SCNet:Spatio-temporal Feature Aggregation and Cross-modal Interactive Encoding Network for DAVIS Object Detection

Yunhua Chen
Guangdong University of Technology
Guangzhou, China
yhchen@gdut.edu.cn

Jinyu Zhong
Guangdong University of Technology
Guangzhou, China
2112305312@mail2.gdut.edu.cn

Pinghua Chen
Guangdong University of Technology
Guangzhou, China
phchen@gdut.edu.cn

Wei Wu
Guangzhou Municipal Planning and
Natural Resources Automation Center
Guangzhou, China
723876878@qq.com

Jinsheng Xiao*
Wuhan University
Wuhan, China
xiaojs@whu.edu.cn

Abstract

DAVIS cameras, which output both event streams and frames simultaneously, are increasingly being used to address the primary object detection challenges posed by complex lighting and motion blur. Nevertheless, fully leveraging the abundant temporal information and effectively fusing data from these two modalities remains a formidable challenge. In this paper, we first design a multi-scale spatio-temporal aggregation (MSTA) module to distill richer semantic information from event frames. Secondly, we assimilate and harness the strengths of YOLOv8 and RT-DETR to develop an innovative encoder with Multi-scale Cross-modal dynamic Interactive fusion and multi-level feature interactive Fusion (MCIF). In MCIF, we propose a dynamic channel switching and spatial attention with learnable fusing factors (DCF-CSSA) to improve the complementary interaction of cross-modal features. Extensive experiments demonstrate that our approach (which we call SCNet) significantly outperforms existing state-of-the-art (SOTA) object detection methods that fuse events and frames, achieving an mAP50 improvement of 6.2% on PKU-DAVIS-SOD and 12% on DESC-MOD, both contain a large number of samples with challenging lighting conditions and motion blur.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Object detection**.

Keywords

Cross-Modal Fusion, Cross attention, Event-based Object Detection, Multimodal Object Detection

*Corresponding author.

ICMR '25, Chicago, IL, USA

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733428

ACM Reference Format:
Yunhua Chen, Jinyu Zhong, Pinghua Chen, Wei Wu, and Jinsheng Xiao. 2025. SCNet:Spatio-temporal Feature Aggregation and Cross-modal Interactive Encoding Network for DAVIS Object Detection. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3731715.3733428

1 Introduction

Object detection [39] is a challenging core task in the field of computer vision with a widespread of application scenarios, such as autonomous driving [2], video surveillance [1], and robot navigation [27]. Since object detectors based on single modal data [15, 31, 44, 47] can hardly achieve satisfactory performance in real scenes with challenging illumination and object motion, object detection based on multi-modal signal fusion [5, 9, 33, 35, 40, 42, 43] is a more realistic solution.

The Dynamic Active Pixel Vision Sensor (DAVIS) [3, 26] composed of a traditional frame-based camera and a novel bio-inspired event camera brings a new perspective to the field of multi-modal object detection. Event cameras[20, 25] generate event streams by triggering events in response to light changes in each single pixel, thus they have ultra-low latency of μs level and high dynamic range of up to 140 dB, which enables them to not only cope with the lighting and high-speed motion challenges faced by traditional frame-based cameras, but also provide rich temporal information that other cameras cannot provide. Nevertheless, event cameras also have their limitations: they are insensitive to static or extremely slow-moving objects and lack color and texture information, which can be supplemented by visible (RGB) images, but some key issues remain unsolved.

One issue is how to make full use of the rich temporal information of events. DAVIS object detection methods usually first segment the event stream into fixed time intervals and perform event aggregation [11, 22] to generate event frames aligned to image frames, and then input both frames into the subsequent backbone network to extract features. Recent studies mainly capture temporal dependencies by adding lateral recurrent connections [19, 29, 32] or turning to the Transformer architecture [10]. Although recurrent networks can capture long-term dependencies, they face challenges such as high computational cost, low parallelism, and difficulty in

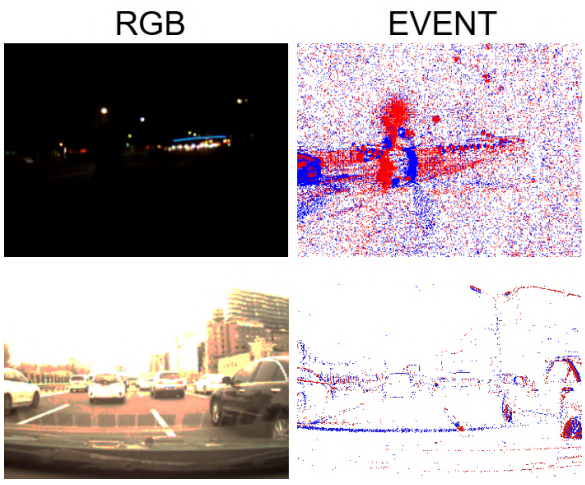

Figure 1: Samples from the PKU-DAVIS-SOD dataset under low-light conditions (first row) and slow-motion scenarios (second row). It can be observed that in different environments, one modality can complement the other, providing more detailed information.

training. The Transformer architecture is memory intensive and not good at capturing local temporal dependencies.

The second issue is how to effectively fuse the extracted features of events and RGB images. The fusion methods can be divided into two categories according to their strategies: decision-level fusion [6, 14, 18, 37] and feature-level fusion [4, 17, 23, 36, 46]. Decision-level fusion directly combines the detection results of the two modalities, lacks interaction between different modalities, and cannot completely eliminate the problems of single-modal detection. Feature-level fusion is mainly achieved in the feature extraction stage through feature concatenation, addition or attention mechanisms, which can achieve better interaction between different modalities. Nevertheless, most of the existing methods still suffer from problems such as insufficient cross-modal interaction, failure to fully utilize complementarity, and lack of interaction between low-level and high-level features.

In this paper, we propose a novel network that fully exploits the advantages of YOLOv8 [15] and RT-DETR [44] to achieve high-precision multi-modal object detection across events and images (SCNet) . As shown in Figure 2, SCNet consists of two key modules: Multi-scale Spatio-Temporal Aggregation (MSTA) and Multi-scale Cross-modal Interactive feature Fusion (MCIF) encoder. (1) In MSTA, we use 3D convolutions with different kernel sizes to efficiently capture the temporal dependencies between bins in parallel. Compared with Transformer-based and recurrent network-based architectures, MSTA can capture more fine-grained (bin-based vs. event frame-based) temporal dependencies, has higher computational efficiency and parallelism, and is easier to train. (2) In MCIF, we fuse the feature maps of the two modalities generated by the last three stages of the backbone network using our low-level feature fusion (LLFF) and high-level feature fusion (HLFF) modules, respectively. In LLFF, we design channel switching and feature fusion with learnable factors and combine them with spatial attention to achieve dynamic complementary feature fusion (DCF-CSSA) with improved interaction and complementarity between features of different modalities. In HLFF, we first capture the correlation between different modalities and perform feature enhancement through cross-modal cross attention and cross-modal self-attention (CMA), and then send the enhanced results to DCF-CSSA to get dynamic interactive complementary fusion results. Finally, PAFPN is used to perform the interactive fusion of high-level semantics and low-level features to further improve the encoding effect.

The main contributions can be summarized as:

- We propose a novel SCNet network that leverages the strengths of YOLOv8 and RT-DETR for multimodal object detection across events and images, and achieve 6.2% and 12% mAP50 improvements on two challenging datasets.
- We propose a dynamic channel switching and spatial attention with learnable mixing factors (DCF-CSSA) to improve the complementary interaction of cross-modal features.
- We design a plug-in module (MSTA) to effectively aggregates spatio-temporal information at different scales of bins to obtain fine-grained spatio-temporal features.

2 Related Works

2.1 Object Detection Based on Events

Thanks for the outstanding characteristics of event cameras, many event-based object detection methods have been proposed [10, 11, 13, 19, 22, 29]. The network grafting algorithm (NGA) [11] employed deep neural networks to extract features from event frames and achieved enhanced object detection performance. RED [29] proposed a recurrent network architecture that uses ConvLSTM [32] to extract spatio-temporal features from the event stream. ASTM-Net [19] introduced a temporal attention convolution module to learn event feature embeddings from continuous event streams, as well as a lightweight spatio-temporal memory module to extract temporal cues. RVT [10] proposed a Transformer-based object detection backbone that significantly reduces inference time while maintaining performance similar to previous works.

2.2 Object Detection Fusing Events and Images

Object detection by combining events and RGB images is a more reliable solution for challenging scenarios. According to different fusion strategies, it can be mainly divided into two solutions: decision-level fusion [6, 14, 18] and feature-level fusion [4, 17, 23, 36, 46].

In decision-level fusion, Chen et al. [6] used Non-Maximum Suppression (NMS) to merge detection results from both modalities. Li et al. [18] employed Dempster-Shafer theory to fuse detection results from the two modalities. Although Jiang et al. [14] proposed a confidence map to fuse the two modalities, these decision fusion methods lack interaction between modalities and cannot effectively utilize the complementary characteristics of the two modalities.

Recently, several feature fusion methods have been proposed to guide the fusion of the two modalities at the feature level. Liu et al. [23] calculated a channel attention map to guide the learning of event-based and image-based features. Cao et al. [4] generated pixel-level attention maps based on the features of the two modalities and multiplied them with image features to obtain fused features. Tomy et al. [36] used a simple concatenation operation to combine

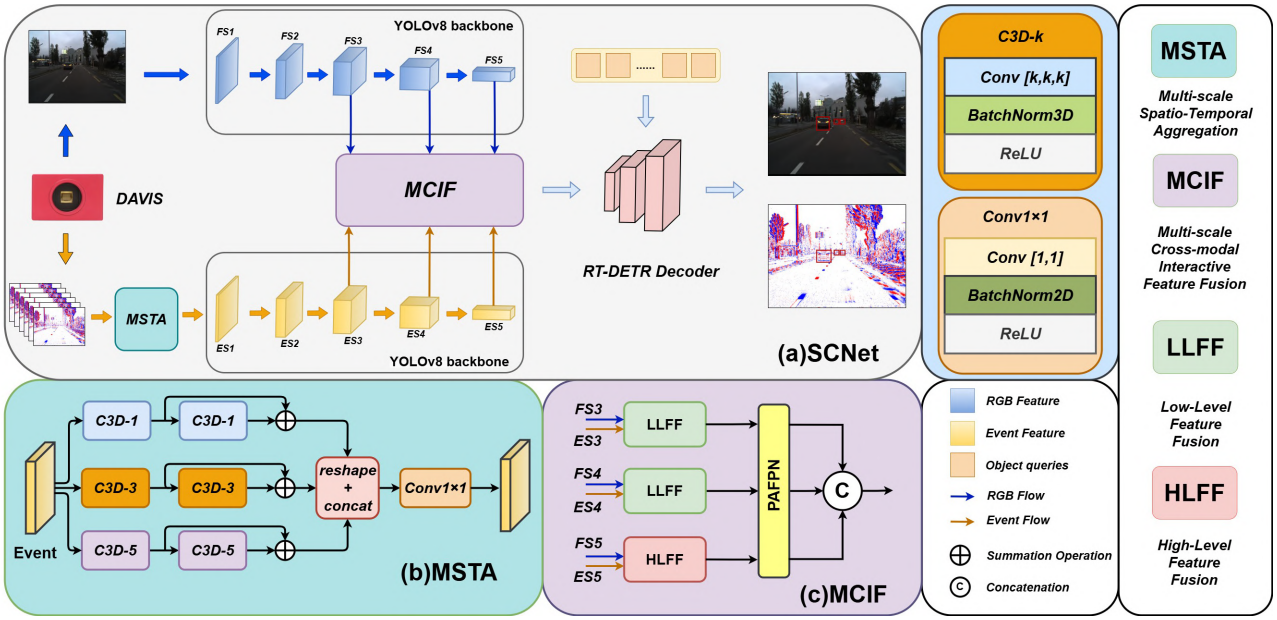

Figure 2: Overview of the proposed model. (a) The structure of SCNet. Event frames are first input into MSTA and then sent to the YOLOv8 backbone, while RGB images are directly sent to the YOLOv8 backbone. The features of the last three stages of the two backbone networks are comprehensively fused through the MCIF module. The RT-DETR decoder is employed to output the objects' classes and bounding boxes. (b) The structure of MSTA (see Section 5), which mainly consists of multiple 3D convolutions with different kernel sizes. (c) The structure of MCIF (see Section 3.4), which consists of two LLFFs (see Section 3.4.1), one HLFF (see Section 3.4.2), and one PAFPN.

features from the two modalities at different resolutions. Zhou et al. [46] proposed a bidirectional fusion module that models bimodal features in both spatial and channel dimensions to form a shared representation. Nevertheless, these methods do not fully utilize the spatio-temporal information in event streams. Li et al. [17] proposed a temporal Transformer model to leverage the rich spatio-temporal information from continuous events and adjacent frames. However, due to the use of a Transformer encoder, their inference speed is very slow.

3 Methodology

The primary objective of this work is to design a fast and efficient object detection network that integrates events and image frames. The overall framework of the network is shown in Figure 2(a).

3.1 Network Overview

Our network consists of four stages: event representation, feature extraction, cross-modal multi-scale feature fusion, and detection. In the first stage, to make asynchronous events compatible with traditional deep learning methods, we divide the continuous event stream into multiple time bins $B = \{B_1, B_2, B_3, ... B_n\}$. Each bin is converted into an event frame using a Time Surface-based event representation method. In the second stage, we use two YOLOv8 backbones [15] to extract features from the event frames and RGB images, respectively. Where we first input the event frames into an additional multi-scale spatio-temporal aggregation (MSTA) module to obtain finer-grained spatio-temporal features before inputting them into the backbone network. In the third stage, we design a multi-scale cross-modal interactive feature fusion (MCIF) encoder to comprehensively fuse the features generated by the last three stages of the backbone network. In the fourth stage, we use the decoder of RT-DETR [44] for detection, which does not require post-processing and offers higher accuracy.

3.2 Event Representation

When the logarithmic brightness change in the environment exceeds a specific threshold, each pixel in the event camera can independently trigger an event. An event typically contains positive polarity (p=+1) and negative polarity (p=-1), depending on the sign of the brightness change. This process can be represented as:

$$p = \begin{cases} +1 & L(x,y,t) - L(x,y,t-\Delta t) \geqslant C \\ -1 & L(x,y,t) - L(x,y,t-\Delta t) \leqslant -C \\ 0 & \text{other} \end{cases}. \quad (1)$$

Where (x, y) are pixel coordinates, t is a timestamp, Δt is a time interval since the last event at pixel (x, y), L represents logarithmic strength, C represents a preset threshold. Therefore, an event can be represented as a tuple (x, y, p, t).An event stream ε can be represented as:

$$\varepsilon = \{e_1 ... e_k ... e_N\}_{k=1}^{N} = \{(x_k, y_k, p_k, t_k)\}_{k=1}^{N}. \quad (2)$$

In modern event cameras, an event stream can contain up to 10 million events per second, making asynchronous event-by-event processing infeasible on traditional processing units. Most existing

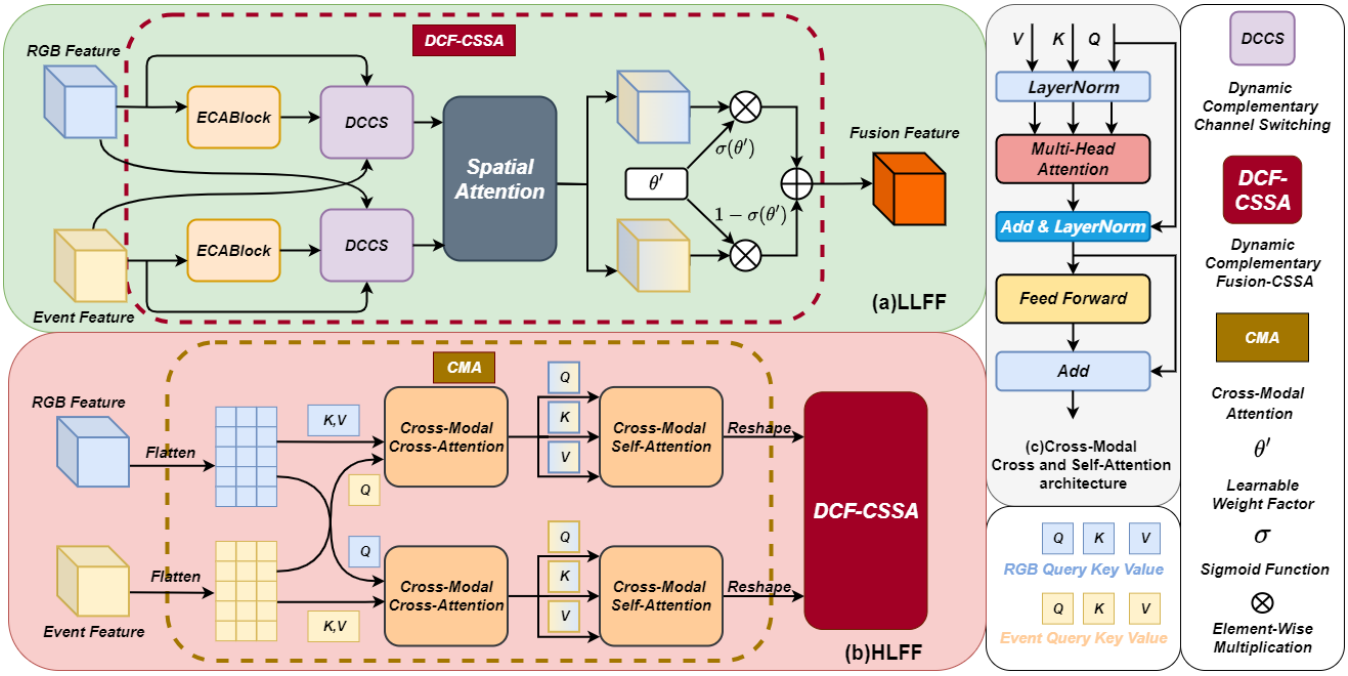

Figure 3: Structure of LLFF and HLFF. (a) LLFF. LLFF consists of a DCF-CSSA network, which contains two dynamic complementary channel switching (DCCS) blocks, a spatial attention block, and a dynamic addition block of feature maps. (b) HLFF. HLFF consists of two modules: the CMA and the DCF-CSSA module. (c) The detailed structure of Cross-modal Cross-attention and Cross-modal Self-attention.

methods convert asynchronous event streams into frame-like representations [16, 28, 41, 45] or use spike neural networks (SNNs) [7, 8] to process. Here, we use Time Surfaces (TS) as our event representation method, which can preserve temporal information while capturing the details of object movement. We first create a four-dimensional tensor TS. The first dimension consists of polarity, the second dimension consists of B components related to the division of the event stream into B time bins, and the third and fourth dimensions represent the height and width of the event camera. Thus, TS can be represented as:

$$TS = \{ts_1, ts_2, ts_3 \cdots ts_B\}, \tag{3}$$

$$ts_i = \sum_{e_k \in B_i} \delta(x - x_k, y - y_k)\delta(p - p_k)\left(e^{-\frac{t - t_{\text{end}}}{S}}\right), \tag{4}$$

where t_{end} is the end time of the current time bin, S is a time constant and $\delta(\cdot)$ refers to the Dirac delta function. The time of events contained in each time bin is in the interval $\left[t_0, \frac{t_N - t_0}{B} \cdot b\right]$, where t_0 is the start time of the current event stream and t_N is the end time of the current event stream, $b \in (0, B)$. In other words, we create B dual-channel time surfaces, resulting in an event representation in the shape of a four-dimensional tensor $(2, B, H, W)$. This four-dimensional tensor can be directly input into our MSTA module.

3.3 MSTA: Multi-scale Spatio-Temporal Aggregation

We design a module named MSTA, which can extract the spatio-temporal features of event frames and the temporal dependencies of event frames at different scales before they are fed into the backbone network.

The overall structure of MSTA is shown in Figure 2(b). It primarily uses 3D convolutions with different kernel sizes to perform convolution operations in both spatial and temporal dimensions, capturing the dynamic changes of multi-channel time surfaces and learning rich spatiotemporal features. First, we use 3D convolutions with kernel sizes of 1, 3, and 5 to aggregate and extract local spatiotemporal information at different scales. The 3D convolution with a kernel size of 1 captures spatiotemporal information at the scale of a single time bin, while the 3D convolutions with kernel sizes of 3 and 5 extract spatiotemporal features from adjacent time bins, capturing contextual information in both dimensions. We then use residual 3D convolutions with corresponding kernel sizes to further aggregate and extract spatiotemporal features. Finally, we reshape and concatenate the features at different scales and adjust the number of channels using a 2D convolution with a kernel size of 1, We express the computation of MSTA as follows:

$$TS'_k = C3D\text{-}k(TS), \tag{5}$$

$$\widetilde{TS'_k} = TS'_k + C3D\text{-}k\left(TS'_k\right), \tag{6}$$

$$\widetilde{TS} = Conv1{\times}1(RAC(\widetilde{TS'_1}, \widetilde{TS'_3}, \widetilde{TS'_5})), \tag{7}$$

where RAC represents the Reshape And Concat operation.

3.4 MCIF: Multi-scale Cross-modal Interactive Feature Fusion Encoder

Efficiently fusing event features and image features is a key step in integrating these two modalities for target detection. To comprehensively merge semantic information across different scales between the two modalities, we designed a multi-scale cross-modal interactive feature fusion encoder, as shown in Figure 2(c).

The entire module consists of three parts: low-level feature fusion, high-level feature fusion, and progressive asymmetric feature pyramid network (PAFPN) [12]. Specifically, we use two LLFFs to merge shallow features extracted by the backbone network ($ES3, FS3, ES4, FS4$), and a HLFF to merge deep features ($ES5, FS5$). For the three different scales of fused features obtained, we use PAFPN for multi-scale fusion to capture and utilize contextual information at different feature levels, enhancing the flow of information between features.

3.4.1 LLFF: Low-Level Feature Fusion. As shown in Figure 3(a), LLFF consists of a dynamic complementary fusion module with channel switching and spatial attention (DCF-CSSA), which is an improved version of CSSA [5] with learnable factors for channel switching and feature addition.

Cao et al. [5] proposed channel switching and spatial attention (CSSA) to perform multimodal object detection on infrared (IR) and visible light (RGB) images and achieved SOTA accuracy. In CSSA, when the weight of the i th channel of modality m is lower than a preset threshold k, the current channel is switched to the channel of another modality, otherwise the current channel is retained. CSSA enables better complementary interaction between the two modalities through channel switching. However, when the weight of the i th channel of modality m is lower than the preset threshold k, it does not mean that the weight of the switched modality m' is higher than the threshold k or is superior to modality m.

To address this issue, we designed a dynamic complementary channel switching (DCCS). By using a learnable mixing factor, the channels that need to be switched are adaptively weighted and fused with the channels of another modality, resulting in the current channel being switched to the fused result. This process can be described as:

$$\begin{cases} x_{m,i} & \text{if } w_{m,i} \geqslant k \\ x_{m,i} \times \sigma(\theta) + x_{m'i} \times (1 - \sigma(\theta)) & \text{if } w_{m,i} < k \end{cases}, \tag{8}$$

where $x_{m,i}$ is the i-th channel of modality m, $w_{m,i}$ is the weight of the i-th channel of modality m after passing through the ECA-Block [38], m' is another modality, σ is sigmoid function, and θ is a learnable factor for dynamic channel-mixing.

After the dynamic complementary channel switching, we apply spatial attention [5] as a supplement to our dynamic complementary channel switching. Finally, in the process of generating the final fusion result, we use learnable dynamic weight fusion instead of the average fusion of CSSA. This calculation process can be represented as:

$$X_{fused} = X_{RGB} \times \sigma(\theta') + X_{Event} \times \left(1 - \sigma(\theta')\right), \tag{9}$$

where X_{fused} represents the final fused feature, $X_{R\,G\,B}$ and X_{Event} are the feature maps after dynamic complementary channel switching and spatial attention, representing image features and temporal features, respectively. θ' is a learnable factor for dynamic weight fusion.

3.4.2 HLFF: High-Level Feature Fusion. The overall structure of HLFF is shown in Figure 3(b). Unlike LLFF, HLFF first passes through a cross-modal attention (CMA) and then through DCF-CSSA to obtain the fused features. This is because applying attention mechanisms to high-level features with richer semantic concepts is beneficial for object recognition and localization. In contrast, low-level features lack semantic concepts, making intra-scale interactions (cross and self-attention) unnecessary and potentially causing redundancy and confusion with high-level feature interactions. In CMA, cross-modal cross-attention is first performed to capture the correlations between different modalities and enhance feature associations. Then, cross-modal self-attention is applied to further smooth and coordinate the fused features, optimizing the features within a single modality after fusion. The architecture of cross-modal cross and self attention is shown in Figure 3(c). The calculation of HLFF can be represented as:

$$Q_E, K_E, V_E = \text{Flatten}\left(ES_5\right), \tag{10}$$

$$Q_F, K_F, V_F = \text{Flatten}\left(FS_5\right), \tag{11}$$

$$ES_5' = \text{Reshape}\left(SA\left(CA\left(Q_F, K_E, V_E\right)\right)\right), \tag{12}$$

$$FS_5' = \text{Reshape}\left(SA\left(CA\left(Q_E, K_F, V_F\right)\right)\right), \tag{13}$$

$$X_{fused} = DCF\text{-}CSSA\left(ES_5', FS_5'\right), \tag{14}$$

where CA stands for cross-modal cross-attention, and SA stands for cross-modal self-attention, Reshape represents restoring the shape of the flattened feature to the same shape as ES_5 and FS_5. The calculation process of MCIF can be represented as:

$$X_{fused}^3 = LLFF(ES_3, FS_3), \tag{15}$$

$$X_{fused}^4 = LLFF(ES_4, FS_4), \tag{16}$$

$$X_{fused}^5 = HLFF(ES_5, FS_5), \tag{17}$$

$$\Omega = PAFPN(X_{fused}^3, X_{fused}^4, X_{fused}^5), \tag{18}$$

where Ω represents the final fused result.

4 Experiments

4.1 Experiment Settings

Implementation details: For the PKU-DAVIS-SOD dataset, we adopt the AdamW optimizer for training, starting with a 3-epoch warm-up phase. During the warm-up, the learning rate quickly increases from a small value to 3e-4, and then gradually decreases to 3e-6 over a total of 35 epochs.

For the DSEC-MOD dataset, we also adopt the AdamW optimizer with a 3-epoch warm-up phase. In this phase, the learning rate rapidly increases to 4e-4 and then slowly decreases to 4e-6 over a total of 120 epochs.

In both datasets, the batch size is set to 16, and the image size is adjusted to 384×384. For event representation, the number of time bins is set to 6. We employ random horizontal flipping as our data augmentation method. The weights for classification loss, L1 loss,

Table 1: Comparison with SOTA. Our method outperforms existing methods on both datasets.

Modality	Method	Input represention	backbone	mAP50	
				PKU-DAVIS-SOD	DSEC-MOD
Events	SSD-event [13]	Event image	SSD	0.221	-
	NGA-event [11]	Voxel grid	YOLOv3	0.232	-
	YOLOv3-RGB [30]	Reconstructed image	YOLOv3	0.244	-
	Faster R-CNN [31]	Event image	R-CNN	0.251	-
	Deformable DETR [47]	Event image	DETR	0.307	-
	LSTM-SSD [24]	Event image	SSD	0.273	-
	ASTMNet [19]	Event embedding	Rec-Conv-SSD	0.291	-
	SODFormer [17]	Event image	Deformable DETR	0.334	-
	Our baseline	**Time surface**	**YOLOv8+RT-DETR**	**0.407**	**0.293**
Frames	Faster R-CNN [31]	RGB frame	R-CNN	0.443	-
	YOLOv3-RGB [30]	RGB frame	YOLOv3	0.426	-
	Deformable DETR [47]	RGB frame	DETR	0.461	-
	LSTM-SSD [24]	RGB frame	SSD	0.456	-
	SODFormer [17]	RGB frame	Deformable DETR	0.489	-
	Our baseline	**RGB frame**	**YOLOv8+RT-DETR**	**0.539**	**0.461**
Events+Frames	MFEPD [14]	Event image + RGB frame	YOLOv3	0.438	-
	JDF [18]	Channel image + RGB frame	YOLOv3	0.442	-
	SODFormer [17]	Event image + RGB frame	Deformable DETR	0.504	-
	FPN-Fusion [36]	Channel image + RGB frame	RetinaNet	-	0.323
	EFNet [34]	Channel image + RGB frame	UNet	-	0.353
	RENet [46]	Channel image + RGB frame	MOC-Detector	-	0.384
	Our SCNet	**Time surface + RGB frame**	**YOLOv8+RT-DETR**	**0.566**	**0.504**

and Giou loss are set to 2, 5, and 2, respectively. The weight decay of 1e-4. The hyper-parameter k mentioned in Equation 8 is set to 0.5, and the hyper-parameter S mentioned in Equation 4 is set to 50ms. All experiments are conducted on a single NVIDIA GeForce RTX 3090.

Evaluation Metrics: To compare different approaches, the mean average precision (e.g., COCO mAP [21]) and running time (ms) are selected as two evaluation metrics, which are the most broadly utilized in the object detection task.

4.2 Comparison with SOTA

We compare our method with nine event-based methods: SSD-event [13], NGA-event [11], YOLOv3-RGB [30], Faster R-CNN [31], Deformable DETR [47], LSTM-SSD [24], ASTMNet [19], SODFormer [17] and ours baseline; six frame-based methods: Faster R-CNN [31], YOLOv3-RGB [30], Deformable DETR [47], LSTM-SSD [24], SODFormer [17] and ours baseline; as well as six RGB-Event fusion-based object detection method: MFEPD [14], JDF [18], SODFormer [17], FPN-Fusion [36], EFNet [34] and RENet [46].

As shown in Table 1, our SCNet outperforms existing methods, including single-modality and RGB-Event fusion-based methods. For instance, in event-based methods, our baseline (i.e., SCNet without MSTA and MCIF) achieves a 7.3% higher mAP50 on the PKU-DAVIS-SOD dataset compared to the state-of-the-art (SOTA) method SODFormer. In RGB image-based methods, our baseline surpasses the SOTA by 5%. Furthermore, in methods based on the fusion of event and RGB images, our SCNet significantly outperforms existing SOTA methods, achieving a 6.2% higher mAP50 on the PKU-DAVIS-SOD dataset compared to SODFormer. Additionally, our SCNet achieves a 12% higher mAP50 on the DSEC-MOD dataset compared to RENet.

4.3 Performance Evaluation in Various Scenarios

We comprehensively evaluate our SCNet on the PKU-DAVIS-SOD dataset. We report quantitative results (see Table2) and representative visual results (see Figure4).

We observe that under normal conditions, our baseline using RGB frames outperforms the baseline using event data. This is because event cameras struggle to produce fine-grained textures in static or slow-motion scenarios (see Figure 4(a)), whereas RGB frames can provide detailed static textures. In motion blur scenarios, the detection performance using frames degrades more significantly than using events. This is because motion-blurred images cause details to become fuzzy (see Figure 4(b)), reducing the contrast between objects and the background, making it difficult to accurately locate and identify objects. Event-based methods, on the other hand, can capture motion information of objects, providing richer spatiotemporal features to address this issue.In low-light conditions, the performance of our baseline using RGB frames drops sharply by 21.7%, compared to a 12.8% drop for the event-based baseline. Methods using RGB frames struggle to detect all objects in low-light environments. This is because, under low-light conditions, the dynamic range of image sensors is limited, making it difficult to capture details in both bright and dark areas simultaneously, leading to degraded image quality (see Figure 4(c)).

Although event-based methods can accurately locate and identify objects in motion blur and low-light scenarios, they still struggle to detect all objects due to the lack of fine-grained texture information. This is particularly challenging when two small objects partially overlap. Event-based methods find it difficult to detect

Table 2: Performance evaluation of our method under various scenarios on the PKU-DAVIS-SOD dataset. Including event-based baseline, RGB image-based baseline, and RGB-Event fusion.

Scenario	Modality	AP50			mAP50	mAP50:95	Runtime(ms)
		Car	Pedestrain	Two-wheeler			
Normal	Events	0.54	0.287	0.515	0.447	0.213	5.6
	Frames	0.834	0.383	0.595	0.604	0.319	5.5
	Frames+Events	0.839	0.403	0.593	0.611	0.321	10.3
Motion blur	Events	0.403	0.209	0.445	0.352	0.16	5.7
	Frames	0.628	0.343	0.46	0.477	0.244	5.6
	Frames+Events	0.657	0.369	0.514	0.513	0.247	10.5
Low-light	Events	0.591	0.018	0.35	0.319	0.143	5.0
	Frames	0.631	0.179	0.315	0.387	0.167	5.0
	Frames+Events	0.665	0.195	0.444	0.435	0.187	10.6
All	Events	0.516	0.247	0.457	0.407	0.189	5.6
	Frames	0.781	0.35	0.488	0.539	0.279	5.5
	Frames+Events	0.792	0.375	0.531	0.566	0.288	10.8

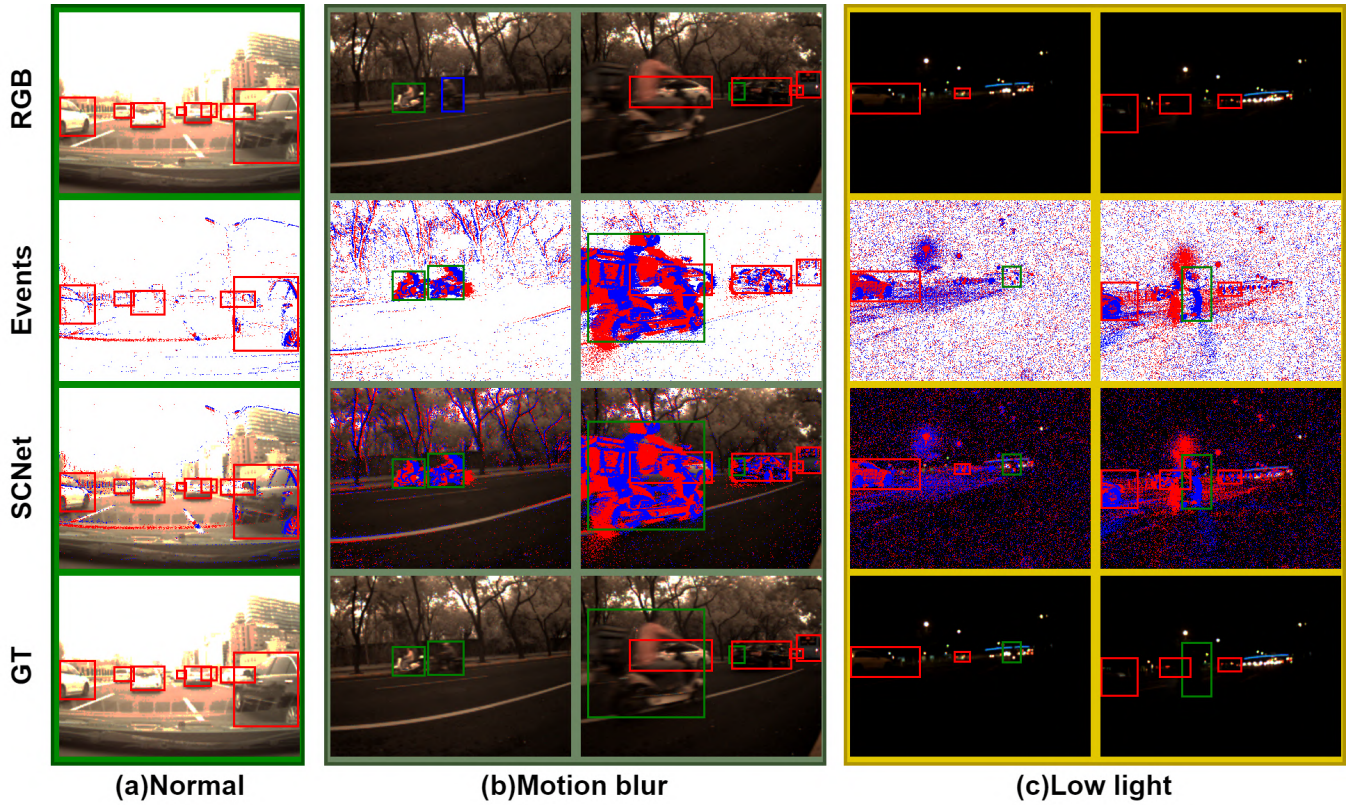

Figure 4: Visualization of detection results for various scenes in the PKU-DAVIS-SOD dataset. Cars, pedestrians, and two-wheelers are represented by red, blue, and green bounding boxes, respectively.

both objects. Therefore, our SCNet extracts rich spatiotemporal information from events through MSTA and efficiently fuses features from both modalities through MCIF, consistently outperforming single-modality methods in all three scenarios.

4.4 Ablation Study

Comparison With CSSA: To validate that our proposed DCF-CSSA outperforms CSSA, we conducted comparisons on our SCNet across two datasets. We replaced the DCF-CSSA in the HLFF and

Table 3: Comparison results of our DCF-CSSA and CSSA on two datasets.

Method	mAP50		Params(M)
	PKU-DAVIS-SOD	DSEC-MOD	
CSSA [4]	0.551	0.483	73.7
DCF-CSSA	0.559	0.495	73.7

Table 4: Effects of LLFF, HLFF and MSTA on two datasets. Ours Baseline* indicates the method using average fusion and without MSTA.

LLFF	HLFF	MSTF	mAP50		Params(M)
			PKU-DAVIS-SOD	DSEC-MOD	
Our Baseline*			0.548	0.476	65.3
✓			0.553	0.489	65.3
	✓		0.554	0.488	73.7
✓	✓		0.559	0.495	73.7
✓	✓	✓	0.566	0.504	73.7

Table 5: The impact of different combinations of three-dimensional convolutions (C3D) with varying kernel sizes in our MSTA module on the results of two datasets.

C3D-k			mAP50	
1	3	5	PKU-DAVIS-SOD	DSEC-MOD
✓			0.56	0.497
	✓		0.562	0.501
		✓	0.563	0.502
✓	✓		0.563	0.501
	✓	✓	0.564	0.503
✓	✓	✓	0.566	0.504

LLFF modules of the MCIF with CSSA, and neither used MSTA. As shown in Table 3, our proposed DCF-CSSA outperforms CSSA on both datasets. This is because our method fully considers the complementary information between the two modalities and dynamically fuses the features from both modalities.

Contribution of Our SCNet Components: Table 4 shows the results of different module combinations in our method. Compared to baseline (without MSTA and only PAFPN in MCIF module), we can observe that the fusion of low-level features and high-level features is crucial for providing good performance in the proposed SCNet. More specifically, LLFF can learn and provide the primary features of objects, including edges, textures, colors, and shapes, while HLFF can capture the global contextual information in the image, helping the model to process and recognize objects and relationships in complex scenes. By combining both, we can utilize low-level detail information and high-level semantic information simultaneously, improving the accuracy and robustness of object detection. Compared to not using MSTA, the accuracy of the model improved further after using MSTA, indicating that MSTA can extract and aggregate rich spatiotemporal information from events.

Further research on MSTA. Finally, we conducted an ablation study to examine the effects of different convolution kernel sizes and their various combinations on the network's performance. As shown in the table 5, convolution kernels of varying sizes all contributed to performance improvements. This enhancement is attributed to their ability to capture spatio-temporal information at different scales within the event flow. The 3D convolutional kernel with size 1×1×1 captures fine-grained spatial features within individual frames, while kernels of size 3×3×3 extract local motion patterns through short-term temporal cycles. Larger 5×5×5 kernels identify global dynamic correlations via long-range dependencies spanning multiple event frames.

5 Conclusion

In this paper, we propose a novel RGB-Event fusion architecture, SCNet, for object detection under challenging lighting and high dynamic range conditions. It consists of two key components: MSTA and MCIF. MSTA extracts and aggregates spatiotemporal information of events at different temporal dimensions, capturing the dynamic changes in the event stream. MCIF enables dynamic complementary fusion of the two modalities at different scales, leveraging the advantages of both events and RGB images to compensate for the missing information in each modality. We thoroughly validate our method on two existing large-scale datasets based on events and RGB images. Experimental results demonstrate that our method outperforms state-of-the-art (SOTA) methods by effectively utilizing the complementarity of the two modalities, enhancing the robustness of object detection under challenging lighting and high dynamic range conditions.

Acknowledgments

This work was supported by the Major Program (JD) of Hubei Province (2023BAA026), and the Natural Science Foundation of Guangdong Province, China (No: 2025A1515012243).

References

[1] Sani Abba, Ali Mohammed Bizi, Jeong-A Lee, Souley Bakouri, and Maria Liz Crespo. 2024. Real-time object detection, tracking, and monitoring framework for security surveillance systems. *Heliyon* 10, 15 (2024).

[2] Daniel Bogdoll, Maximilian Nitsche, and J Marius Zöllner. 2022. Anomaly detection in autonomous driving: A survey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4488–4499.

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

[4] Hu Cao, Guang Chen, Jiahao Xia, Genghang Zhuang, and Alois Knoll. 2021. Fusion-based feature attention gate component for vehicle detection based on event camera. *IEEE Sensors Journal* 21, 21 (2021), 24540–24548.

[5] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. 2023. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 403–411.

[6] Nicholas FY Chen. 2018. Pseudo-labels for supervised learning on dynamic vision sensor data, applied to object detection under ego-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 644–653.

[7] Yunhua Chen, Weijie Bai, Qingkun Huang, and Jinsheng Xiao. 2021. Efficient Motion Symbol Detection and Multikernel Learning for AER Object Recognition. *IEEE Transactions on Cognitive and Developmental Systems* 14, 4 (2021), 1544–1552.

[8] Yunhua Chen, Ren Feng, Zhimin Xiong, Jinsheng Xiao, and Jian K Liu. 2024. High-performance deep spiking neural networks via at-most-two-spike exponential coding. *Neural Networks* 176 (2024), 106346.

[9] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. 2022. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2022), 12878–12895.

[10] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. 2023. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*. PMLR, 694–710.

[11] Yuhuang Hu, Tobi Delbruck, and Shih-Chii Liu. 2020. Learning to exploit multiple vision modalities by using grafted networks. In *European Conference on Computer Vision*. Springer, 85–101.
[12] Muhammad Hussain. 2023. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines* 11, 7 (2023), 677.
[13] Massimiliano Iacono, Stefan Weber, Arren Glover, and Chiara Bartolozzi. 2018. Towards event-driven object detection with off-the-shelf deep learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1–9.
[14] Zhuangyi Jiang, Pengfei Xia, Kai Huang, Walter Stechele, Guang Chen, Zhenshan Bing, and Alois Knoll. 2019. Mixed frame-/event-driven fast pedestrian detection. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8332–8338.
[15] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics
[16] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. 2016. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence* 39, 7 (2016), 1346–1359.
[17] Dianze Li, Yonghong Tian, and Jianing Li. 2023. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 14020–14037.
[18] Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. 2019. Event-based vision enhanced: A joint detection framework in autonomous driving. In *2019 ieee international conference on multimedia and expo (icme)*. IEEE, 1396–1401.
[19] Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. 2022. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing* 31 (2022), 2975–2987.
[20] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. 2008. A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits* 43, 2 (2008), 566–576.
[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.
[22] Bingde Liu, Chang Xu, Wen Yang, Huai Yu, and Lei Yu. 2023. Motion robust high-speed light-weighted object detection with event camera. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–13.
[23] Mengyun Liu, Na Qi, Yunhui Shi, and Baocai Yin. 2021. An attention fusion network for event-based vehicle object detection. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3363–3367.
[24] Mason Liu and Menglong Zhu. 2018. Mobile video object detection with temporally-aware feature maps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5686–5695.
[25] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. 2022. Multi-bracket high dynamic range imaging with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 547–557.
[26] Diederik Paul Moeys, Federico Corradi, Chenghan Li, Simeon A Bamford, Luca Longinotti, Fabian F Voigt, Stewart Berry, Gemma Taverni, Fritjof Helmchen, and Tobi Delbruck. 2017. A sensitive dynamic and active pixel vision sensor for color or neural imaging applications. *IEEE transactions on biomedical circuits and systems* 12, 1 (2017), 123–136.
[27] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *The International Journal of Robotics Research* 36, 2 (2017), 142–149.
[28] Yansong Peng, Yueyi Zhang, Peilin Xiao, Xiaoyan Sun, and Feng Wu. 2023. Better and faster: Adaptive event conversion for event-based object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2056–2064.
[29] Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. 2020. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems* 33 (2020), 16639–16652.
[30] Joseph Redmon. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).
[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.
[32] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015).
[33] Rahman Soroush and Yasser Baleghi. 2023. NIR/RGB image fusion for scene classification using deep neural networks. *The Visual Computer* 39, 7 (2023), 2725–2739.
[34] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. 2022. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*. Springer, 412–428.
[35] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 10 (2022), 6700–6713.
[36] Abhishek Tomy, Anshul Paigwar, Khushdeep S Mann, Alessandro Renzaglia, and Christian Laugier. 2022. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 933–939.
[37] Fei-yue Wang. 2023. Drive like a machine: Remembering the origin and goal of autonomous driving and intelligent vehicles. *IEEE Transactions on Intelligent Vehicles* (2023).
[38] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11534–11542.
[39] Jinsheng Xiao, Haowen Guo, Jian Zhou, Tao Zhao, Qiuze Yu, Yunhua Chen, and Zhongyuan Wang. 2023. Tiny object detection with context enhancement and feature purification. *Expert Systems with Applications* 211 (2023), 118665.
[40] Jinsheng Xiao, Shurui Wang, Jian Zhou, Ziyue Tian, Hongping Zhang, and Yuan-Fang Wang. 2025. MIM: High-Definition Maps Incorporated Multi-View 3D Object Detection. *IEEE Transactions on Intelligent Transportation Systems* 26, 3 (2025), 3989–4001.
[41] Bochen Xie, Yongjian Deng, Zhanpeng Shao, Qingsong Xu, and Youfu Li. 2024. Event voxel set transformer for spatiotemporal representation learning on event streams. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
[42] Jiacheng Ying, Can Tong, Zehua Sheng, Bowen Yao, Si-Yuan Cao, Heng Yu, and Hui-Liang Shen. 2023. Region-aware RGB and near-infrared image fusion. *Pattern Recognition* 142 (2023), 109717.
[43] Hongcheng Zhang, Liu Liang, Pengxin Zeng, Xiao Song, and Zhe Wang. 2025. SparseLIF: High-Performance Sparse LiDAR-Camera Fusion for 3D Object Detection. In *European Conference on Computer Vision*. Springer, 109–128.
[44] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16965–16974.
[45] Jinyu Zhong, Weiming Zeng, Yunhua Chen, Jinsheng Xiao, and Irwin King. 2024. Efficient Spatio-temporal Event Representation Based on Kalman Filtering and Linear Weighted Timestamps. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.
[46] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. 2023. Rgb-event fusion for moving object detection in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7808–7815.
[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).