

proves performance to some extent. When k is predicted using FEs, the mAP further increases to 49.7, achieving the best performance. Dynamic learning and prediction of thermal diffusivity k , particularly through frequency embedding, significantly improve the model’s accuracy and allow it to better adapt to frequency feature variations in the input data.

[Note] More experimental results and visualizations can be found in our **Supplementary Materials** due to the limited space in this paper.

6. Conclusion

In this paper, we introduce a novel approach to event stream-based object detection, termed MvHeat-DET, which leverages a MoE-based heat conduction framework for efficient and interpretable feature extraction. This method balances performance, efficiency, and interoperability, offering a promising solution to the challenges faced by existing event-based detectors. Additionally, we propose a new high-definition dataset, EvDET200K, designed to advance research in this field by providing a comprehensive benchmark with diverse object categories and samples. We conclude by re-training more than 15 state-of-the-art object detectors on this dataset, paving the way for future advancements in event-based object detection.

Acknowledgment: This work is supported by the National Natural Science Foundation of China (No. 62102205, 62302041, 62332002, 62027804, 61825101); Anhui Provincial Natural Science Foundation 2408085Y032; Natural Science Foundation of Anhui Province 2408085J037; Key Technologies R & D Program of Anhui Province (202423k09020039); Major Key Project of PCL (PCL2021A13) and the project of Peng Cheng Laboratory (PCL2023A08). The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

- [1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.
- [2] Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [4] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2988–2997, 2021.
- [5] Pierre De Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.
- [7] Yimeng Fan, Wei Zhang, Changsong Liu, Mingyang Li, and Wenrui Lu. Sfod: Spiking fusion object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17191–17200, 2024.
- [8] Daniel Gehrig and Davide Scaramuzza. Pushing the limits of asynchronous graph-based object detection with event cameras. *arXiv preprint arXiv:2211.12324*, 2022.
- [9] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024.
- [10] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023.
- [11] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, page 1440. IEEE, 2015.
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [13] Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11270–11277, 2020.
- [14] Tomasz Kryjak. Optimising graph representation for hardware implementation of graph convolutional networks for event-based vision. *Design and Architectures for Signal and Image Processing*, page 110.
- [15] Trung-Nghia Le, Shintaro Ono, Akihiro Sugimoto, and Hiroshi Kawasaki. Attention r-cnn for accident detection. In *2020 IEEE intelligent vehicles symposium (IV)*, pages 313–320. IEEE, 2020.
- [16] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.
- [17] Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Graph-based asynchronous event processing for rapid object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 934–943, 2021.
- [18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [19] Bingde Liu, Chang Xu, Wen Yang, Huai Yu, and Lei Yu. Motion robust high-speed light-weighted object detection with event camera. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023.
- [20] Chang Liu, Yongsheng Yuan, Xin Chen, Huchuan Lu, and Dong Wang. Spatial-temporal initialization dilemma: towards realistic visual tracking. *Visual Intelligence*, 2(1):35, 2024.
- [21] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.
- [22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
- [24] Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. *arXiv preprint arXiv:2407.20708*, 2024.
- [25] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021.
- [26] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.
- [27] Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6038–6048, 2023.
- [28] Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based

- object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2024.
- [29] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.
- [30] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10213–10224, 2021.
- [31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- [32] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Sae-hoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.
- [33] Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, and Guoqi Li. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6555–6565, 2023.
- [34] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021.
- [35] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024.
- [36] Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4661–4670, 2021.
- [37] Xiao Wang, Yao Rong, Fuling Wang, Jianing Li, Lin Zhu, Bo Jiang, and Yaowei Wang. Event stream based sign language translation: A high-definition benchmark dataset and a new algorithm. *arXiv preprint arXiv:2408.10488*, 2024.
- [38] Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19248–19257, 2024.
- [39] Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5615–5623, 2024.
- [40] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2567–2575, 2022.
- [41] Zeyu Wang, Chen Li, Huiying Xu, and Xinzhong Zhu. Mamba yolo: Ssms-based yolo for object detection. *arXiv preprint arXiv:2406.05835*, 2024.
- [42] Zhaozhi Wang, Yue Liu, Yunfan Liu, Hongtian Yu, Yaowei Wang, Qixiang Ye, and Yunjie Tian. wheat: Building vision models upon heat conduction. *arXiv preprint arXiv:2405.16555*, 2024.
- [43] Ziming Wang, Ziling Wang, Huaning Li, Lang Qin, Runhao Jiang, De Ma, and Huajin Tang. Eas-snn: End-to-end adaptive sampling and representation for event-based detection with recurrent spiking neural networks. *arXiv preprint arXiv:2403.12574*, 2024.
- [44] Peilei Yan, Xuehu Liu, Pingping Zhang, and Huchuan Lu. Learning convolutional multi-level transformers for image-based person re-identification. *Visual Intelligence*, 1(1):24, 2023.
- [45] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.
- [46] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [47] Pengyu Zhang, Hao Yin, Zeren Wang, Wenyue Chen, Shengming Li, Dong Wang, Huchuan Lu, and Xu Jia. Evsign: Sign language recognition and translation with streaming events. In *European Conference on Computer Vision*, pages 335–351. Springer, 2025.
- [48] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detsr beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.
- [49] Aihua Zheng, Juncong Liu, Zi Wang, Lili Huang, Chenglong Li, and Bing Yin. Visible-infrared person re-identification via specific and shared representations learning. *Visual Intelligence*, 1(1):29, 2023.
- [50] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020.
- [51] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Ultra-high temporal resolution visual reconstruction from a fovea-like spike camera via spiking neuron model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1233–1249, 2022.
- [52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.
- [53] Zhuofan Zong, Guanglu Song, and Yu Liu. Detsr with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.
- [54] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5819–5828, 2024.