2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Attentional Event-RGB Sensor Fusion for Fast Drone Detection

Antoine Zundel[1,2], Cédric Demonceaux[2], Nicolas Hueber[1], Guillaume Strub[1], Sébastien Changey[1]

***Abstract*— This paper presents an embedded multi-modal vision system for drone detection, combining an event-based camera, an IMU, and an RGB sensor. The method leverages an attentional mechanism on the event stream and is robust to rotations along all three axes (roll, pitch, and yaw) of a rotating platform. The event-based sensor enables localization of fast moving objects, while the RGB camera provides classification, with the entire system optimized for embedded computational constraints. Performance analysis, with and without attention mechanisms and across various algorithmic variants, assesses the trade-off between computational cost and detection accuracy. The study identifies optimal operating situation for each configuration, validated on an outdoor test data samples.**

I. INTRODUCTION

Drones, with their high agility, versatility, and speed, are increasingly being adopted across a multitude of fields. However, their widespread availability has also led to unauthorized and malicious uses, necessitating the development of detection and interception methods for unmanned aerial vehicles (UAVs). Among the crucial tasks for UAV countermeasures are detection, classification and tracking. Accurately identifying the type and intentions of drones as soon as possible is essential to determine appropriate responses and to allow sufficient time for reaction (interception).

In this context, this work focuses on detecting fast-moving objects using a bimodal system combining the high temporal resolution of the event sensor with the high spatial and color information of the RGB sensor, to reduce the latency of the situational analysis to allow more time for an appropriate response.

II. PREVIOUS WORK

A. Event-based object detection

Recent studies have explored object detection using event sensors and RGB fusion. They can be classified as follows.

a) Event-Only Methods: Approaches relying solely on event-based sensors, such as [10] and [1], demonstrate strong performance in high dynamic range (HDR) conditions and they offer a clear advantage for fast motion detection compared to traditional RGB sensors where standard RGB sensors typically fail. However, these methods tend to lose effectiveness for detecting small and slow-moving objects under standard conditions.

[1]French-German Research Institute of Saint-Louis (ISL), France. antoine.zundel@isl.eu
[2]Université Bourgogne Europe, CNRS, Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), UMR 6303, 21000 Dijon, France.

b) Bimodal Fusion: Another solution is to combine the high temporal resolution of event sensors with the spatial and color information provided by RGB sensors. Recent works, such as [15], [5], [9], and [19], present fusion methods based on deep learning methods. These approaches achieve strong detection performance, particularly in HDR and for fast-moving objects, outperforming unimodal methods. However, high computational costs , and the need for precise sensor calibration limit the use of these techniques for embedded systems and fast drone detection.

c) Attentional Systems: Previous works [8], [20] investigated collaborative sensors to enhance real-time processing: An event-driven attention mechanisms guides RGB processing to reduce computational load. These methods achieve low-latency tracking or detection in low background and/or static scenarios, but falter under ego-motion due to the huge number of events generated.

To address this gap, we propose an enhanced attentional mechanism that dynamically compensates for ego rotational movements and returns areas where objects are in motion. Our approach is grounded in the observation that aerial threats (e.g., other UAVs) are often in motion.

B. Ego-Motion Filtering

Several works have addressed background filtering in event-based sensors under ego-motion, typically falling into two main categories.

a) Optimization and AI-based methods: [11], [14], [18] aim to distinguish moving objects from the background and remain effective under full 6-DoF motion, but require significant computational resources.

b) Sensor-fusion approaches: [17], [3], [16] integrate event cameras with IMUs, using gyroscope data to compensate for rotation, offering linear complexity suitable for embedded systems, albeit limited to 3-DoF correction.

In both cases, event-based motion compensation often yields a partial object reconstruction, mainly highlighting high-event-rate regions (e.g., front and rear along motion direction) rather than capturing the complete structure.

C. Our contributions

This work presents an attentional methods combining three sensors, event-based and RGB sensor with IMU, to reduce computation time for drone detection. Inspired by a biological visual system, the event-based sensor provides

rapid detection of moving areas at low computational cost, while the RGB sensor, offering higher spatial resolution but at a higher processing cost, focuses on the identified regions of interest. To the best of our knowledge, we present the first implementation of an event-based attentional mechanism robust to rotations around all three axes, designed to guide the detection and classification module on an RGB sensor. The three-modal platform is shown in figure 1a.

III. The Bimodal Vision System

A. Software

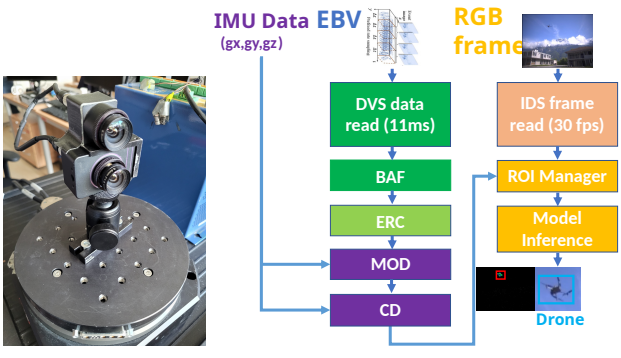
The proposed system enhances our previous framework [20] by completing the event processing pipeline with 4 modules : the Reading module, the Ego Rotation Compensation (ERC), the Moving Object Detection (MOD) and the Clustering Detection module (CD). The complete pipeline is illustrated on figure 1b.

- The Reading module reads events from the sensor in 11 ms packets, applying Background Activity Filtering (BAF) [7] to remove sensor noise. The BAF parameter settings are: minimum neighboring active pixels = 3 and no maximum density threshold. The choice of 11 ms packets is motivated by its proximity to time windows commonly used in the literature [16], providing sufficient spatial coherence and enabling easier synchronization with the RGB frames at 30 Hz, allowing one event packet per image to be processed. The results of this module is presented in figure 2a.
- The Event Rotation Compensation (ERC) module compensates sensor motion using the method from [16]. It employs IMU rotation rates for rotation compensation and computes a rotation-compensated event list.
- The Motion Object Detection (MOD) module, adapted from [16], generates normalized time images from motion-compensated events. A dynamic thresholding step detects moving regions; the threshold (Th) is defined as a linear function of scene dynamics:

$$Th = a \times |W| + b \tag{1}$$

Where a and b are constants determined empirically ($a = 0.15$; $b = 0.03$) and W is the angular velocity over the three IMU axes expressed in the event sensor reference frame. Morphological filtering (3x3 open filter) reduces false positives and noise, producing a spatial mask that can isolates events corresponding to moving object.

- The Cluster Detection module (*CDgrid*) applies a 2D grid-based spatial clustering [13] to the filtered event stream, with parameters set to a minimum of one event for an active cell, a cell size of 32×32 px, and a merge distance of two cells (64 px in the event domain). This distance closely matches the projected YOLOv5 [6] input size (160×160 px), enabling efficient grouping of events into coherent moving-object clusters. Unlike the method employed in [16], which often produces partial object reconstructions, our grid strategy dilates and merges clusters, reducing redundant ROIs for the same object and improving event aggregation. Here, the objective is not to maximize the intersection-over-union (IoU) between the generated ROI and the moving object, but rather to maximize the probability that the object is fully contained within the ROI (Fig.2b).
- The ROI Manager projects detected clusters into the RGB frame and generates fixed-size ROIs equal to the model's input size. This resolution was chosen as it preserves sufficient detail for reliable classification without degrading detection performance on objects larger than the YOLOv5 [6] model's input size. Larger clusters produce ROIs matching the largest dimension of their bounding box to ensure complete coverage, while smaller clusters are dilated to model's input size, ensuring both full object representation and compatibility with the network input. These steps are illustrated on Fig. 2c.
- Classification – The extracted ROIs are processed by a YOLOv5-medium model [6] in ONNX format with input size 160×160 px, which outputs the object class (drone, bird, plane, helicopter, etc.) and refines object localization. The results of this module is presented in figure 2d.



(a) (b)

Fig. 1: General system architecture. (a) Acquisition platform comprising RGB and event-based sensors with integrated IMU in the DVXplorer. (b) Software architecture with rotation filtering compensation based on IMU.

IV. Moving Object Detection Performance in Static and Ego-Rotation Scenarios

A. Experimental Setup and Dataset Acquisition

To validate and quantify the effect of motion compensation on attentional mechanisms under both rapid sensor rotations and static conditions, we conduct dedicated test acquisitions.

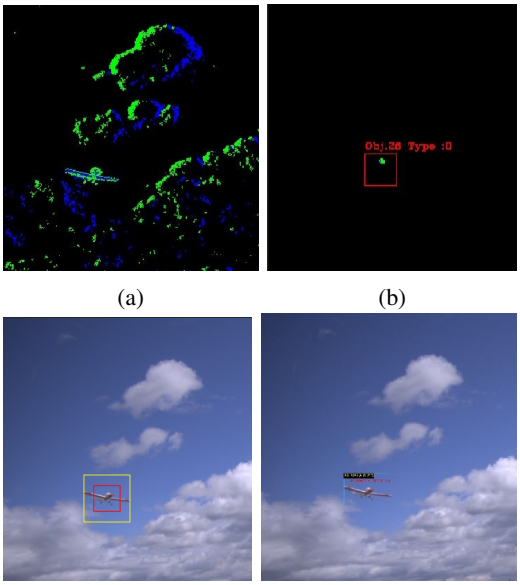


Fig. 2: Detection results during outdoor tests in the rotation scenario. (a) Event accumulation prior to motion compensation. (b) After *CDgrid* clustering, clusters detected are indicated by red bounding boxes. (c) Output of the ROI manager: red bounding boxes correspond to *CDgrid* clusters in the RGB reference frame; yellow bounding boxes represent the predicted regions of interest (ROI) forwarded by the ROI manager to the YOLO model. (d) Final detection result: blue bounding box identifies the airplane as detected by the YOLO detection model.

A dedicated acquisition system has been developed for experimental validation.

a) Acquisition system: The acquisition setup presented in Fig. 1a consists of an IDS RGB camera (resolution: 2056×1542 px) and a DVXplorer event-based sensor [4] (resolution: 640×480 px) equipped with an IMU. Both sensors have identical optics with a focal length of 5.5 mm, providing a horizontal field of view (FoV) of 55°. The three sensing modalities are calibrated following the procedure described in [12]. For a fair baseline comparison, only the overlapping FoV between the two sensors (event and RGB) was considered. In the RGB camera, this corresponds to a 1712×1289 px subregion, closely matching the FoV of the event sensor and providing an effective spatial resolution approximately 2.7 times higher in the RGB modality.

b) Dataset acquisition design: The dataset has been collected using a model airplane (Bidule ECOTOP) flying towards the acquisition system at varying speeds, resembling a small aircraft detectable by the YOLOv5 model. Two experimental conditions are evaluated:

- Static sensor — The acquisition system remained stationary during the approach of the drone.
- Rotating sensor — The acquisition system executed ±50° scanning motions with sinusoidal modulation and a maximum yaw rate of 60°/s.

c) Annotation and distance estimation: The distance of the drone from the event sensor was estimated using its known physical dimensions (length = 1.80 m, wingspan = 2.46 m) and the size of the box of limits in pixels annotated in RGB frames.

The following methods have been compared:

- *ERC+MOD+CDgrid*: Our method.
- *ERC+MOD+CDdbscan*: Our method with DBSCAN (Density-Based Spatial Clustering of Applications with Noise) described in [2] clustering algorithm applied to event data, replacing grid-based clustering (*CDgrid*) to evaluate clustering impact.
- *CDgrid* (without ERC+MOD): removing rotation compensation to assess its contribution, retaining event filtering (BAF) and grid-based clustering.
- *SW+CDdbscan* baseline: Attentional mechanism from [20], using DBSCAN clustering and a BAF filter on high-density events within a small time window equal to 1 ms (SW), without rotation compensation.
- *full_image*: Standard YOLOv5 medium (trained on the same dataset and exported in ONNX format, input size 1696×1248 px) applied to RGB frames cropped to the event sensor's FoV (1712×1289 px), representing conventional frame-based object detection without event-driven preprocessing.

B. Detection performance results

a) Method: To evaluate detection performance in both scenarios, all methods were assessed by computing the F1-score as a function of the drone–sensor distance. The estimated distance was discretized into 10 m intervals, and the F1-score was computed separately for each interval using an IoU threshold of 0.3 (higher scores indicate better performance). Results are presented as histograms, one for each scenario on figure 3 and 4.

b) Static scenario performance: The comparison of the five methods on the static scenario, shown on figure 3, reveals several key trends. Without any event-based attentional mechanism, i.e. with inference directly on the full RGB frame (*full_image* method) the first detection occurs at approximately 45 m, with an F1-score of 0.29. For distances below 30 m, the system achieves an F1-score of 1.0, as the drone occupies a large portion of the image and can be classified reliably.

Among attentional-mechanism approaches without rotation compensation, *CDgrid* achieves detection performance similar to the *full_image* method and outperforms other event-based methods in this static-background scenario. This is because motion compensation is unnecessary when the

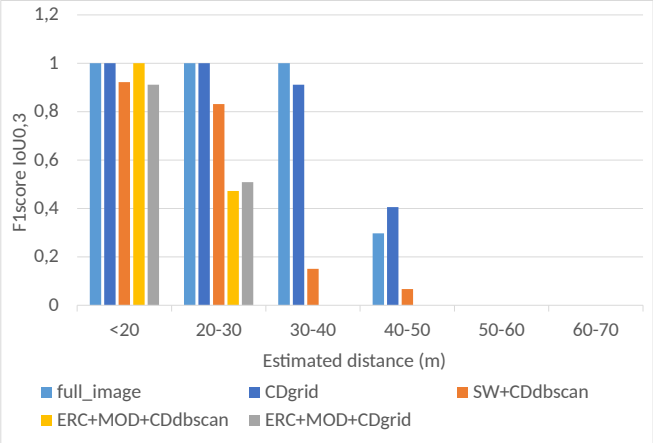


Fig. 3: F1-score comparison in function of the distance of the drone in static acquisition scenario for several methods.

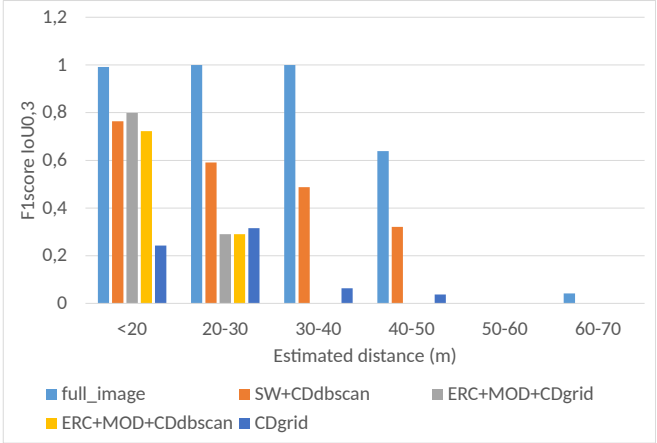


Fig. 4: F1-score comparison in function of the distance of the drone in rotation acquisition scenario for several methods.

background is static: high-density event regions can be directly considered as potential moving objects. *CDgrid* also outperforms *SW+CDdbscan* [20], primarily because the latter uses a very short 1 ms event-integration window. Such a small window produces fewer events, reducing the spatial coherence for slowly moving small objects at long distances. Increasing the integration window improves spatial coherence, aiding the separation of events corresponding to moving objects that their from noise. The method in [20] is better suited to detecting very fast targets; in this case, the drone's approach direction reduces its apparent speed in the sensor frame, further limiting the number of detected events. Its detection performance improves progressively thereafter, but it never reaches the accuracy obtained by full-image inference.

c) Rotation scenario performance : The comparison of the five methods in rotation scenario, shown in Fig. 4, reveals several key trends : rotation scenario *full_image* inference achieves the best performance, with good performance of detection at 45 m (F1 = 0.61) and near-perfect accuracy below 40 m. Ref [20] (*SW+CDdbscan*) without rotation compensation ranks second (first detection F1 = 0.32) but never matches full-image accuracy. *CDgrid* clustering without rotation compensation performs poorly ($F1 \leq 0.31$) due to large ROIs being resized, reducing accuracy, whereas *SW+CDdbscan* maintains higher performance by generating more, smaller ROIs. Rotation-compensated methods *ERC+MOD+CDgrid* and *ERC+MOD+CDdbscan* achieve similar results to the static case, peaking at F1 = 0.79 for ≤ 20 m, but with shorter first-detection distances due to need for more spatial information.

C. Computational Performance Evaluation

The performance evaluation is conducted on a NVIDIA Jetson Orin Nano platform, equipped with a 6-core CPU and an integrated GPU. All compared methods is implemented in C++. YOLOv5 inference is executed using GPU acceleration, while the event-based attention system is implemented as a multithreaded CPU process.

1) Hardware-in-the-loop Simulation : A stream of events, randomly distributed in space, is simulated to measure the processing time of each module for a fixed number of input events. This setup enables validation of algorithmic complexity independently of sensor constraints.

Figure 5 shows that the measured runtimes align with the expected theoretical complexities.

- Event Rotation Compensation (ERC): $\mathcal{O}(N)$
- Motion Object Detection (MOD): $\mathcal{O}(N + R)$
- Clustering Detection (CDgrid): $\mathcal{O}(N + C^2)$

where N is the number of processed events, R is the sensor resolution, and C is the number of active cells in the grid-based clustering stage.

The grid-based clustering approach outperforms DBSCAN [2], which has $\mathcal{O}(N^2)$ complexity, providing superior robustness and scalability for high volume events.

The MOD module exhibits a higher variance in processing times, mainly due to temporal event ordering that causes cache misses.

We evaluated the inference time of the YOLOv5-medium model (ONNX format) using two different input sizes:

- Attentional mechanism size: $160 \times 160 \rightarrow 25.6\,\text{ms}$ average
- Full-image inference input size: $1696 \times 1248 \rightarrow 365.7\,\text{ms}$ average

This corresponds to a 14.3× speed-up. The results also indicate that, to achieve real-time operation (30 ms per frame), the system can process up to 200,000 events per packet per thread. This is achieved through parallel processing across

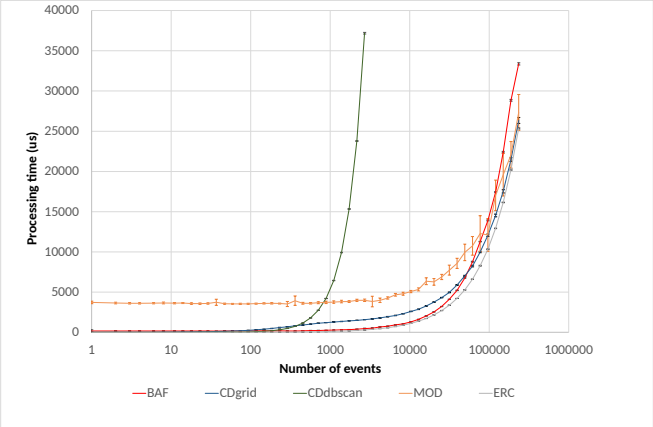


Fig. 5: Computational time of each processing module function of the number of events.

five threads dedicated to event reading/filtering, *ERC*, and the *MOD+CDgrid* pipeline, plus one dedicated thread for ROI management and YOLOv5 inference.

2) Static and rotating scenarios: The average delay of each methods is computed, considering only time intervals when at least one inference model is active coresponding to periods where lest one regions of interest were identified by the attentional system to avoid biasing comparisons for methods that do not involve inference.

a) Static Scenario: In static conditions, the event sensor generates on average only 0.3 k events per packet. Figure 6 and Table I show that methods *ERC+MOD+CDgrid*, *ERC+MOD+CDdbscan*, and *CDgrid* process approximately 3,300 kevents every 11 ms, while *SW+CDdbscan* handles ∼300 events every 30 ms.

The figure 6 shows that the process can run without adding delay.

In this context:

- *SW+CDdbscan* achieves the highest efficiency (15× faster than other methods) because DBSCAN complexity is mitigated by the low number of events, with very few false positive ROIs.
- Rotation-compensated methods ERC+MOD+CDdbscan and *ERC+MOD+CDgrid* display low computational cost, reducing inference time by ∼12× compared to full-image inference and halving variance relative to *CDgrid*.
- Full-image processing remains significantly slower than all event-based approaches, despite low variance.

b) Rotation Scenario: In rotation, the sensor produces ∼5 M events/s (about 100 k events per packet), increasing processing demands. Figure 7 and Table II show that:

- Methods without motion compensation such as *SW+CDdbscan* see drastic inference time increases, approaching full-image performance with a huge variance (std. dev. ≈ 237 ms) and worst-case time up to 965 ms, corresponding to multiple ROIs (∼40) processed sequentially.
- *CDgrid* uses a longer temporal window to merge events into larger clusters, reducing ROI count and maintaining average latency close to MOD-based pipelines. However, clustering over larger windows reduces the detection accuracy.
- Motion-compensated methods (*ERC+MOD+CDgrid*, *ERC+MOD+CDdbscan*) achieve better robustness: lower variance, smaller worst-case times, and improved real-time stability. Among them, grid-based clustering is faster than DBSCAN thanks to its $\mathcal{O}(N + C^2)$ complexity.

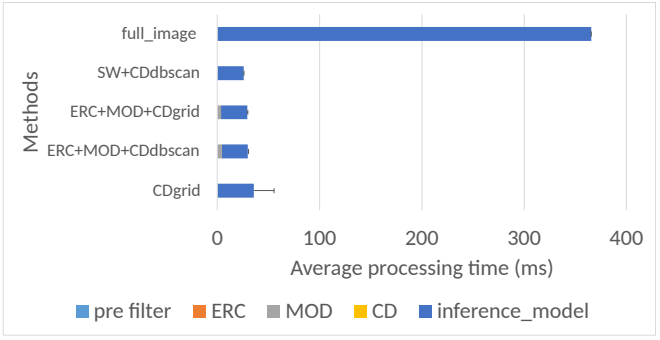


Fig. 6: Latency distribution on the static scenario for different methods

TABLE I: Mean, standard deviation, and worst-case execution times in the static scenario in (ms).

Methods	Mean	Std. dev	Worst case (ms)
CDgrid	35.57	19.79	134.27
ERC+MOD+CDdbscan	29.68	0.74	59.86
ERC+MOD+CDgrid	29.34	0.53	40.27
SW+CDdbscan	25.76	0.02	37.58
Full_image	365.74	1.12	377.21

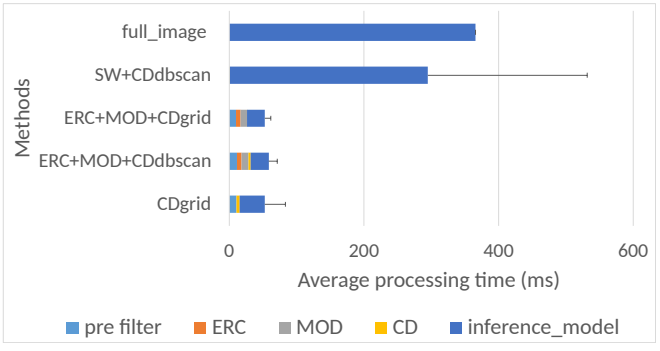


Fig. 7: Latency distribution on the rotation scenario for different methods

TABLE II: Mean, standard deviation, and worst-case execution times in the rotation scenario in ms.

Methods	Mean	Std. dev	Worst case (ms)
CDgrid	48.01	30.74	363.51
ERC+MOD+CDdbscan	55.58	12.56	160.84
ERC+MOD+CDgrid	52.71	8.74	139.44
SW+CDdbscan	294.40	237.08	965.91
Full_image	365.4	1.12	365.74

c) Key Insights: From both scenarios, three conclusions can be drawn:

- In low-event static conditions, *CDgrid* without (ERC+MOD) can be particularly efficient (very low computational cost, equivalent detection performace to full image).
- Under high rotation and event rates, motion compensation becomes critical for maintaining low latency.
- *ERC+MOD+CDgrid* shows good temporal performance, but its short detection range results from the MOD module and the event sensor's low resolution. This scenario where the drone moves directly toward the sensor is the worst case for motion detection and can greatly reduce the detection distance.

V. Conclusion and perspectives

This paper presents an embedded multi-modal vision system combining event-based and RGB sensing for efficient drone detection on low-power hardware. In static conditions, a simple attentional module (CDgrid) achieves comparable detection accuracy with significantly reduced computation time. By exploiting event-driven attention and motion compensation, the system reduces inference latency under ego-rotation. Despite its power efficiency, detection range remains limited in rotational scenarios due to ego-motion filtering. Future work will enhance the attentional mechanism and extend motion compensation to translation for improved robustness and long-range performance.

Acknowledgments

The authors wish to thank the French Agence de l'Innovation de Défence (AID) for its financial support. The authors also wish to thank Rémy Sautot and Damien Delmas for their help with the acquisitions and for piloting the drones.

References

[1] Loïc Cordone, Benoît Miramond, and Philippe Thierion. Object Detection with Spiking Neural Networks on Automotive Event Data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[2] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[3] Davide Falanga, Kevin Kleber, and Davide Scaramuzza. Dynamic obstacle avoidance for quadrotors with event cameras. *Science Robotics*, 5(40):eaaz9712, 2020.

[4] Inivation. Specifications – current models inivation. https://inivation.com/wp-content/uploads/2021/08/2021-08-iniVation-devices-Specifications.pdf.

[5] J. Cao, X. Zheng, Y. Lyu, J. Wang, R. Xu, and L. Wang. Chasing Day and Night: Towards Robust and Efficient All-Day Object Detection Guided by an Event Camera. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9026–9032, May 2024.

[6] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO. https://github.com/ultralytics/ultralytics, January 2023.

[7] A. Linares-Barranco, F. Gómez-Rodríguez, V. Villanueva, L. Longinotti, and T. Delbrück. A USB3.0 FPGA event-based filtering and tracking framework for dynamic vision sensors. In *2015 IEEE international symposium on circuits and systems (ISCAS)*, pages 2417–2420, 2015.

[8] Hongjie Liu, Diederik Paul Moeys, Gautham Das, Dan Neil, Shih-Chii Liu, and Tobi Delbruck. Combined frame- and event-based detection and tracking. In *International Symposium on Circuits and Systems*, May 2016.

[9] Gabriele Magrini, Federico Becattini, Pietro Pala, Alberto Del Bimbo, and Antonio Porta. Neuromorphic Drone Detection: an Event-RGB Multimodal Approach. 2024. Accepted at NeVi Workshop at ECCV 2024.

[10] Jakub Mandula, Jonas Kühne, Luca Pascarella, and Michele Magno. Towards real-time fast unmanned aerial vehicle detection using dynamic vision sensors. *IEEE International Instrumentation and Measurement Technology Conference*, 2024.

[11] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. *CoRR*, abs/1803.04523, 2018.

[12] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. How to Calibrate Your Event Camera. *CoRR*, abs/2105.12362, 2021. arXiv: 2105.12362.

[13] Carlos Ordonez, Edward R. Omiecinski, Shamkant B. Navathe, and Norberto F. Ezquerra. A clustering algorithm to discover low and high density hyper-rectangles in subspaces of multidimensional data. Technical report, College of Computing, Georgia Institute of Technology, Atlanta, GA, 1999.

[14] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7243–7252, 2019.

[15] Abhishek Tomy, Anshul Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, and Christian Laugier. Fusing Event-based and RGB camera for Robust Object Detection in Adverse Conditions. In *ICRA 2022 - IEEE International Conference on Robotics and Automation*, Philadelphia, United States, May 2022.

[16] Chunhui Zhao, Yakun Li, and Yang Lyu. Event-based Real-time Moving Object Detection Based On IMU Ego-motion Compensation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 690–696, 2023.

[17] Hanyu Zhou, Zhiwei Shi, Hao Dong, Shihan Peng, Yi Chang, and Luxin Yan. JSTR: Joint Spatio-Temporal Reasoning for Event-based Moving Object Detection. Accepted at NeVi Workshop, ECCV 2024.

[18] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4868–4880, August 2023.

[19] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. RGB-event fusion for moving object detection in autonomous driving, 2023. arXiv: 2209.08323 [cs.CV].

[20] Antoine Zundel, Cédric Demonceaux, Nicolas Hueber, Damien Spittler, Guillaume Strub, and Sébastien Changey. Bimodal vision system combining standard camera and dynamic vision sensor for detecting and tracking fast uavs. In *Proc. SPIE, Electro-Optical Remote Sensing*, 2024.