# Advancing Object Detection for Autonomous Vehicles via General Purpose Event-RGB Fusion

Hajer Fradi, Panagiotis Papadakis

# Advancing Object Detection for Autonomous Vehicles via General Purpose Event-RGB Fusion

Hajer Fradi
*IMT Atlantique, Lab-STICC, UMR CNRS 6285*
*F-29238 Brest, France*
hajer.fradi@imt-atlantique.fr

Panagiotis Papadakis
*IMT Atlantique, Lab-STICC, UMR CNRS 6285*
*F-29238 Brest, France*
panagiotis.papadakis@imt-atlantique.fr

***Abstract*—Real-time vision applications such as object detection for autonomous navigation have recently witnessed the emergence of neuromorphic or event cameras, thanks to their high dynamic range, high temporal resolution and low latency. In this work, our objective is to leverage the distinctive properties of asynchronous events and static texture information of conventional frames. To handle that, asynchronous events are first transformed into a 2D spatial grid representation, which is carefully selected to harness the high temporal resolution of event streams and align with conventional image-based vision. Via a joint detection framework, detections from both RGB and event modalities are fused by probabilistically combining scores and bounding boxes. The superiority of the proposed method is demonstrated over concurrent Event-RGB fusion methods on DSEC-MOD and PKU-DDD17 datasets by a significant margin.**

***Index Terms*—Object detection, Event cameras, RGB, Fusion, Bayes.**

## I. Introduction

Event cameras are newly emerging bio-inspired sensors that asynchronously measure per-pixel brightness changes. Known also as neuromorphic sensors, they are different from conventional frame-based sensors, where images at a fixed frame rate are captured. Their resulting output is a stream of events encoding multiple types of information at once such as time, position, and sign of the brightness changes. Event sensors have several advantages compared to conventional cameras including high microsecond temporal resolution, very high dynamic range (140 dB vs. 60 dB), low power consumption, high pixel bandwidth (in the order of kHz) and robustness to motion blur [1]. Given all these advantages, event cameras have attracted considerable attention from academia to industry. They have particularly witnessed an increasing interest in autonomous vehicles [2]–[4], where rapid responses to sudden changes, adaptability to weather and illumination changes, and the acquisition of robust visual information at high speed are crucial.

This novel sensing paradigm, however, poses certain challenges that need to be accounted for. In contrast to conventional frame-based cameras, event cameras yield spatially sparse and asynchronous output data. Additionally, grayscale or color information produced by conventional cameras is substituted with binary information and a corresponding sign that together indicate per-pixel illumination changes. To match this unconventional output, current mainstream approaches consist in employing deep architectures that can associate sparse representations of event streams as in the case of Spiking Neural Networks (SNNs) and Graph Neural Networks (GNNs) [5]. However, it is worth noting that even though these algorithms preserve the sparsity aspect of event streams, they are computationally less efficient and are not as effective as conventional Deep Neural Networks (DNNs), consequently, they do not scale well to handle complex tasks [5].

Instead of developing deep learning models that could better match the unconventional output of event cameras, our main proposal in this paper is to leverage the high microsecond temporal resolution inherent in event sensors to construct multi-range representations. This representation aims to preserve temporal precision effectively, compared to common practices that involve aggregating polarity over a time interval [6]. Furthermore, the chosen representation acts as a 2D dense grid, enabling the utilization of conventional image-based vision methods, which would otherwise be incompatible with raw event streams. The chosen event representation could be significantly enriched by incorporating additional visual information derived from RGB inputs. This supplementary data includes details regarding color and texture, providing a more comprehensive understanding of the events being analyzed. Such a fusion of visible and event inputs has been investigated in different applications including visual SLAM, image super-resolution and depth estimation [7]–[10], but only few works for the task of object detection have been conducted.

Our contributions in this paper can be summarized as follows:

- An appropriate event representation, leveraging the high temporal resolution of event streams, is carefully chosen. This representation operates as a 2D dense grid, enabling the feeding of event data input to conventional deep neural networks.
- To handle and process events alongside frames, information from both modalities is fused together using a very late fusion scheme, which proves to be an effective solution for adapting to changing conditions.
- The joint detection framework, utilizing YOLOv7 as the baseline detector, achieves state-of-the-art performance in object detection on two publicly available datasets, surpassing other existing methods by a significant margin.

The remainder of the paper is organized as follows: Section II provides an overview of existing event representations and event-based object detectors, along with current fusion methods between events and frames. In the sequel, our proposed approach of fusion for event and RGB cameras for object detection is detailed in Section III. The conducted experiments and the obtained results are discussed in Section IV. Finally, we briefly conclude and give an outlook of possible future works in Section V.

## II. Related Work

### A. Event Representation

One important aspect of using event cameras is figuring out the most suitable solution to represent event streams in a specific task. According to existing literature, these representations can be mainly grouped into sparse and dense categories [5]. Sparse representations have the advantage of preserving the sparsity aspect of event streams. It is the case of SNNs, which are theoretically capable of learning end-to-end representations. In these networks, events are represented as spikes emitted at a certain time without the need of any pre-processing step. However, SNNs have not yet achieved the performance levels of DNNs due to the lack of specialized hardware and a computationally efficient backpropagation algorithm. For these reasons, most of reported SNN-based works are applied to classification tasks, but do not yet scale to more complex tasks [11]. Compared to SNNs, graph-based representations are more scalable and have demonstrated good performance across various tasks. However, they are still less accurate compared to DNNs in event-based vision [12]–[14].

Alternatively, to make event streams compatible with conventional computer-vision algorithms designed for standard images, dense grid-like representations currently bear a higher potential. The first attempts consist in integrating asynchronous events into a 2D spatial representation during a predefined temporal window. These early representations have been extensively applied in various computer vision applications [6], [15]. Dense representations include as well histograms, time surfaces or combinations of both [16]. Late dense representations attempt to capture more event information by stacking multiple time windows [17]. Since stacking events based on time can be problematic when the event rate is too high or too low depending on scene content and/or ego-motion, these observations have led to the introduction of stacking based on the number of events [18], [19].

### B. Event-based Object Detection

To overcome the limitations of conventional frame-based cameras in challenging situations, some event-based object detectors have recently been proposed. These detectors can be categorized based on their event representations. For example, a SNN composed of a spiking backbone and Single Shot Multibox Detector (SSD) bounding box regression heads is proposed in [20] to perform object detection on a real-world event dataset. Similarly, Zhang *et al.* introduced STNet in [21], where both temporal and spatial cues are dynamically extracted and fused. GNNs have been used in few other works, such as [22], where graph spectral clustering is employed to detect moving objects in event data.

Alternatively, some dense neural networks have been applied, often inferring detections from short temporal windows of events [23], [24]. Recent advancements in this field involve the incorporation of learned representations into end-to-end architectures. For instance, Perot *et al.* in [25] introduced a ConvLSTM-based object detection framework that integrates spatio-temporal information, starting with a pre-processing stage to construct tensor maps as dense event representations at each time step. Furthermore, Recurrent Vision Transformers have been employed as a novel backbone for object detection with event cameras in [26], aiming to reduce inference time while maintaining high performance. ASTMNet [27] is another learnable spatio-temporal representation from asynchronous events for object detection, featuring adaptive temporal sampling, a temporal attention convolutional network, and spatio-temporal memory modules.

### C. Fusion of Event and Intensity Information

Fusing event camera data with conventional frames is advantageous, as both provide complementary information, improving performance in tasks like visual SLAM, super-resolution, optical flow, depth estimation, object detection, and tracking [7]–[10], [28]–[30]. This section focuses on joint works for object detection, aligning with the scope of the paper. Few works have used such fusion schemes for object detection. Related work includes [31], where event-based and frame-based streams are incorporated into a neural network. SNN generates visual attention maps, and a joint decision is made using Dempster-Shafer theory. In [32], an LSTM processes events to learn salient objects, which are then fused with RGB inputs via multi-level feature fusion. [33] adopts as input a voxel grid representation for events and proposes two-parallel feature extractor networks for frames and events. Features from both inputs are combined before being fed to the FPN network. Another mid-level feature fusion scheme has been recently proposed in [17], where a temporal multi-scale aggregation module followed by a bi-direction feature fusion module are employed.

In contrast to prior works such as [17], [32], [33], we adopt a very-late fusion scheme based on Bayes' rule. This approach probabilistically combines scores and bounding boxes obtained from individual modalities. Additionally, instead of employing predefined time windows for event representation, as proposed in [17] at three time intervals, we capitalize on the high temporal resolution provided by asynchronous events within multi-range representations. The hyperparameters associated with this representation are automatically determined using Gromov-Wasserstein Discrepancy (GWD), a recently introduced discrepancy measure between raw events and the corresponding representations [5].

## III. METHOD

In this paper, we propose a joint detection framework that integrates both input RGB and event streams for detection. The overall proposed framework is depicted in Fig. 1, and the remainder of this section describes each of its components.

### A. Input Event Representation

Compared to conventional cameras capturing images at a fixed frame rate, event cameras respond to brightness changes for every pixel asynchronously and independently. This results in a stream of events that are spatially sparse and asynchronous. Each event can be expressed as $e_k = (x_k, y_k, t_k, p_k)$, where $x_k$ and $y_k$ denote spatial coordinates, $t_k$ represents the timestamp at which the event is triggered, and $p_k \pm 1$ is the polarity indicating the sign of the change. To align with conventional image-based vision, events need to be transformed into a 2D spatial grid representation. The transformation strategy aims to map asynchronous event streams onto image planes before applying frame-based object detectors. Various event representations have been developed in this direction, such as event frames [15], time surface and voxel grids [34], yet the challenge persists in determining the most effective representation.

In this study, we represent the event stream through a stacking method based on the number of events (SBN). This involves counting the number of events in reverse order from the current timestamp to a pre-defined number. This stacking strategy has demonstrated its effectiveness in prior research [10], [35] when compared to stacking based on time (SBT), which involves incorporating all events within a time period. In our specific task of object detection, the perceived object motion can vary, especially in real-world scenarios like driving cars or flying drones, where objects move at different speeds and the camera is mostly in motion as well. In such situations, stacking based on a pre-defined number of events could exclude some objects with low movement if the chosen number is small. In contrast, a large number of events may overwrite previous events in the case of highly moving objects.

For these reasons, we have opted for Mixed-Density Event Stacking (MDES), as presented in [19] for depth estimation. This representation involves using multiple event sequences with different event counts. This multi-channel representation is highly detailed without suppressing previous details or missing fine information. It proceeds by initially aggregating a large number of events $N$ and creating a multi-channel tensor by progressively reducing the number of events by half. For $M$ stacks, the number of events, denoted as $n_p$, is calculated as $n_p = N/2^{(p-1)}$ at each stack $p$, with $p \in \{1, ..., M\}$.

For the choice of some parameters such as the number of channels $M$ and the number of events $N$, we employ a recently proposed measure called Gromov-Wasserstein Discrepancy [5] which measures the distortion between an ordered set of raw events $\mathcal{E}_i = \{e_k\}$ and their representations $\mathcal{F}_i$. This measure has been shown to preserve the performance ranking for object detection task. For $K$ given samples, optimal parameters of $M$ and $N$ are determined as follows:

$$M^*, N^* = \arg\min_{M,N} \frac{1}{K} \sum_i L(\mathcal{E}_i, \mathcal{F}_i(M, N)), \tag{1}$$

where $L$ is the distortion function from raw events $\mathcal{E}_i$ to event representations $\mathcal{F}_i$, following the same notations used in [5].

### B. Very-Late Event-RGB Fusion

While the chosen representation proves useful for event-based object detection, integrating conventional frames could offer advantages by providing complementary static texture information. We anticipate that leveraging both sensor modalities would enhance the detection performance. For instance, in scenarios where no events occur, even if the representation retains some information, high-quality frames captured at a low frame-rate could complete the unclear or missing information from event streams. Additionally, event-based cameras have proven their edge in adverse conditions, while conventional cameras perform better in normal conditions. To deal with changing conditions, a fusion scheme that accounts for both modalities holds promise as an effective approach.

Strategies that address the integration of information from the two modalities can be categorized into early, mid, and late fusion approaches. Following [36], we apply the recently proposed very-late fusion, initially trained on separate RGB-thermal modalities through detector ensembling. This choice is motivated by the fact that it is a lightweight, non-learned method that has been successfully employed on two RGB-thermal benchmarks, showing good performance compared to prior works in the field.

The probabilistic ensembling is derived from Bayes' rule. Given multimodal measurements $x_1$ and $x_2$ from RGB and event inputs, an object label $y$ is inferred as follows:

$$\begin{aligned} p(y|x_1, x_2) &= \frac{p(y, x_1, x_2)}{p(x_1, x_2)} \\ &\propto p(y, x_1, x_2) \\ &\propto p(x_1|y)p(x_2|y)p(y) \\ &\propto \frac{p(y|x_1)p(y|x_2)}{p(y)}, \end{aligned} \tag{2}$$

where $\propto$ refers to "proportional to". Practically, pre-trained single-modal detectors predict the distributions over the label $y$: $p(y|x_1)$ and $p(y|x_2)$. The fused score is obtained by multiplying the two distributions and dividing by the class prior distribution. Similar to the fusion of the class posteriors, bounding boxes are probabilistically fused. For a given detection, the bounding box is defined as $z$ representing the center coordinates with the corresponding width and height. Single-modal detections outcome a posterior $p(z|x_i) \sim \mathcal{N}(\mu_i, \sigma_i^2 I)$, $i = 1, 2$, where $\mu_i$ are bounding box coordinates predicted from each modality and $\sigma_i^2$ determines the corresponding

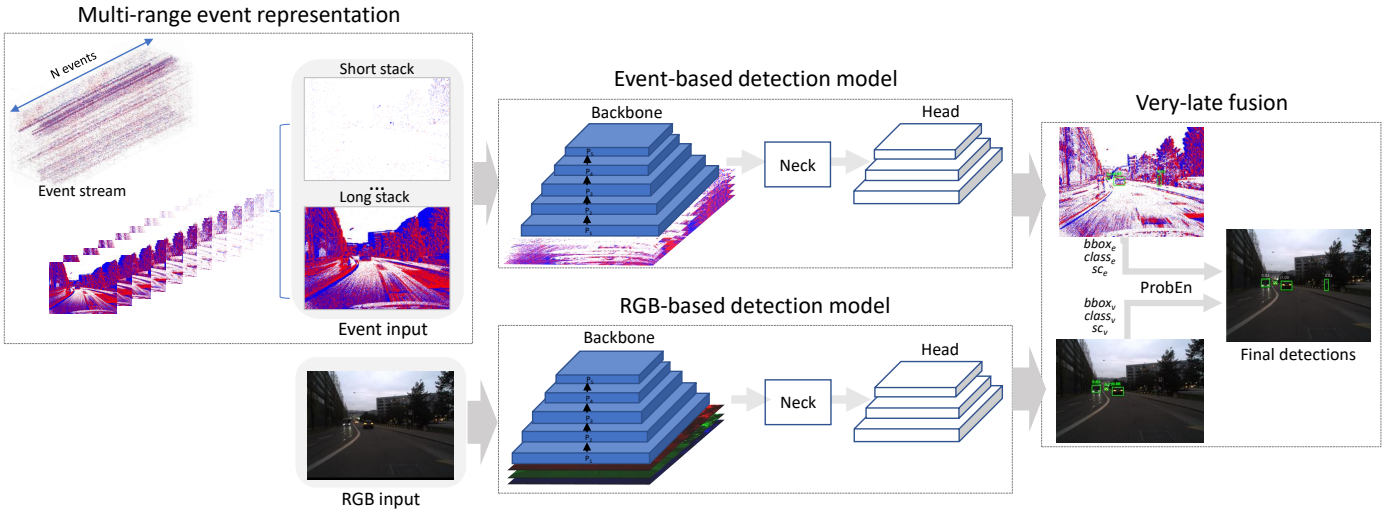



Fig. 1. The proposed flowchart for very late fusion of event and RGB detections, with YOLOv7 as the baseline detector. The input multi-range representation of event streams precedes the fusion process.

uncertainty degree. The bounding box coordinates are consequently fused as:

$$\begin{aligned} p(z|x_1,x_2) &= \frac{p(z,x_1,x_2)}{p(x_1,x_2)} \\ &\propto p(z|x_1)p(z|x_2) \\ &\propto \exp\left(\frac{\|z-\mu\|^2}{-2(1/\sigma_1^2+1/\sigma_2^2)}\right), \end{aligned} \tag{3}$$

where $\mu = \frac{\mu_1/\sigma_1^2+\mu_2/\sigma_2^2}{1/\sigma_1^2+1/\sigma_2^2}$. This means that $p(z|x_1,x_2) \sim \mathcal{N}(\mu, (1/\sigma_1^2+1/\sigma_2^2)^{-1}I)$ since the product of two Gaussian distributions is a Gaussian one. The fusion implies calculating a weighted average of box coordinates, with the weights determined by the inverse covariance.

## IV. Experiments

We conduct experiments on two public datasets **DSEC-MOD** and **PKU-DDD17**. We choose YOLOv7 [37] as a baseline detector, representing a good compromise between accuracy and time complexity. Given the generic nature of our work, any other detector could serve as an alternative. However, it is worth noting that while the key outcome lies in the joint late decision, achieving optimal overall performance depends on the careful selection of the detection model and event representation. Datasets, parameters, experiments and results analysis are detailed in the following subsections.

### A. Datasets

To validate the joint detection framework, we conduct experiments on two publicly available datasets in the field, namely, DSEC-MOD and PKU-DDD17. DSEC-MOD [17] is a new dataset dedicated to moving object detection, originally extracted from the widely known RGB-Event dataset DSEC [2]. The dataset contains moving objects belonging to a total of 8 classes with automatically labeled and manually checked annotations. In total, DSEC-MOD dataset contains 13314 frames of 640x480 resolution, with 10495 frames for training and 2819 frames for testing. PKU-DDD17 [31] is another synchronized RGB-Event dataset recorded with a 346x260 pixel DAVIS sensor. The dataset is manually annotated. It records data from highways and city driving scenarios under various timing conditions. The size of this dataset is relatively small, with 2241 frames for training and 912 frames for testing.

### B. Implementation Details

In the training phase, we use a mini-batch size of 64 for 30 epochs with DSEC-MOD dataset. For PKU-DDD17 dataset, the mini-batch size is 32, with 100 epochs. The training is conducted on NVIDIA RTX A6000 GPU. For event representation, we use 12 channels to stack 5 million events for DSEC-MOD dataset, following [19]. For PKU-Car dataset, we adopt a basic frame accumulation representation, as the multi-range representation is not feasible due to the limited information in the event data within short temporal window of only 20 ms. We modified the YOLOv7 code to support multi-channel inputs, accommodating both multi-range representations and the early fusion scheme by concatenating RGB and event streams. The detection performance is evaluated in terms of mean Average Precision (mAP) at Intersection Over Union (IOU) equal to 0.5 regarding the ground truth boxes.

For RGB and event synchronization, since the two experimented datasets provide annotations on calibrated RGB videos, we perform detections at the RGB frame rate. For event-based detections, event streams are sliced using the same temporal windows as the annotations. However, since the chosen representation is number-based and the target count may exceed the current slice, events are counted backward from the current timestamp. Detections between event and RGB are also synchronized by maintaining the same spatial resolution, as confirmed by close results in runtime comparisons.

### C. Comparison with State-of-the-Art

Our detection results on DSEC-MOD dataset are presented in Table I. Two settings of the baseline detector are considered for comparisons: RGB only using RGB images and Event only using event frames by aggregating polarity over a time interval. Additionally, we compare our proposed approach to the state-of-the-art RGB-Event fusion detectors, namely, FPN-Fusion [33], EFNet [38] and RENet [17]. Early fusion scheme is considered for comparisons by concatenating RGB and event channels as input for the YOLOv7 detector. The results of late fusion using non-maximum suppression and returning the average score are also compared.

| Methods | mAP @0.5 |
|---|---|
| RGB only | 0.55 |
| Event only (frame) | 0.36 |
| Event only (MDES) | 0.39 |
| FPN-Fusion [33] | 0.32 |
| EFNet [38] | 0.35 |
| RENet [17] | 0.38 |
| Early fusion | 0.50 |
| Late fusion | 0.56 |
| **Ours** | **0.58** |

TABLE I
DETECTION RESULTS ON DSEC-MOD DATASET IN TERMS OF MAP@0.5

As depicted in the table, our proposed method demonstrates the highest detection performance, achieving a mAP of 0.58 with a significant margin compared to the state-of-the-art fusion detectors and to both of early and late fusion schemes. The overall performance of our proposed method relies on three key factors: the use of YOLOv7 as baseline detection model, the appropriate choice of event representation, and the joint late decision strategy. First, the choice of YOLOv7 as baseline detection model is justified by the results obtained using each single modality. Second, the relevance of the chosen event representation is demonstrated by comparing the selected MDES representation to the event frame representation, resulting in a 3% improvement. Lastly, the joint late decision strategy contributes to the overall effectiveness of our method by combining insights from both modalities, further improving our detection reliability.

Similarly, we report our results on PKU-DDD17 dataset in Table II, where we compare our proposed method to single modality either using RGB only or event only data as input, as well as to early and late fusions. Additionally, we consider comparisons with the Joint Detection Framework (JDF) [31] on which the same dataset was initially tested. As shown in the table, our method exhibits superior performance, in both lighting conditions. It is worth noting that the significant difference of 0.40 in performance between RGB and event results can be largely attributed to the limited information provided by the event data within the short temporal window of only 20 ms in this dataset. Additionally, the comparisons between results in day and night conditions have to be carefully considered since the number of frames is not balanced between the two settings.

| Methods | day | night-fall | all |
|---|---|---|---|
| RGB only | 0.89 | 0.86 | 0.88 |
| Event only (frame) | 0.44 | 0.65 | 0.48 |
| JDF [31] | 0.91 | 0.83 | 0.84 |
| Early fusion | 0.73 | 0.83 | 0.75 |
| Late fusion | 0.91 | 0.88 | 0.90 |
| **Ours** | **0.92** | **0.89** | **0.91** |

TABLE II
DETECTION RESULTS ON PKU-DDD17 DATASET IN TERMS OF MAP@0.5 FOR DAY, NIGHT-FALL AND ALL CONDITIONS.

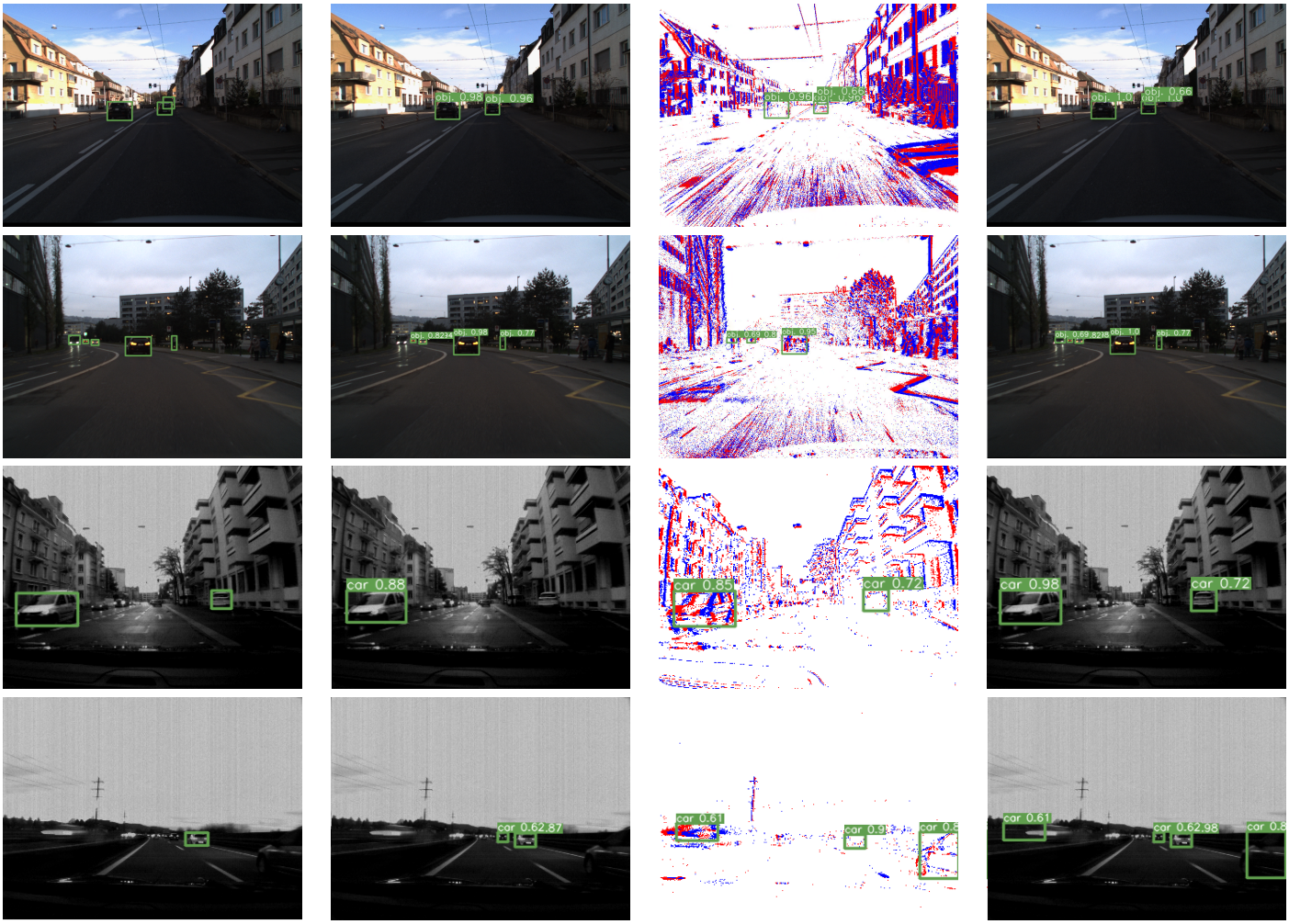
### D. Qualitative Results

In addition to quantitative results, we provide visual examples to further illustrate our proposed method's effectiveness. Figure 2 showcases four qualitative results comparing our method with single-modality results. As shown, our method can produce more accurate bounding boxes thanks to the complementary nature between the two sensors. For example, in the first row showing a sample frame from DSEC-MOD dataset, our method identified a true positive in the event data that the RGB data missed. When both inputs agree, the confidence scores increase, leading to more precise bounding box coordinates. This observation is shown in most samples, for example, in the second row of DSEC-MOD dataset. In the same example, the complementary aspect between the two inputs is demonstrated by showing improved detection results compared to those obtained with a single modality. In the third row, which shows a sample frame from PKU-DDD17 dataset, the RGB data results in a false negative because a car is mistaken for a part of the building. This error is corrected by the event data. In the fourth row, which represents a night-fall scenario on PKU-DDD17 dataset, a moving car is correctly detected by the event data despite the presence of motion blur. In such a situation, it is difficult to rely on RGB input to detect objects.

From the obtained results, it is clearly shown that it is highly advantageous for enhancing the performance of object detection to incorporate both inputs. Conventional frames play a crucial role when event streams lack texture information for precise recognition, while event streams become more useful in challenging scenarios where RGB images are affected. These findings substantiate our initial proposition regarding the complementary nature of events and frames, as outlined at the outset of this paper.

### E. Discussion and Limitations

Our discussion is mainly about data challenges. It is important to highlight the size of the experimented datasets remains relatively small compared to certain event-based datasets, such as the 1 Megapixel dataset [25], considered as one of the largest event-based autonomous driving datasets, containing over 25 million bounding boxes across seven object classes.




(a) Ground-truth (b) RGB results (c) Event results (d) Fusion results

Fig. 2. Qualitative results of our proposed method compared to single modality results and to the ground-truth annotations. From top to bottom: 2 results on DSEC-MOD, PKU-DDD17 (day), and PKU-DDD17 (night-fall), respectively.

Unfortunately, the corresponding RGB data has not been made publicly available, limiting its usefulness for broader research.

Following the previous discussion on data challenges, another issue concerns the failure cases encountered during the annotation process for event-based datasets. These cases include geometric errors from imprecise rectification between event and RGB inputs, as well as some missing detections when objects are not clearly visible in the annotated images but can be seen in the event streams. The last sample in Figure 2 illustrates this situation. In this example, some annotations are missing because the cars are not clearly visible in the images where the annotations were made. However, the event data provide clearer information about these objects. In the particular case of moving object detection using DSEC-MOD dataset, there are instances of erroneous boxes when a slowly moving car is inaccurately identified as a static object.

Another potential extension of this study is to develop a quantized model of YOLOv7. Under the same hardware configuration, the current average frame rate on DSEC-MOD is 8 fps. The runtime is shorter on PKU dataset due to a lower number of events, fewer channels, and lower resolution for PKU-DDD17. While the current runtime is satisfactory, further energy and resource savings can be achieved by implementing the solution on embedded systems using model compression techniques like parameter quantization.

## V. Conclusion

In this paper, we propose a very late Event-RGB fusion detection framework, with key outcomes derived from the joint late decision, the relevance of the chosen event representation, and the appropriate choice of the baseline detection model. These factors collectively contribute to the framework overall performance, which has been demonstrated on commonly used datasets compared to state-of-the-art methods. As discussed at the end of the paper, the development of fusion networks using event and RGB cameras has been relatively hindered by the lack of large-scale datasets with accurate annotations. Therefore, creating new, larger-scale datasets could significantly contribute to further advances in this field.

## ACKNOWLEDGEMENT

This work has received a French government support granted to the Labex CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference ANR-10-LABX-07-01 from September 2022 to December 2024. In this program, the project associated to this work is *LEASARD* (Low-Energy deep neural networks for Autonomous Search-And-Rescue Drones). This work has also received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 899546.

## REFERENCES

[1] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[2] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.

[3] N. J. Sanket, C. M. Parameshwara, C. D. Singh, A. V. Kuruttukulam, C. Fermüller, D. Scaramuzza, and Y. Aloimonos, "Evdodgenet: Deep dynamic obstacle dodging with event cameras," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 651–10 657.

[4] D. Falanga, K. Kleber, and D. Scaramuzza, "Dynamic obstacle avoidance for quadrotors with event cameras," *Science Robotics*, vol. 5, no. 40, p. eaaz9712, 2020.

[5] N. Zubić, D. Gehrig, M. Gehrig, and D. Scaramuzza, "From chaos comes order: Ordering event representations for object recognition and detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 846–12 856.

[6] L. Wang, Y.-S. Ho, K.-J. Yoon *et al.*, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 081–10 090.

[7] A. Safa, T. Verbelen, I. Ocket, A. Bourdoux, H. Sahli, F. Catthoor, and G. Gielen, "Fusing event-based camera and radar for slam using spiking neural networks with continual stdp learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2782–2788.

[8] Y. Lu, Z. Wang, M. Liu, H. Wang, and L. Wang, "Learning spatial-temporal implicit neural representations for event-guided video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 1557–1567.

[9] Y.-F. Zuo, L. Cui, X. Peng, Y. Xu, S. Gao, X. Wang, and L. Kneip, "Accurate depth estimation from a hybrid event-rgb stereo setup," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6833–6840.

[10] M. Mostafavi, K.-J. Yoon, and J. Choi, "Event-intensity stereo: Estimating depth by the best of both worlds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4258–4267.

[11] M. Yao, H. Gao, G. Zhao, D. Wang, Y. Lin, Z. Yang, and G. Li, "Temporal-wise attention spiking neural networks for event streams classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 221–10 230.

[12] Y. Li, H. Zhou, B. Yang, Y. Zhang, Z. Cui, H. Bao, and G. Zhang, "Graph-based asynchronous event processing for rapid object recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 934–943.

[13] S. Schaefer, D. Gehrig, and D. Scaramuzza, "Aegnn: Asynchronous event-based graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 371–12 381.

[14] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based spatio-temporal feature learning for neuromorphic vision sensing," *IEEE Transactions on Image Processing*, vol. 29, pp. 9084–9098, 2020.

[15] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5419–5427.

[16] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "Hats: Histograms of averaged time surfaces for robust event-based object classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1731–1740.

[17] 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)*. IEEE, 2023, pp. 7808–7815.

[18] M. Mostafavi, Y. Nam, J. Choi, and K.-J. Yoon, "E2sri: Learning to super-resolve intensity images from events," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 6890–6909, 2021.

[19] Y. Nam, M. Mostafavi, K.-J. Yoon, and J. Choi, "Stereo depth from events cameras: Concentrate and focus on the future," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6114–6123.

[20] L. Cordone, B. Miramond, and P. Thierion, "Object detection with spiking neural networks on automotive event data," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[21] J. Zhang, B. Dong, H. Zhang, J. Ding, F. Heide, B. Yin, and X. Yang, "Spiking transformers for event-based single object tracking," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 8801–8810.

[22] A. Mondal, J. H. Giraldo, T. Bouwmans, A. S. Chowdhury *et al.*, "Moving object detection for event-based vision using graph spectral clustering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 876–884.

[23] N. F. Chen, "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*, 2018, pp. 644–653.

[24] M. Iacono, S. Weber, A. Glover, and C. Bartolozzi, "Towards event-driven object detection with off-the-shelf deep learning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[25] E. Perot, P. De Tournemire, D. Nitti, J. Masci, and A. Sironi, "Learning to detect objects with a 1 megapixel event camera," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 639–16 652, 2020.

[26] M. Gehrig and D. Scaramuzza, "Recurrent vision transformers for object detection with event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 884–13 893.

[27] J. Li, J. Li, L. Zhu, X. Xiang, T. Huang, and Y. Tian, "Asynchronous spatio-temporal memory network for continuous event-based object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2975–2987, 2022.

[28] J. Zhang, X. Yang, Y. Fu, X. Wei, B. Yin, and B. Dong, "Object tracking by jointly exploiting frame and event domain," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 043–13 052.

[29] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu, "Visevent: Reliable object tracking via collaboration of frame and event flows," *IEEE Transactions on Cybernetics*, 2023.

[30] G. Paikin, Y. Ater, R. Shaul, and E. Soloveichik, "Efi-net: Video frame interpolation from fusion of events and frames," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 1291–1301.

[31] J. Li, S. Dong, Z. Yu, Y. Tian, and T. Huang, "Event-based vision enhanced: A joint detection framework in autonomous driving," in *2019 ieee international conference on multimedia and expo (icme)*. IEEE, 2019, pp. 1396–1401.

[32] X. Jiang, L. Zhu, and H. Tian, "Learning event guided network for salient object detection," *Pattern Recognition Letters*, vol. 151, pp. 317–324, 2021.

[33] 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.

[34] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 989–997.

[35] J. Choi, K.-J. Yoon *et al.*, "Learning to super resolve intensity images from events," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2768–2776.

[36] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *European Conference on Computer Vision*. Springer, 2022, pp. 139–158.

[37] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[38] 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.