Robotics, Artificial Intelligence and Embedded Systems

Embracing Events and Frames with Hierarchical Feature Refinement Network for Object Detection

Authors: **Hu Cao**, Zehua Zhang, Yan Xia, Xinyi Li, Jiahao Xia, Guang Chen, Alois Knoll

Technical University of Munich

UNIVERSITY OF TECHNOLOGY SYDNEY

TONGJI UNIVERSITY 1907

Munich Center for Machine Learning

Challenges

- Challenges in frame-based cameras

➢ The performance of conventional frame-based cameras in object perception often faces a significant decline in challenging conditions, such as overexposure, low light, and motion blur (e.g., high-speed motion).

➢ Overexposure

➢ Low light

➢ Motion blur

Figure 1: The problems in frame-based cameras.

Challenges

- Event camera

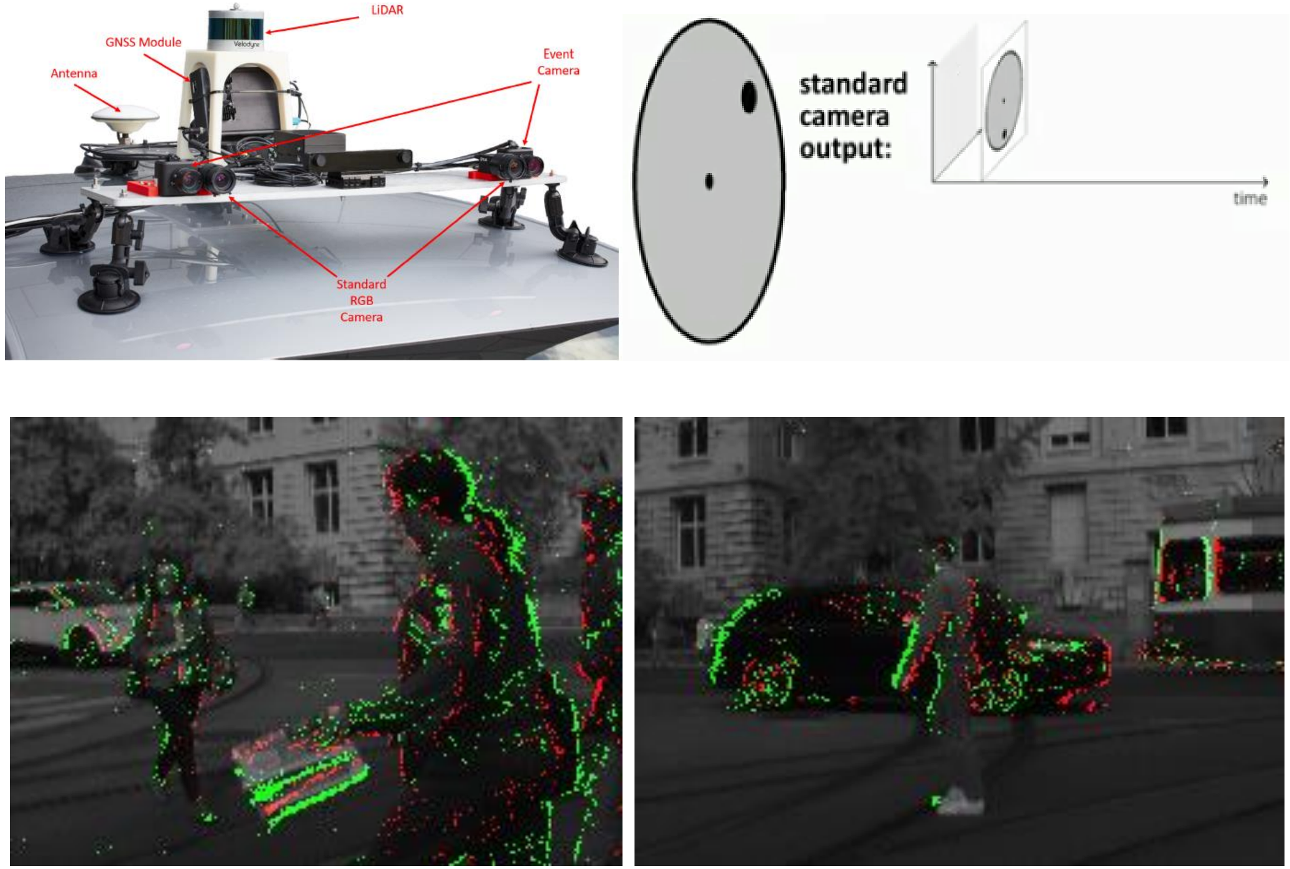


Figure 2: Comparison between standard camera and event camera.

➢ The event camera is a bio-inspired vision sensor that captures dynamic changes in the scene and filters out redundant information.

➢ Strengths:

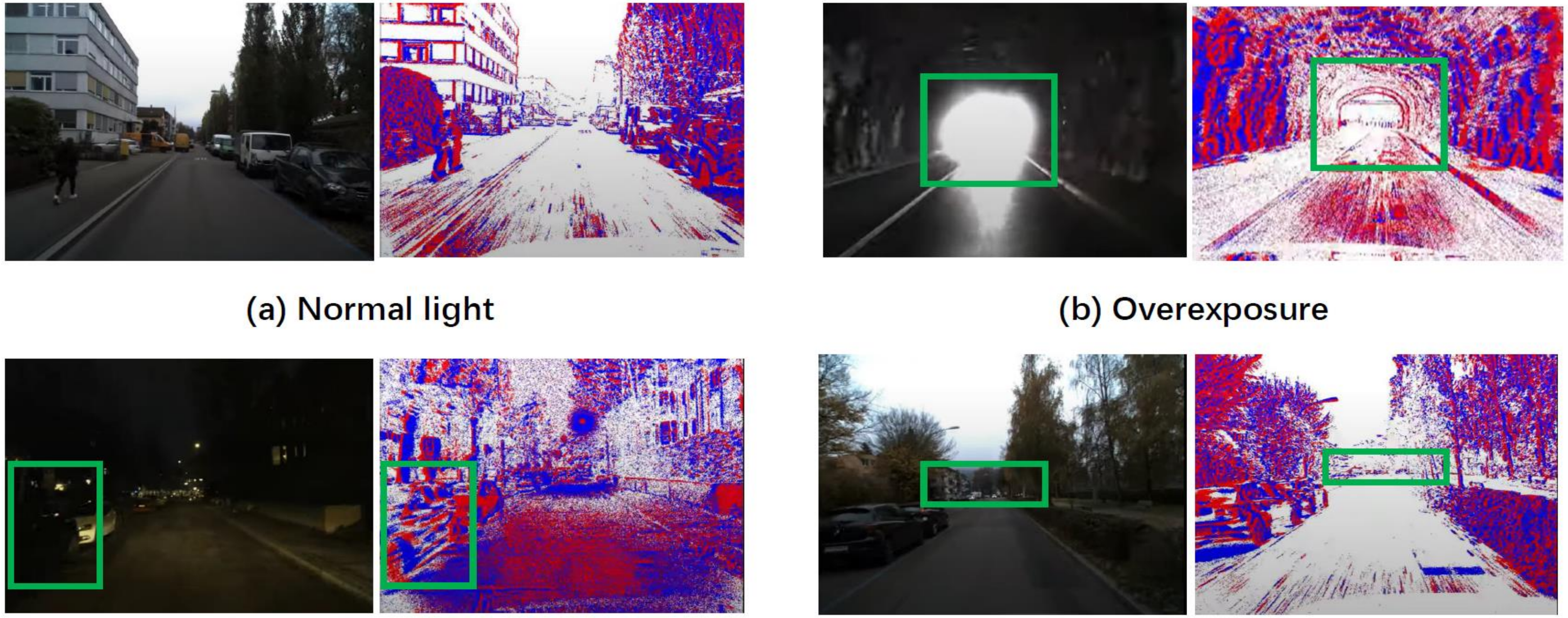
(1) No redundant background information;
(2) Low latency;
(3) High temporal resolution;
(4) High dynamic range.

➢ Weakness:

(1) No color information;
(2) No texture information.

Challenges

- Challenges in frame-based cameras and event cameras



(a) Normal light (b) Overexposure

(c) Low light (d) Remote targets

Figure 3: Challenging scenarios.

➢ Both event cameras and frame-based cameras are complementary, motivating the development of new algorithms for object perception.

Feature Imbalance Problems

- How to fuse these two heterogeneous modalities?

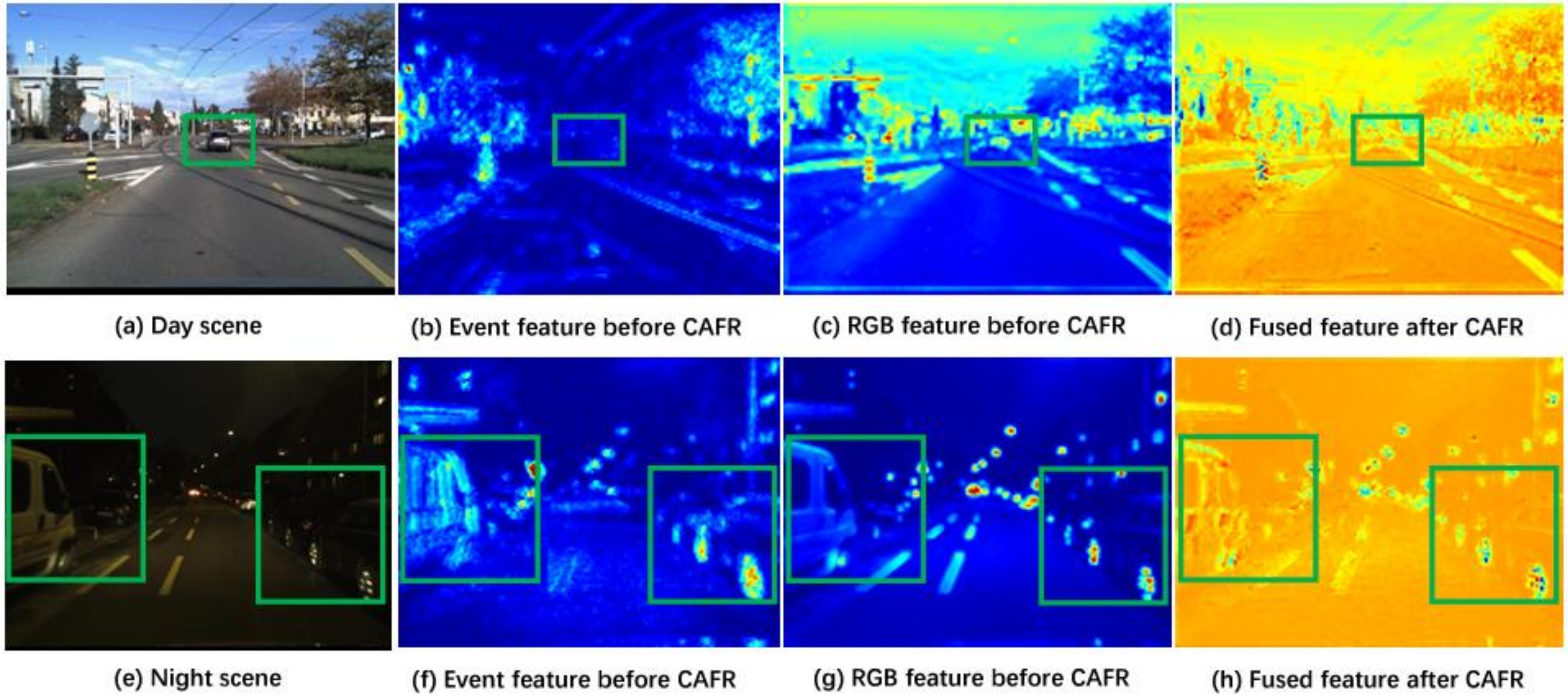


Figure 4: Feature maps of RGB and event modalities before and after CAFR.

➢ We propose a novel hierarchical feature refinement network with CAFR modules for event-frame fusion.

Hierarchical Feature Refinement Network

- Method

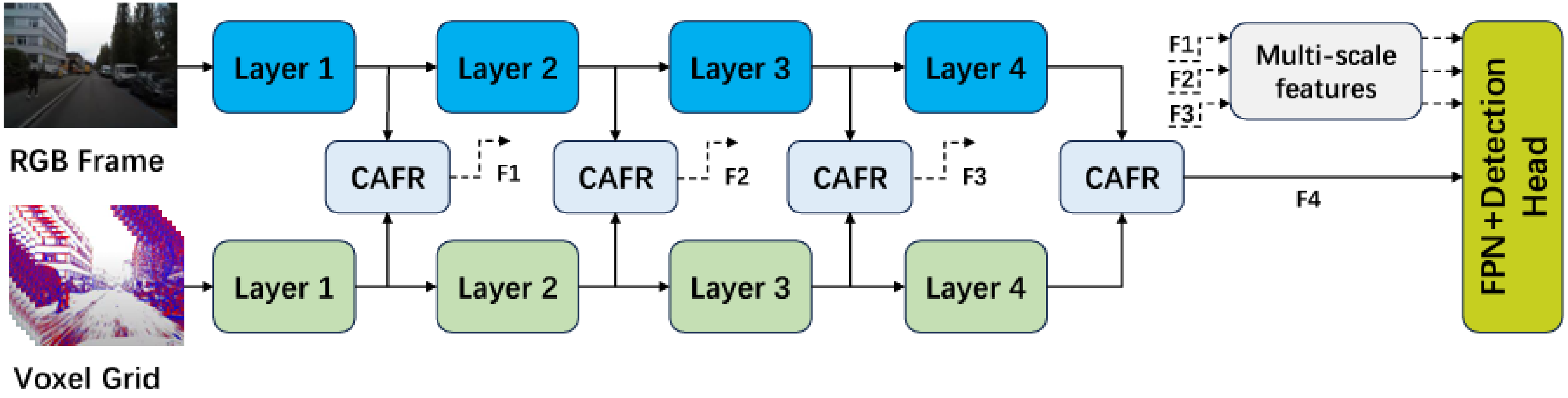


Figure 5: The overall architecture of our hierarchical feature refinement network.

➢ In contrast to the current event-frame fusion methods, our method adopts a dual-branched coarse-to-fine structure. The dual-branch architecture guarantees comprehensive utilization of both event-based and frame-based features.

Hierarchical Feature Refinement Network

- CAFR

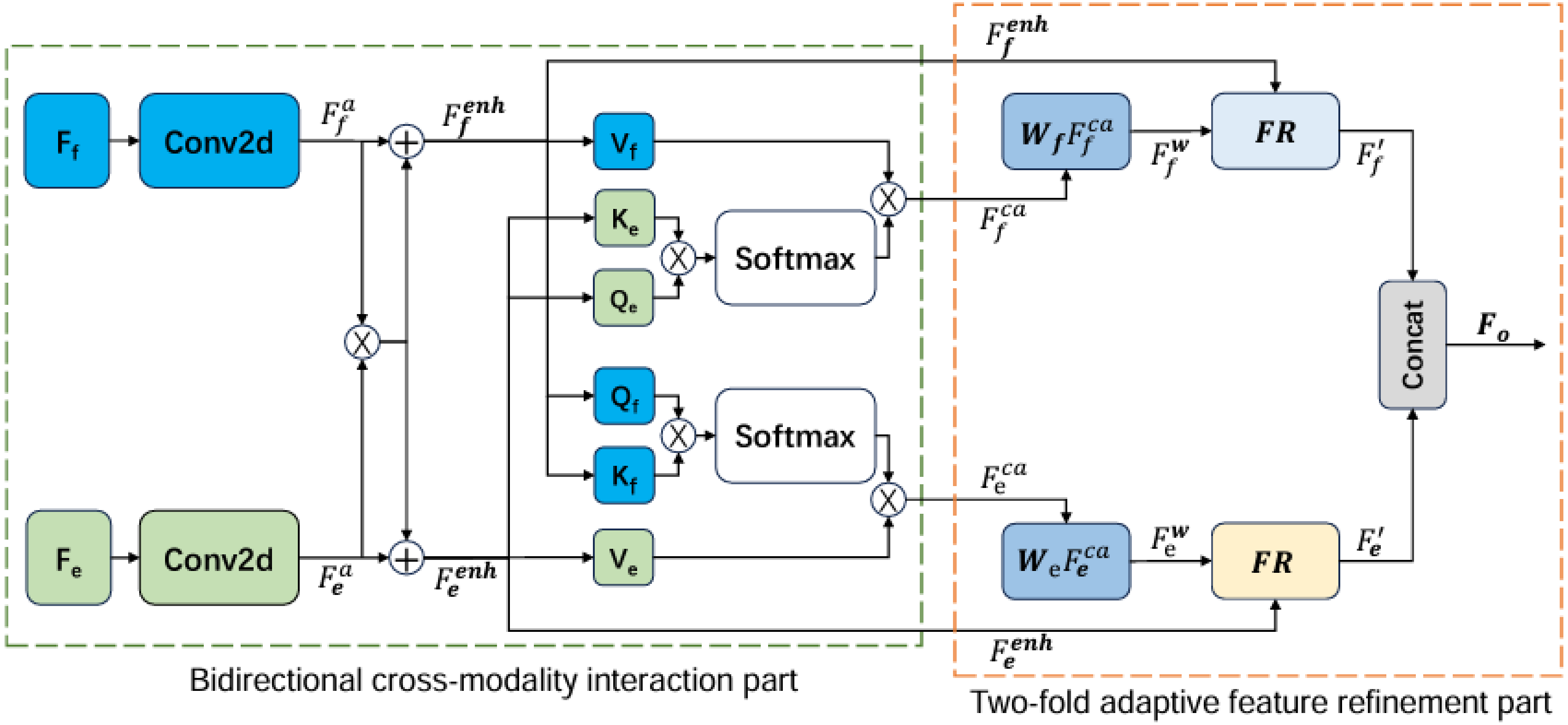


Figure 6: Cross-modality adaptive feature refinement module (CAFR).

➢ For effective information exchange between different modal features, CAFR receive event-based and frame-based features to balance the information flow.

Experimental results

- Comparison with SOTA methods on the DSEC dataset

Table 3: Comparison with SOTA methods on the DSEC dataset.

Modality	Method	Model type	mAP (%) Car	Pedestrian	Large vehicle	Average
Events + Frames	SENet [20]	Attention	38.4	14.9	26.0	26.2
	ECANet [49]	Attention	36.7	12.8	27.5	25.7
	CBAM [50]	Attention	37.7	13.5	27.0	26.1
	SAGate [11]	RGB-D	32.5	10.4	16.0	19.6
	DCF [23]	RGB-D	36.3	12.7	28.0	25.7
	SPNet [57]	RGB-D	39.2	17.8	26.2	27.7
	FPN-Fusion [48]	RGB-E	37.5	10.9	24.9	24.4
	DRFuser [38]	RGB-E	38.6	15.1	30.6	28.1
	RAMNet [15]	RGB-E	24.4	10.8	17.6	17.6
	CMX [55]	RGB-E	41.6	16.4	29.4	29.1
	FAGC [5]	RGB-E	39.8	14.4	33.6	29.3
	RENet [58]	RGB-E	40.5	17.2	30.6	29.4
	EFNet [46]	RGB-E	41.1	15.8	32.6	30.0
	CAFR (Ours)		**49.9**	**25.8**	**38.2**	**38.0**

➢ Compared with other methods, our CAFR achieves significant improvements. Notably, CAFR outperforms the second-best method, EFNet [46], by an impressive margin of **8.0%**.

Experimental results

- Comparison with SOTA methods on the PKU-DDD17-Car dataset

Table 4: Comparison with SOTA methods on the PKU-DDD17-Car dataset.

Modality	Method	Input representation	Model type	mAP_{50} (%)	mAP (%)
Events	MTC [8]	Channel image	Events only	47.8	-
	ASTMNet [28]	Event embedding		46.2	-
Frames	SSD [34]	Frame	Frames only	73.1	-
	Faster-RCNN [43]			80.2	-
	YOLOv4 [2]			81.3	-
Events + Frames	SENet [20]	Voxel grid + Frame	Attention	81.6	42.4
	ECANet [49]			82.2	40.8
	CBAM [50]			81.9	42.8
	SAGate [11]	Voxel grid + Frame	RGB-D	82.0	43.4
	DCF [23]			83.4	42.5
	SPNet [57]			84.7	43.3
	JDF [27]	Channel image + Frame	RGB-E	84.1	-
	FPN-Fusion [48]	Voxel grid + Frame		81.9	41.6
	DRFuser [38]	Voxel grid + Frame		82.6	42.4
	RAMNet [15]	Voxel grid + Frame		79.6	38.8
	CMX [55]	Voxel grid + Frame		80.4	39.0
	FAGC [5]	Voxel grid + Frame		84.8	42.4
	RENet [58]	Voxel grid + Frame		81.4	43.9
	EFNet [46]	Voxel grid + Frame		83.0	41.6
	CAFR (Ours)	Voxel grid + Frame		**86.7**	**46.0**

➢ Our CAFR achieves the best performance in terms of mAP50 and mAP with accuracy of **86.7%** and **46.0%,** respectively.

Experimental results

- Robustness

Table 5: The performance of different methods under various corruption conditions, including noise, blur, weather, and digital.

Method	Model type	$mPC_{50}(\%)$				
		Average	Noise	Blur	Weather	Digital
Frames only [30]	Frames only	38.7	47.6	25.3	28.5	53.0
SENet [20]	Attention	63.6	68.6	56.6	58.9	70.3
ECANet [49]	Attention	67.1	72.6	57.6	66.8	71.4
CBAM [50]	Attention	65.2	69.9	57.2	62.4	70.3
SAGate [11]	RGB-D	63.6	68.1	55.9	61.1	69.4
DCF [23]	RGB-D	65.7	70.9	57.9	62.9	71.1
SPNet [57]	RGB-D	66.6	70.6	58.7	64.8	72.3
FPN-Fusion [48]	RGB-E	64.7	70.0	56.6	63.9	69.4
DRFuser [38]	RGB-E	67.7	72.1	59.4	67.8	71.4
RAMNet [15]	RGB-E	53.9	53.5	43.3	54.3	64.6
CMX [55]	RGB-E	64.2	67.7	56.0	62.9	70.2
FAGC [5]	RGB-E	52.4	62.8	38.5	48.4	59.9
RENet [58]	RGB-E	57.2	58.5	**72.3**	29.9	68.1
EFNet [46]	RGB-E	66.4	67.1	58.2	66.7	73.4
CAFR (Ours)	RGB-E	**69.5**	**73.6**	57.0	**70.6**	**76.7**

➢ In comparison to other fusion methods, our proposed CAFR demonstrates superior performance. These findings highlight the effectiveness of CAFR in strengthening the model against corrupted data across diverse severity levels and types.

Visualization

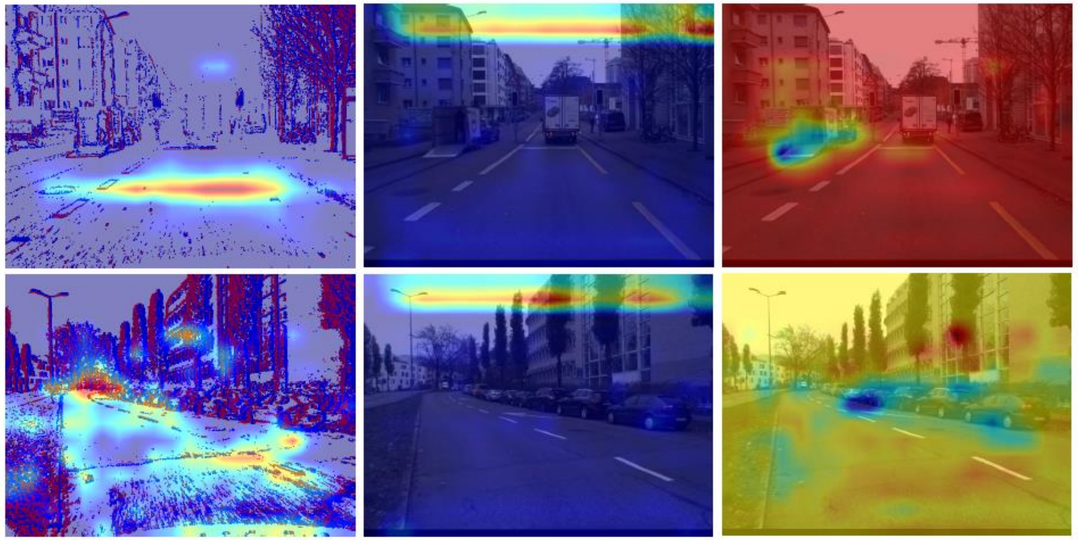


Figure 7: Representative examples of different activation maps.

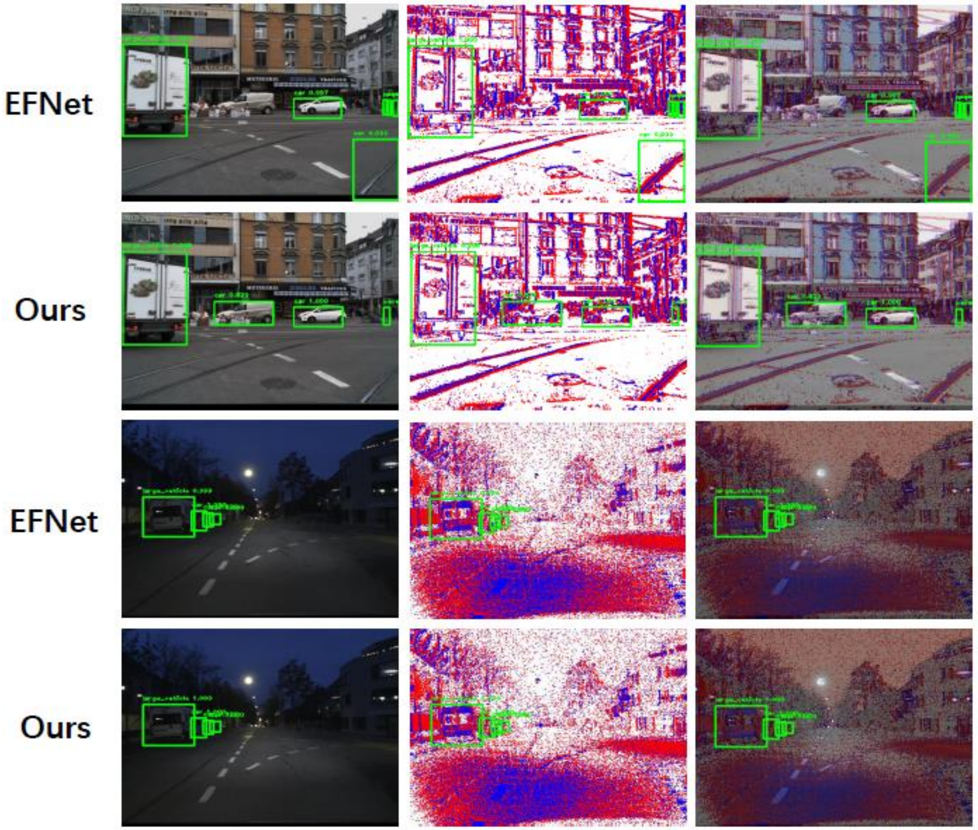


Figure 8: Representative examples of different object detection results on the DSEC dataset.

- After applying CAFR, the model demonstrates enhanced focus on significant regions.

- The detection results demonstrate that the proposed method can consistently produce satisfactory detection results in various challenging scenarios.

Thanks for Your Attention!