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

Hu Cao[1], Zehua Zhang[1], Yan Xia[1,4], Xinyi Li[1], Jiahao Xia[2], Guang Chen[3⋆], and Alois Knoll[1]

[1] Technical University of Munich, Munich, Germany
{hu.cao,zehua.zhang,yan.xia,super.xinyi,k}@tum.de
[2] University of Technology Sydney, Sydney, Australia
Jiahao.Xia@student.uts.edu.au
[3] Tongji University, Shanghai, China
guangchen@tongji.edu.cn
[4] Munich Center for Machine Learning (MCML)

1 More Experimental Details

1.1 Datasets

DDD17. The vehicles in the DDD17 dataset were manually labeled by the authors of [11] to create the PKU-DDD17-Car dataset for object detection tasks. Additional details about the PKU-DDD17-Car dataset are outlined in Tab. 1.
DSEC. The original dataset comprises 53 sequences captured in three distinct regions of Switzerland. RGB frames are captured using the FLIR color camera, which boasts a resolution of 1440 × 1080. The original DSEC dataset [5] lacks the required labels for object detection. The labels introduced in [18] is used in this work. This annotated dataset consists of a total of 41 sequences, allocated for training (33 sequences), validation (3 sequences), and testing (5 sequences). Notably, the dataset covers a wide range of lighting conditions, from ideal to highly challenging. This diversity guarantees comprehensive testing of vision systems, ensuring their robustness and applicability in real-world settings. Additionally, unlike the DDD17 dataset, in which both event data and RGB data are sourced from the same DAVIS camera, the raw event data and RGB data in the DSEC dataset are not aligned; they do not correspond to the same frame. Therefore, additional preprocessing of the raw DSEC data is essential. Further details regarding the preprocessing steps are provided below:

Homographic transformation. To address the misalignment between the event data and RGB data, the two types of data need to be transformed into the same frame. The dataset provides a baseline of 4.5 cm between the two cameras, allowing for the transformation to a common viewpoint. In our pre-processing, we leverage the homographic transformation induced by pure rotation, as derived in Eq. 1, to align the scene from an RGB frame to an event-camera frame. It's

⋆ Corresponding author

Table 1: A detailed description of the recorded data in the PKU-DDD17-CAR dataset

Recorded data (.hdf5)	Condition	Length (s)	Type
1487339175	day	347	test
1487417411	day	2096	test
1487419513	day	1976	train
1487424147	day	3040	train
1487430438	day	3135	train
1487433587	night-fall	2335	train
1487593224	day	524	test
1487594667	day	2985	train
1487597945	night-fall	50	test
1487598202	day	1882	train
1487600962	day	2143	test
1487608147	night-fall	1208	train
1487609463	night-fall	101	test
1487781509	night-fall	127	test

Fig. 1: Examples illustrate the comparison before and after the homomorphic transformation. On the left (a): an overlay map of the original RGB image and event image. On the right (b): an overlay map of the RGB image and event image after the homomorphic transformation.

important to note that, in our scenario, the scene appears far away from the camera, and the baseline of 4.5 cm is smaller than the distances of the scene objects.

$$P_{\text{event,rgb}} = K_{\text{event}} * R_{\text{rgb}} * R_{\text{event,rgb}} * R_{\text{event}}^{T} * K_{\text{rgb}}^{-1}. \tag{1}$$

where K_{rgb} and K_{event} represent the intrinsic camera matrices of the RGB and event cameras, respectively. Similarly, R_{rgb} and R_{event} denote the rotation matrices accounting for the transition from distorted to undistorted frames for each camera. Additionally, $R_{\text{event,rgb}}$ stands for the rotation matrix aligning the RGB camera coordinate system with that of the event camera.

Table 2: Object annotations in the labeled DSEC dataset.

Categories	Car	Pedestrian	Large vehicle (Bus & Truck)
Count	100068	17126	14771
Percentage	0.76	0.13	0.11

Table 3: Data amount in the labeled DSEC dataset.

Type	Train	Val	Test	Total
Sequences	33	3	5	41
Frames	44148	3642	4896	52686

Table 4: The details of different corruption types.

Group	Corruption Type
Noise	Gaussian Noise, Shot Noise, Impulse Noise
Blur	Defocus Blur, Glass Blur, Motion Blur, Zoom Blur
Weather	Fog, Snow, Frost, Brightness
Digital	Contrast, Elastic Transform, Pixelate, Jpeg Compression

Fig. 1 illustrates the impact of the homomorphic transformation. In the left figure, the RGB image and the event image are not aligned. However, after applying the homomorphic transformation, the RGB image and the event image become aligned, sharing the same frame and being suitable for multimodal fusion. Additionally, the size of the RGB images is resized to match the event image, ensuring that both types of data have the same field of view and resolution.

Annotation generation. To facilitate object detection on the DSEC dataset, we employed simulated annotations provided by [18]. YOLOv5 [9] was used to label RGB images. The homographic transformation was applied to transfer labels from RGB images to the event frame, accounting for objects within the event camera's 640×480 resolution. Examples of labeled events and RGB images are depicted in Fig. 2, encompassing three object classes: car, pedestrian, and large vehicle (refer to Tab. 2).

After the aforementioned pre-processing steps, we acquired the labeled DSEC dataset for experimentation. The dataset comprises a total of 41 sequences, with 33 sequences allocated for training, 3 sequences for validation, and the remaining 5 sequences designated for testing. Detailed information about the data is presented in Tab. 3, where the event data file is approximately 300 GB and the RGB data file is around 23 GB.

Corruption data. In this work, we introduced 15 types of corruption, each with five levels of severity, to assess the impact of diverse corruption types on object detection models. The chosen corruption types are categorized into four groups:

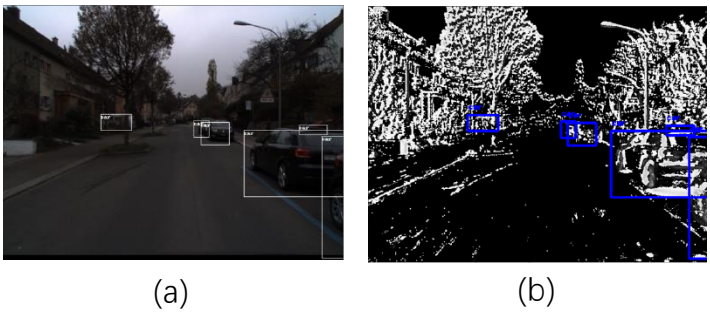

Fig. 2: Examples of events and RGB images with annotations. Left (a): RGB image; Right (b): event frame.

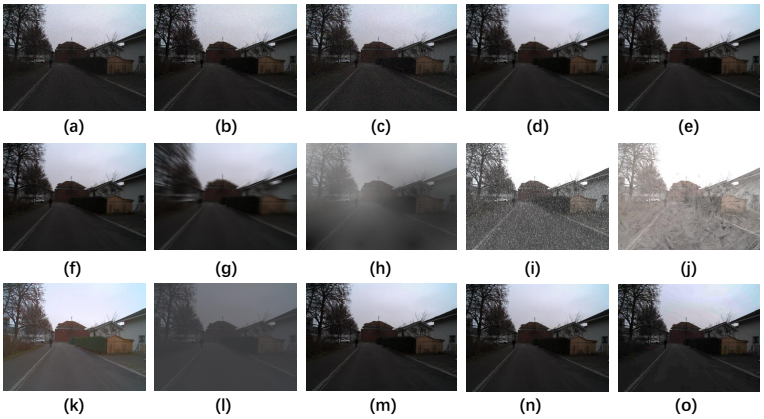

Fig. 3: The dataset encompasses 15 types of algorithmically generated corruptions, categorized into noise, blur, weather, and digital groups. The corrupted images, denoted from (a) to (o), include Gaussian Noise, Shot Noise, Impulse Noise, Defocus Blur, Glass Blur, Motion Blur, Zoom Blur, Fog, Snow, Frost, Brightness, Contrast, Elastic Transform, Pixelate, and Jpeg Compression. Each corruption type exhibits five severity levels, resulting in a total of 75 distinct corruptions. The images presented here correspond to severity level 2.

noise, blur, weather, and digital. The specific corruption types are detailed in Tab. 4. Fig. 3 provides an illustration of these corruption types at severity level 2. All corruption treatments are applied to the test set, enabling us to evaluate a model's robustness against previously unseen corruptions.

Noise. The first corruption type is Gaussian noise. This corruption may occur in low-light conditions. Electronic noise caused by the discontinuous character

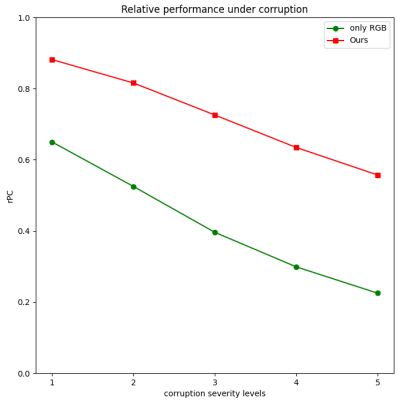

Fig. 4: Relative performance under various severity.

of light is known as shot noise, also referred to as poisson noise. Impulse noise is the color analog of salt-and-pepper noise and can be caused by bit errors.

Blur. Defocus blur occurs when the image is out of focus. Glass blur appears on "frosted glass" windows or panels. When the camera is moving swiftly, motion blur happens. When the camera advances quickly toward an item, zoom blur occurs.

Weather. Snow is a type of precipitation that impairs vision. When there are ice crystals on the lenses or windows, frost occurs. A diamond-square method is used to render the fog that surrounds the items. The brightness varies with the intensity of daylight.

Digital. Based on the lighting and the subject's color, contrast can be either high or low. The elastic transform enlarges or reduces picture regions. Pixelate occurs when upsampling low-resolution images. Jpeg compression is a lossy image compression format that produces compression artifacts.

1.2 Training details.

Our model is implemented using PyTorch [14], and we initialize the event-based and frame-based backbone branches with pre-trained ResNet-50 [6]. The input data for the PKU-DDD17-Car dataset and DSEC dataset are resized to 346×260 and 640×480, respectively. Furthermore, we train the model using the Adam optimizer [10] with an initial learning rate of 1×10^{-4}. The experiments are carried out on an Nvidia RTX 3090 GPU with a system running Ubuntu 20.04. During training, the input from the RGB image is set to blank (zero) with a 15% probability. This strategy compels the sensor fusion model to extract information primarily from the second modality, the event camera, thereby enhancing the model's robustness against corruption in the frame-based camera. The total

Table 5: Comparison with SOTA fusion alternatives: COCO mAP@0.50:0.95 on the PKUDDD17-CAR dataset for the different methods.

Method	Model type	Test(All day)	Test(Day)	Test(Night)	FPS
SENet [7]	Attention	42.4	43.7	37.0	8.0
ECA [19]		40.8	42.2	36.1	7.6
CBAM [20]		42.8	44.2	38.0	10.3
SAGate [2]	RGB-D	43.4	44.9	38.0	11.8
DCF [8]		42.5	43.4	39.0	13.8
SPNet [25]		43.3	44.9	37.1	9.1
FPN-Fusion [18]	RGB-E	41.6	43.2	35.7	12.0
DRFuser [13]		42.4	43.3	38.8	11.5
RAMNet [4]		38.8	39.2	36.9	11.5
CMX [23]		39.0	40.2	35.4	2.8
FAGC [1]		42.4	43.7	36.7	5.3
RENet [26]		43.9	45.4	39.1	5.0
EFNet [16]		41.6	43.4	35.1	9.7
CAFR (Ours)		**46.0**	**46.9**	**42.1**	6.4

number of epochs is set to 200, with a batch size of 8 for the PKU-DDD17-Car dataset and 1 for the DSEC dataset, respectively.

2 More Experimental Analysis

Performance under different lighting conditions. To analyze the contribution of the event camera, we assess the performance gain under different illumination conditions. Detailed comparisons are provided in Tab. 5. The results illustrate that our proposed CAFR, capitalizing on the complementary nature of events and frames, consistently improves detection performance and is better than other fusion methods across various lighting conditions. This analysis provides valuable insights into the effectiveness of incorporating event data in improving overall model performance under diverse illumination scenarios.

Efficiency analysis. As detailed in Tab. 5, the running speeds of various methods are presented. In comparison with other fusion methods, such as CBAM [20], SAGate [2], and RENet [26], the proposed method exhibits comparable running speed while significantly enhancing detection accuracy.

More robustness analysis. Fig. 4 illustrates the relative performance under corruption (RPC) at severity levels ranging from 1 to 5. Across all models, there is a consistent decline in relative performance as the severity of corruption increases. Notably, the model relying solely on RGB data exhibits the steepest decline, indicating its lower robustness. In contrast, the proposed fusion method significantly improves robustness across different severity levels.

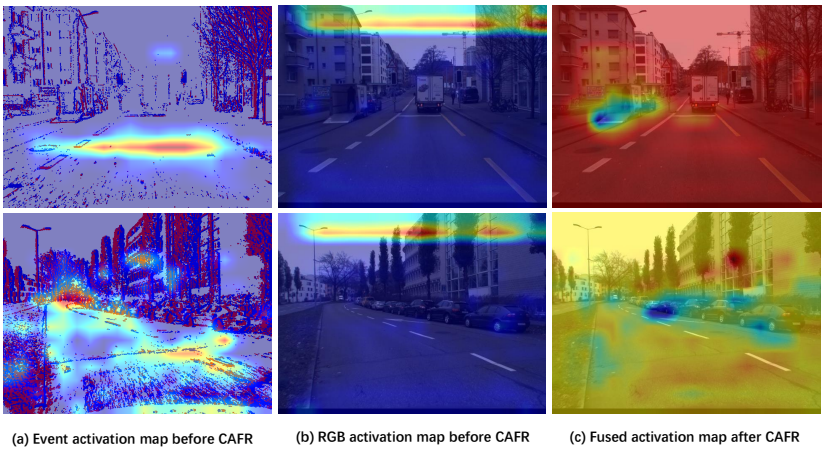

Fig. 5: Representative examples of different activation maps on the DSEC dataset are: (a) event activation map before CAFR; (b) RGB activation map before CAFR; and (c) fused activation map after CAFR.

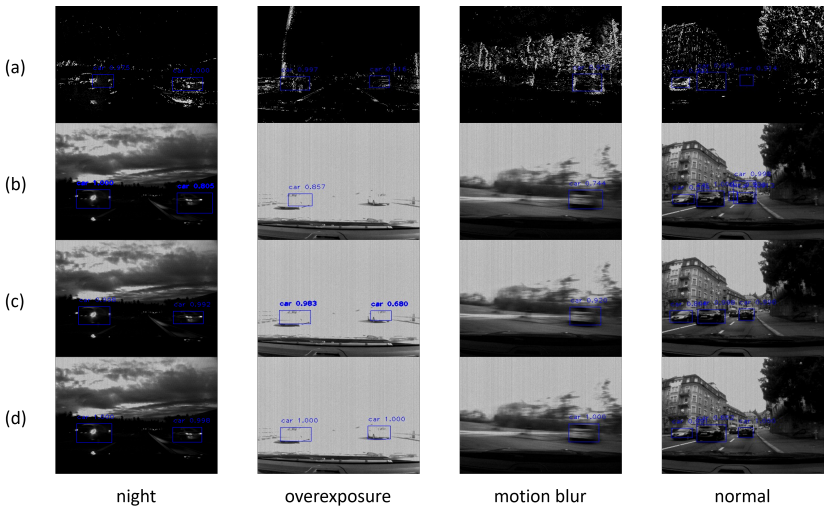

Fig. 6: Representative examples of different object detection results on the PKU-DDD17-Car dataset: (a) our baseline using event images; (b) our baseline using frames; (c) SPNet [25] (the second-best model in terms of mAP_{50}% and mAP%) using frames and events. (d) our method using frames and events.

Visualization of activation maps. In the fig. 5, we visualize the activation maps of RGB and event modalities before and after CAFR. After applying CAFR, the model demonstrates enhanced focus on significant regions.

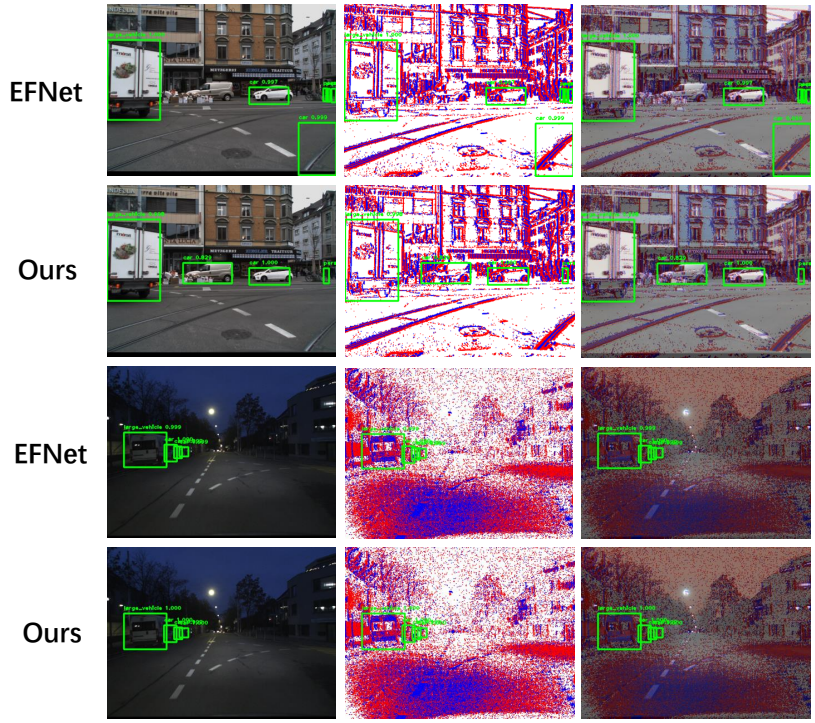

Fig. 7: Representative examples of different object detection results on the DSEC dataset. The first two rows are daytime scenes, and the last two rows are nighttime scenes. Each column represents RGB images, event frames, and merged RGB event frames, respectively.

Results on detection predictions. In Fig. 6 and Fig. 7, we visualize the detection results selected from the PKU-DDD17-Car dataset and DSEC dataset, respectively. The results demonstrate that the proposed method can consistently produce satisfactory detection results in various challenging scenarios. Compared with the second-best methods, SPNet [25] and EFNet [16], our method performs better prediction results.

3 Limitations

While our CAFR has demonstrated efficacy in the context of object detection, it is essential to acknowledge a current limitation. The scope of our evaluations has been confined to this specific task, and extrapolating the performance of CAFR to other prevalent perception tasks, such as semantic segmentation, depth prediction, and steering angle prediction, remains unexplored. Future investigations could delve into the broader applicability of CAFR, providing insights into its adaptability and potential limitations across diverse perception domains.

4 More Related Works

In other areas of cross-modal fusion, such as depth, thermal, and event data, we discuss relevant works below. In the field of RGB-D salient object detection (SOD), the joint learning and densely cooperative fusion framework introduced in [3] aims to improve performance. The authors of [12] developed graph-based techniques to design network architectures for RGB-D SOD. Additionally, an automatic architecture search approach for RGB-D SOD is presented in [17]. For semantic segmentation, the authors of [15] proposed the efficient scene analysis network (ESANet) for RGB-D semantic segmentation, utilizing channel attention [7] for RGB-D fusion. Furthermore, an uncertainty-aware self-attention mechanism is employed in [22] for indoor RGB-D semantic segmentation. The adaptive-weighted bi-directional modality difference reduction network proposed in [24] addresses RGB-T semantic segmentation. Recently, a multi-modal fusion network (EISNet) introduced in [21] aims to enhance semantic segmentation performance using events and images.

References

1. Cao, H., Chen, G., Xia, J., Zhuang, G., Knoll, A.: Fusion-based feature attention gate component for vehicle detection based on event camera. IEEE Sensors Journal **21**(21), 24540–24548 (2021)
2. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: ECCV (2020)
3. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: CVPR (2020)
4. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotics and Automation Letters **6**(2), 2822–2829 (2021)
5. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3), 4947–4954 (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
8. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al.: Calibrated rgb-d salient object detection. In: CVPR (2021)
9. Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., Ingham, F., Poznanski, J., Fang, J., Yu, L.U.: Yolov5: V3. 1-bug fixes and performance improvements. Zenodo (2020)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint (2014)
11. Li, J., Dong, S., Yu, Z., Tian, Y., Huang, T.: Event-based vision enhanced: A joint detection framework in autonomous driving. In: ICME (2019)
12. Luo, A., Li, X., Yang, F., Jiao, Z., Cheng, H., Lyu, S.: Cascade graph neural networks for rgb-d salient object detection. In: ECCV (2020)

13. Munir, F., Azam, S., Yow, K.C., Lee, B.G., Jeon, M.: Multimodal fusion for sensorimotor control in steering angle prediction. Engineering Applications of Artificial Intelligence **126**, 107087 (2023)
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
15. Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.M.: Efficient rgb-d semantic segmentation for indoor scene analysis. In: ICRA (2021)
16. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Van Gool, L.: Event-based fusion for motion deblurring with cross-modal attention. In: ECCV (2022)
17. Sun, P., Zhang, W., Wang, H., Li, S., Li, X.: Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: CVPR (2021)
18. Tomy, A., Paigwar, A., Mann, K.S., Renzaglia, A., Laugier, C.: Fusing event-based and rgb camera for robust object detection in adverse conditions. In: ICRA (2022)
19. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: CVPR (2020)
20. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV (2018)
21. Xie, B., Deng, Y., Shao, Z., Li, Y.: Eisnet: A multi-modal fusion network for semantic segmentation with events and images. IEEE Transactions on Multimedia pp. 1–12 (2024)
22. Ying, X., Chuah, M.C.: Uctnet: Uncertainty-aware cross-modal transformer network for indoor rgb-d semantic segmentation. In: ECCV (2022)
23. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems **24**(12), 14679–14694 (2023)
24. Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., Han, J.: Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In: CVPR (2021)
25. Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.P., Shao, L.: Specificity-preserving rgb-d saliency detection. In: ICCV (2021)
26. Zhou, Z., Wu, Z., Boutteau, R., Yang, F., Demonceaux, C., Ginhac, D.: Rgb-event fusion for moving object detection in autonomous driving. In: ICRA (2023)