Channel-based complementary fusion network for RGB-Event sensing and moving-object detection

1st Ziqiang Ye
College of Information Science and Technology
Beijing University of Technology
Beijing, China
yzq4964@gmail.com

2nd Erjun Xiao
Center for Excellence in Brain Science and Intelligence Technology
Chinese Academy of Sciences
Shanghai, China
xej0106@163.com

3rd Liyuan Han
Center for Excellence in Brain Science and Intelligence Technology
Chinese Academy of Sciences
Shanghai, China
hanliyuan@ion.ac.cn

4th Tielin Zhang*
Center for Excellence in Brain Science and Intelligence Technology
Chinese Academy of Sciences
Shanghai, China
zhangtielin@ion.ac.cn

5th Kebin Jia*
College of Information Science and Technology
Beijing University of Technology
Beijing, China
kebinj@bjut.edu.cn

***Abstract*—Moving object detection (MOD) is essential for autonomous driving. Although conventional frame-based methods can achieve high detection accuracy under favorable illumination conditions, challenges remain in low-light, overexposed, and fast-motion scenarios. Dynamic vision sensors (DVS), also known as event cameras, provide high temporal resolution and high dynamic range, making them complementary to RGB cameras. In this paper, we propose a channel-based complementary fusion network, termed CCFNet, to jointly process RGB frames and event signals for moving object detection. The proposed CCFNet mainly introduces a Bidirectional Channel Interaction Module (BCIM), which consists of: 1) a channel switching mechanism that estimates channel-wise importance and performs bidirectional cross-modal channel exchange to enhance complementary information; and 2) a spatial attention refinement mechanism that reweights fused features in the spatial dimension to highlight informative object regions and suppress redundant background responses. Experimental results on RGB-Event moving object detection datasets show that CCFNet achieves competitive performance compared with representative fusion methods and improves robustness under challenging illumination conditions.**

***Index Terms*—Moving object detection, RGB-Event fusion, Multimodal fusion, Channel fusion**

I. Introduction

Moving object detection (MOD) plays an important role in autonomous driving and intelligent transportation systems [1]. Although deep learning-based detection methods have achieved significant progress, most existing approaches rely on conventional frame-based cameras, whose performance can degrade under challenging conditions such as low light, overexposure, and fast motion [2], [3]. This is because frame-based cameras are limited by fixed frame rates and are highly dependent on image quality, making it difficult to capture reliable motion and appearance information in complex driving environments [4], [5].

Event cameras have recently attracted increasing attention as a complementary sensing modality for visual perception [6]. Unlike conventional cameras, event cameras asynchronously record pixel-level brightness changes and output sparse event streams [7]. Owing to their high temporal resolution and high dynamic range, event cameras are robust to fast motion and illumination changes [8]. However, they usually provide limited texture and semantic information, and their responses can become weak in static or extremely slow-motion scenes [9]. Therefore, RGB images and event streams are naturally complementary, and their effective fusion is crucial for robust moving object detection.

Existing RGB-Event fusion methods usually combine features by direct concatenation, summation, or attention-based interaction [10], [11]. However, these strategies may not fully address the modality discrepancy between RGB frames and event streams. On the one hand, the reliability of the two modalities varies across different illumination and motion conditions. On the other hand, event data and RGB images have different temporal and spatial distributions, which may lead to feature misalignment and modality bias during direct

fusion [12], [13]. Therefore, it is necessary to adaptively evaluate the importance of modality-specific features and exploit their complementary information in both channel and spatial dimensions.

To address these issues, we propose a channel-based complementary fusion network, termed CCFNet, for RGB-Event moving object detection. The core component of CCFNet is a Bidirectional Channel Interaction Module (BCIM), which contains a channel switching mechanism and a spatial attention refinement mechanism. The channel switching mechanism evaluates channel-wise importance and exchanges complementary features between RGB and event modalities, while the spatial attention refinement mechanism further emphasizes informative regions such as motion boundaries and object contours. In this way, CCFNet enhances cross-modal feature interaction and improves the robustness of multimodal detection.

In summary, the main contributions of this paper are as follows:

- We propose a channel switching module that enables deep fusion between RGB and event features by evaluating channel-wise importance and exchanging complementary features to improve cross-modal information interaction.
- We design a spatial attention fusion mechanism that further integrates complementary features in the spatial dimension, improving the accuracy of target localization.
- We conduct comparative experiments, ablation studies, threshold sensitivity analysis, and illumination-condition analysis on RGB-Event moving object detection datasets to evaluate the effectiveness and robustness of the proposed fusion strategy.

II. Method

This section first introduces the event camera representation. Then, the overall architecture of CCFNet is described. Finally, the proposed Bidirectional Channel Interaction Module is presented in detail.

A. Event Camera Representation

The event camera responds to the light intensity change $R(u,t)$ of each individual pixel and outputs it as an event stream. Specifically, an event e_n is a quadruple (x_n, y_n, p_n, t_n), represented using Address Event Representation (AER). For a pixel $u = (x_n, y_n)$, an event is triggered at the timestamp t_n when the logarithmic change in light intensity exceeds a preset threshold θ_{th}. This process can be expressed as:

$$\ln R\left(u_n, t_n\right) - \ln R\left(u_n, t_n - \Delta t_n\right) = p_n \theta_{th}, \tag{1}$$

where Δt_n is the time sampling interval at the pixel location, and polarity p indicates whether the light intensity is decreasing or increasing. Intuitively, asynchronous events manifest as sparse and discrete points in the spatiotemporal domain. This can be expressed as:

$$S(x,y,t) = \sum_{n=1}^{N_c} p_n \delta\left(x - x_n, y - y_n, t - t_n\right), \tag{2}$$

where N_c is the number of events within the time window. The Dirac delta function, denoted as $\delta(\cdot)$, satisfies $\int \delta(t)dt = 1$ and $\delta(t) = 0$ for $t \neq 0$.

B. Framework Overview

This section details the network architecture of the Channel-exchange-based Complementary Fusion Network (CCFNet), which is designed to fuse bimodal RGB-Event data for moving object detection. The model is designed to alleviate RGB feature degradation caused by sudden illumination changes.

As illustrated in Figure 1, CCFNet takes multi-temporal frames from both an RGB camera and an event camera as input. It incorporates the Event-based Temporal Multi-scale Aggregation (E-TMA) module [15] to leverage multi-scale temporal information from the event stream.

The network first employs a lightweight bimodal feature extraction structure, using CSPDarkNet [14] as the backbone network to process inputs from both modalities. To mitigate the information loss in the RGB modality under challenging environmental conditions, an Adaptive Feature Completion Module (AFCM) [10] is adopted to enable explicit feature interaction between the two modalities. This module adaptively supplements missing information in the RGB feature maps caused by environmental disturbances using features from the event modality. Subsequently, the adaptively complemented bimodal features are fed into a Bidirectional Channel Interaction Module (BCIM) for effective cross-modal feature fusion. The fused features are further enhanced by a Feature Pyramid Network (FPN) [16] for multi-scale feature extraction, and finally, the detection head decodes the output to generate bounding box information for each detected moving target.

C. Bidirectional Channel Interaction Module

RGB-Event feature fusion is a relatively unexplored area compared to other multimodal fusion tasks such as RGB-Depth or RGB-Point Cloud. Existing studies often rely on straightforward concatenation or convolution-based fusion, or alternatively leverage transformer attention that primarily emphasizes the channel dimension. However, these approaches tend to overlook the crucial spatial cues that are indispensable for tasks such as detection and segmentation. To address this limitation, we propose a Bidirectional Channel Interaction Module (BCIM), as illustrated in Fig. 2. BCIM is designed to enable cross-modal interaction in both channel and spatial dimensions. Specifically, it incorporates two stages: channel switching and spatial attention refinement. This design ensures that the two modalities can exchange complementary information during the encoding stage while preserving modality-specific discriminative cues.

For simplicity, we take the highest-level fusion as an example. Formally, let the RGB features be denoted as $f_R \in \mathbb{R}^{C\times h\times w}$ and the event features $f_E \in \mathbb{R}^{C\times h\times w}$. Channel exchange is first performed, followed by spatial attention refinement. Specifically, we first use element-wise multiplication and addition to mutually enhance the most informative common features in the dual modalities. This preliminary fusion

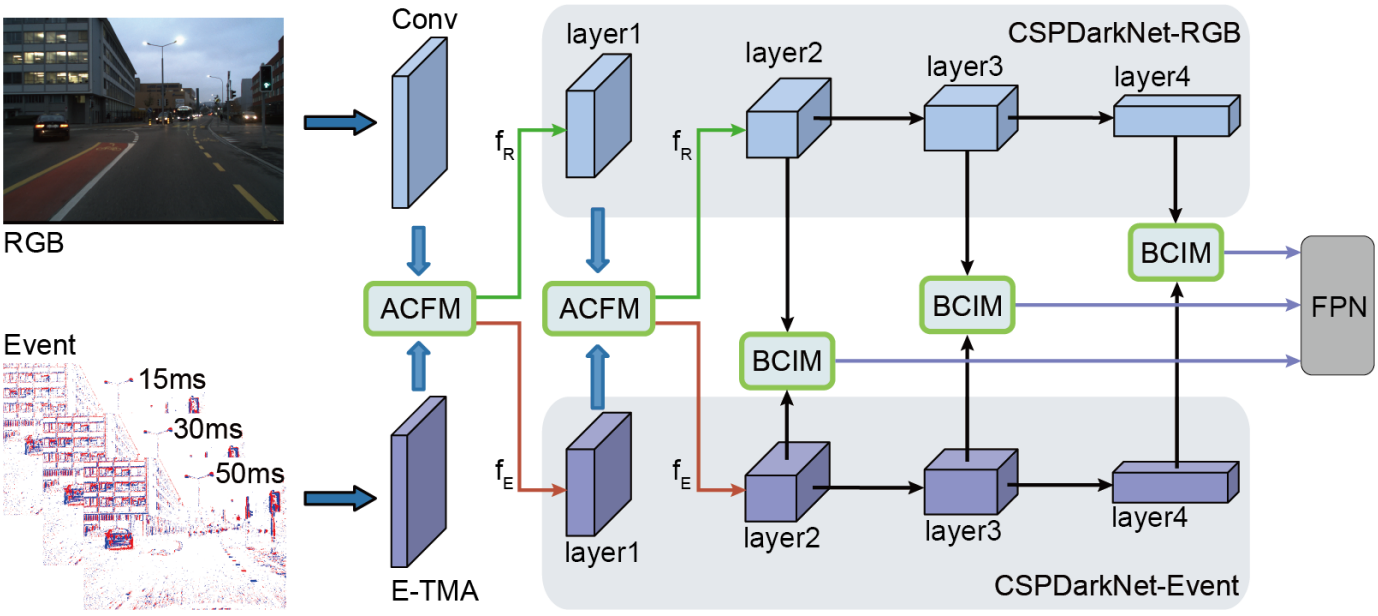

Fig. 1. The overall architecture of the proposed CCFNet. The network is built upon a CSPDarkNet [14] backbone for feature extraction, incorporates a bidirectional channel interaction fusion module, and utilizes an Adaptive Feature Completion Module (AFCM) adapted from [10]. The decoding and loss functions are also derived from [15].

step enhances shared informative responses before channel exchange, reducing the negative influence of large modality discrepancies on subsequent fusion. The enhanced features are formulated as follows:

$$f_r' = f_r \otimes f_e + f_r; \quad f_e' = f_r \otimes f_e + f_e. \tag{3}$$

In our module, we introduce a channel exchange mechanism, motivated by the observation that the information strength carried along the channel dimension is often imbalanced across different modalities (RGB and Event). Specifically, certain channels in one modality may contain highly discriminative cues, while the corresponding channels in the other modality may primarily encode noise or irrelevant information. Conventional fusion strategies, such as direct summation or concatenation, tend to preserve these weak channels, thereby diluting the contribution of informative signals. The proposed channel switching module is designed to selectively replace or reallocate weak channels by importing highly informative channels from the complementary modality, thus enhancing the overall discriminative representation. To achieve this, we first perform a channel-wise importance evaluation on the feature maps of both modalities. This evaluation is implemented through global average pooling (GAP), followed by a one-dimensional convolution and a sigmoid activation function. Based on the resulting channel weights, discriminative channel substitution is carried out to accomplish effective channel exchange. Formally, the computed weights after weight evaluation ω_r^{cs} and ω_e^{cs} are as follows:

$$\begin{aligned} \omega_{\mathrm{r}} &= \sigma\left(\mathrm{Conv1D}\left(\mathrm{GAP}\left(\mathrm{f}_{\mathrm{r}}'\right)\right)\right); \\ \omega_{\mathrm{e}} &= \sigma\left(\mathrm{Conv1D}\left(\mathrm{GAP}\left(\mathrm{f}_{\mathrm{e}}'\right)\right)\right). \end{aligned} \tag{4}$$

where

$$\mathrm{GAP(f)} = \frac{1}{\mathrm{HW}} \sum_{\mathrm{i}=0}^{\mathrm{H}-1} \sum_{\mathrm{j}=0}^{\mathrm{W}-1} \mathrm{f_{ij}} \tag{5}$$

where Conv1D denotes a one-dimensional convolution along the channel dimension, and σ denotes the sigmoid activation function. The channel weights ω_r^{cs} and ω_e^{cs} first use GAP (Global Average Pooling) to reduce the input RGB and Event feature maps to $1 \times 1 \times C$ vectors, then after alignment through $1D$ convolution, the weights are passed through a sigmoid function. The resulting channel weights ω_r^{cs} and ω_e^{cs} are used for channel exchange. The channel exchange process is defined as follows:

$$\begin{aligned} f_{r,c}^{out} &= \begin{cases} f_{r,c}', & \text{if } \omega_{r,c}^{cs} \geq k_r, \\ f_{e,c}', & \text{if } \omega_{r,c}^{cs} < k_r, \end{cases} \\ f_{e,c}^{out} &= \begin{cases} f_{e,c}', & \text{if } \omega_{e,c}^{cs} \geq k_e, \\ f_{r,c}', & \text{if } \omega_{e,c}^{cs} < k_e. \end{cases} \end{aligned} \tag{6}$$

where c represents the $c-th$ channel, and k is a predetermined threshold. When the weight of the $c-th$ channel is lower than the threshold k, the model replaces it with the corresponding channel from the other modality.

After completing the channel-level information interaction, to further achieve fine-grained fusion and semantic alignment between RGB and event features, this study introduces a spatial attention mechanism to enhance the network's responsiveness to key spatial regions. This module performs channel concatenation and spatial weighting on the fused feature maps, enabling the model to identify critical regions shared across both modalities (such as motion edges and object boundaries).

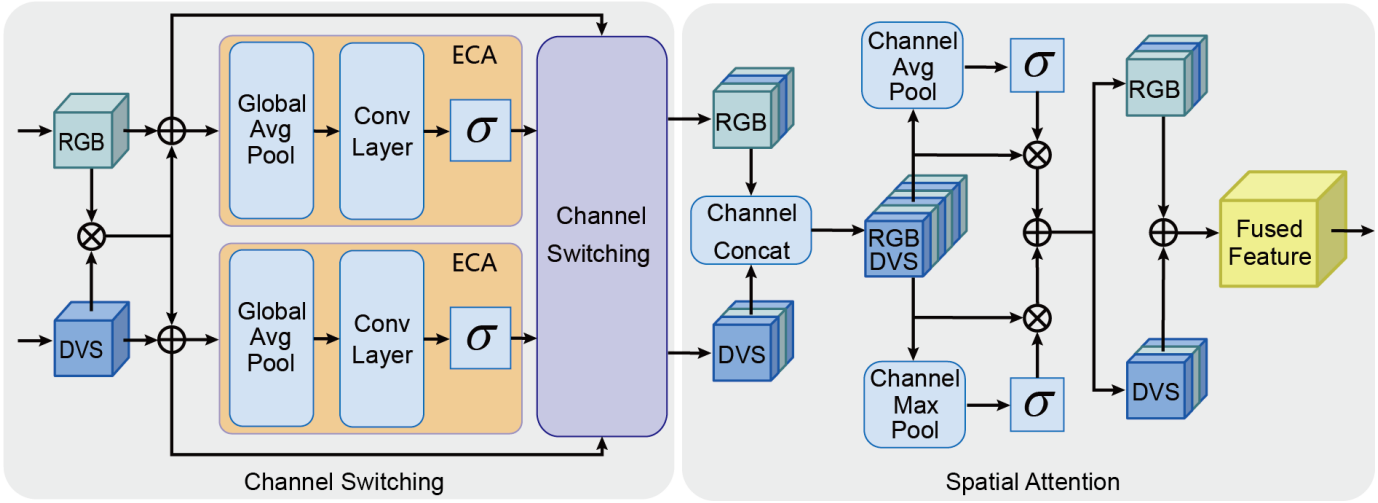

Fig. 2. Structure of the proposed Bidirectional Channel Interaction Module (BCIM). It consists of channel switching and spatial attention refinement to perform channel-wise cross-modal exchange and spatial feature enhancement.

Consequently, the final fusion output preserves both static semantics and dynamic information. This mechanism effectively alleviates the semantic misalignment between RGB and event data, providing the subsequent detection branch with more consistent and higher-quality fusion feature inputs.

As shown in Figure 2, after channel exchange, the RGB and event features are concatenated along the channel dimension to form a joint multimodal representation. To generate spatial attention maps, the concatenated features are processed by channel-wise average pooling and channel-wise max pooling, producing two spatial descriptors. The channel-wise max pooling operation emphasizes the most discriminative response regions, such as moving-object boundaries and high-confidence motion cues, while the channel-wise average pooling operation preserves the overall spatial structure and contextual information. The two descriptors are then passed through convolutional layers and normalized by a sigmoid function to obtain pixel-level spatial attention maps. These attention maps are used to reweight the concatenated features, enabling the network to suppress redundant background responses and enhance informative regions. The computation process can be expressed as:

$$\begin{aligned} M_a &= \sigma(\text{Conv}(\text{CAP}(f_{cat}))); \\ M_m &= \sigma(\text{Conv}(\text{CMP}(f_{cat}))). \end{aligned} \tag{7}$$

$$f'_{\text{cat}} = \frac{M_a \otimes f_{\text{cat}} + M_m \otimes f_{\text{cat}}}{2}, \tag{8}$$

where f_{cat} denotes the concatenated RGB and event features. Conv represents the convolution operation, and σ denotes the sigmoid activation function. CMP and CAP denote channel-wise max pooling and channel-wise average pooling along the channel dimension, respectively. The spatial attention maps M_a and M_m, generated through CAP and CMP, are utilized to modulate the fused features, enabling the network to focus on the most informative spatial regions. f'_{cat} represents the final concatenated feature after attention modulation, and $\otimes$ denotes element-wise multiplication.

Finally, the refined feature map is split back into RGB and event feature components, which are then added element-wise to obtain the final fused representation, as formulated below:

$$f_r^{SA}, f_e^{SA} = \text{Split}(f'_{\text{cat}}), \tag{9}$$

$$f = \frac{f_r^{SA} + f_e^{SA}}{2}, \tag{10}$$

where f_r^{SA} and f_e^{SA} represent the RGB and event features obtained after spatial attention output, respectively. The weighted feature maps $f_r^{SA}, f_e^{SA} \in \mathbb{R}^{H\times W\times C}$ have the same dimensions as the input feature maps.

III. Experiments

A. Experimental Setup

1) Datasets: Experiments are conducted on DSEC-MOD and PKU-DAVIS-SOD. DSEC-MOD is built from the DSEC benchmark and contains frame-level annotations for eight moving object categories under morning, afternoon, and night conditions. It is used as the primary dataset for quantitative comparison, ablation study, and illumination-related analysis.

PKU-DAVIS-SOD contains three common moving object categories and is used as an auxiliary dataset to further evaluate the robustness of the proposed method under illumination variations and motion degradation conditions.

2) Data Preprocessing: Due to the asynchronous nature of event data, the event stream is first converted into frame-like event representations. The RGB frames are processed by convolutional layers, while the event stream is represented at multiple temporal scales. Specifically, the E-TMA module [15] is used to extract and aggregate informative event features across different temporal windows. For DSEC-MOD, the temporal windows are set to 15 ms, 30 ms, and 50 ms, while for PKU-DAVIS-SOD, they are set to 15 ms, 25 ms, and 40 ms.

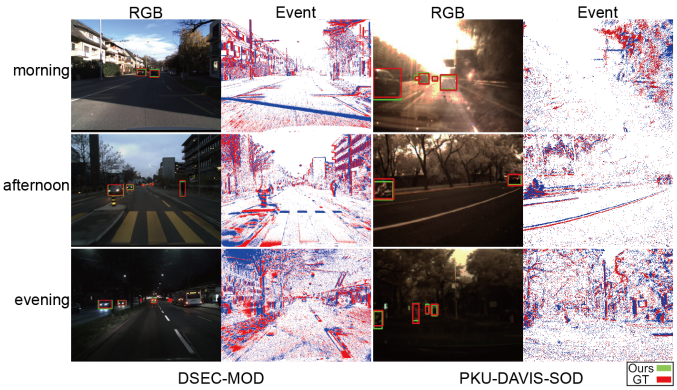

Fig. 3. Visualization of detection results across different illumination conditions (morning, afternoon, evening) on the DSEC-MOD and PKU-DAVIS-SOD datasets. The number of video sequences in DSEC-MOD for morning, afternoon, and evening is 2:2:1, while in PKU-DAVIS-SOD it is 6:7:15.

3) Training Details: CSPDarkNet [14] is adopted as the backbone network, and all input frames are resized to 288×288. During training, common data augmentation strategies are applied, including photometric transformation, random scaling, and positional jittering. The model is optimized using Adam with an initial learning rate of 5×10^{-4}, which is reduced by a factor of 10 at the 10th and 15th epochs. The entire model is trained for 30 epochs.

B. Experimental Results

1) Comparative Experiments on the DSEC-MOD Dataset: Table I reports the comparison results on the DSEC-MOD dataset. Self-attention based methods improve intra-modal feature representation but lack effective cross-modal interaction, resulting in limited gains in RGB-Event fusion tasks. Cross-attention based methods introduce inter-modal information propagation; however, asymmetric fusion strategies may cause modality bias under challenging conditions.

Compared with existing methods, the proposed CCFNet introduces bidirectional channel exchange and spatial attention refinement for adaptive multimodal fusion. As a result, CCFNet achieves the highest F.mAP of 44.14 on DSEC-MOD, slightly outperforming FRN and clearly surpassing other representative fusion methods. The qualitative results in Fig. 3 further demonstrate its effectiveness under complex illumination and motion conditions.

2) Sensitivity Analysis of the Threshold k on the DSEC-MOD Dataset: Table II shows that the threshold has a noticeable influence on performance rather than being negligible. When the threshold is too low, weak channels may be retained, limiting the benefit of complementary exchange. When the threshold is inappropriate for one modality, the exchange process may introduce noisy or mismatched channels. The best result is obtained when both thresholds are set to 0.5, suggesting that a balanced criterion between feature preservation and cross-modal replacement is beneficial.

3) Performance Analysis across Different Time Periods: Table III presents the performance of CCFNet under different illumination conditions on the DSEC-MOD and PKU-DAVIS-SOD datasets. The proposed method achieves stable performance during daytime conditions and maintains competitive detection capability under nighttime scenarios.

The relatively lower nighttime performance on DSEC-MOD is mainly due to the limited number of nighttime testing samples. In contrast, the more balanced distribution in PKU-DAVIS-SOD leads to more consistent results across different time periods. The qualitative results in Fig. 3 further verify the robustness of the proposed fusion framework.

TABLE I
COMPARISON WITH SOTA FUSION ALTERNATIVES ON THE DSEC DATASET.

Model Type	Fusion Module	F.mAP@0.5 (%)
Self-Attention	SENet [17]	29.28
	CBAM [18]	36.22
	ECANet [19]	34.49
Cross-Attention	SAGate [20]	33.62
	DCF [21]	32.20
	SPNet [22]	32.70
RGB-Event	FPN-Fusion [23]	32.28
	EFNet [24]	35.33
	RENet [15]	38.38
	FRN [25]	43.59
	CCFNet (Ours)	**44.14**

TABLE II
SENSITIVITY ANALYSIS OF THE THRESHOLD PARAMETERS ON THE DSEC-MOD DATASET. THE PERFORMANCE IS REPORTED BY F.MAP@0.5 (%).

Varied Parameter	Fixed Parameter	Threshold Value				
		0.1	**0.2**	**0.3**	**0.4**	**0.5**
k_r	$k_e = 0.5$	40.44	41.74	42.08	38.86	**44.14**
k_e	$k_r = 0.5$	40.73	39.42	40.69	42.00	**44.14**

TABLE III
PERFORMANCE COMPARISON OF CCFNET ACROSS DIFFERENT TIME PERIODS ON DSEC-MOD AND PKU-DAVIS-SOD.

Dataset	F.mAP (%)			
	Morning	**Afternoon**	**Evening**	**All**
DSEC-MOD	46.17	**47.98**	29.36	44.14
PKU-DAVIS-SOD	63.34	57.10	**79.82**	71.02

4) Ablation Study: To evaluate the effectiveness of each component, we conduct ablation experiments on the DSEC-MOD dataset. The results are summarized in Table IV.

The baseline model using direct RGB-Event concatenation achieves only 15.37 F.mAP, indicating that naive fusion cannot effectively exploit multimodal complementary information.

After introducing the channel exchange module, the performance significantly improves to 41.20 F.mAP, demonstrating the effectiveness of adaptive cross-modal feature interaction. Using only the spatial attention module also improves performance to 42.91 F.mAP by enhancing salient spatial responses.

By jointly integrating channel exchange and spatial attention, CCFNet achieves the best performance of 44.14 F.mAP, showing that the two modules provide complementary benefits for multimodal fusion.

TABLE IV
ABLATION STUDY OF KEY COMPONENTS ON THE DSEC-MOD DATASET.

No.	R-E Baseline	Channel Switching	Spatial Attention	F.mAP@0.5 (%)
1	✓	–	–	15.37
2	✓	✓	–	41.20
3	✓	–	✓	42.91
4	✓	✓	✓	**44.14**

IV. CONCLUSION

In this paper, we propose CCFNet, an RGB-Event fusion framework for moving object detection in challenging driving environments. By combining bidirectional channel exchange and spatial attention refinement, CCFNet adaptively exploits complementary information from RGB images and event streams. Experiments on DSEC-MOD demonstrate competitive performance compared with representative fusion methods, while additional analysis on PKU-DAVIS-SOD evaluates its robustness under different illumination conditions. Ablation studies further verify the effectiveness of the proposed components. In future work, we will explore more efficient temporal modeling strategies and lightweight designs for real-world deployment.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (2025ZD0217200), CAS Project for Young Scientists in Basic Research (YSBR-116), Youth Innovation Promotion Association CAS, the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB1010302), the Lingang Laboratory Fund (Grant No. LG-GG-202402-06-07, LGL-1987-09), the Shanghai Municipal Science and Technology Project (Grant No. 25ZR1401370, 25LN3200400), Special Support Project of Guangdong Province (Grant No.0720240209). The numerical calculations in this study were carried out on the ORISE Supercomputer.

REFERENCES

[1] H. Ghahremannezhad, H. Shi, and C. Liu, "Object detection in traffic videos: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6780–6799, 2023.

[2] Q. Guo, W. Feng, R. Gao, Y. Liu, and S. Wang, "Exploring the effects of blur and deblurring to visual object tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 1812–1824, 2021.

[3] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021.

[4] X. Guo, T. Chen, J. Liu, Y. Liu, and Q. An, "Dim space target detection via convolutional neural network in single optical image," *IEEE Access*, vol. 10, pp. 52 306–52 318, 2022.

[5] D. M. Jiménez-Bravo, Á. L. Murciego, A. S. Mendes, H. S. San Blás, and J. Bajo, "Multi-object tracking in traffic environments: A systematic literature review," *Neurocomputing*, vol. 494, pp. 43–55, 2022.

[6] M. Ben Miled, W. Liu, and Y. Liu, "Bioinspired framework for real-time collision detection with dynamic obstacles in cluttered outdoor environments using event cameras," *IET Cyber-Systems and Robotics*, vol. 7, no. 1, p. e70006, 2025.

[7] W. Shariff, M. S. Dilmaghani, P. Kielty, M. Moustafa, J. Lemley, and P. Corcoran, "Event cameras in automotive sensing: A review," *IEEE Access*, vol. 12, pp. 51 275–51 306, 2024.

[8] M. Aitsam, S. Davies, and A. Di Nuovo, "Event camera-based real-time gesture recognition for improved robotic guidance," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[9] D. Gehrig and D. Scaramuzza, "Low-latency automotive vision with event cameras," *Nature*, vol. 629, no. 8014, pp. 1034–1040, 2024.

[10] Z. Liu, N. Yang, Y. Wang, Y. Li, X. Zhao, and F.-Y. Wang, "Enhancing traffic object detection in variable illumination with rgb-event fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 20 335–20 350, 2024.

[11] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, "Multi-modal 3d object detection in autonomous driving: a survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 2122–2152, 2023.

[12] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications." *Computers, Materials & Continua*, vol. 80, no. 1, 2024.

[13] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[14] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[15] Z. Zhou, Z. Wu, R. Boutteau, F. Yang, C. Demonceaux, and D. Ginhac, "Rgb-event fusion for moving object detection in autonomous driving," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7808–7815.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[18] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[19] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[20] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European conference on computer vision*. Springer, 2020, pp. 561–577.

[21] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9471–9481.

[22] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4681–4691.

[23] A. Tomy, A. Paigwar, K. S. Mann, A. Renzaglia, and C. Laugier, "Fusing event-based and rgb camera for robust object detection in adverse conditions," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 933–939.

[24] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *European conference on computer vision*. Springer, 2022, pp. 412–428.

[25] H. Cao, Z. Zhang, Y. Xia, X. Li, J. Xia, G. Chen, and A. Knoll, "Embracing events and frames with hierarchical feature refinement network for object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 161–177.