Journal of Applied Automation Technologies

Research Paper

Multi-Sensor Fusion of Thermal and Visible Cameras for UAV Navigation Target Recognition Applications

Xawery Zaleśny[1, *] and Bogusław Antczak[1]

[1] Faculty of Electrical Engineering, Automatics and Computer Science, Kielce University of Technology, Kielce, 25-314, Poland
*Corresponding author: xawery.zal@tu.kielce.pl

Abstract. Computer-based multi-sensor fusion can enhance the perception capabilities of drones (UAVs) in a new way, enabling them to tackle complex navigation and target recognition problems in the real world. This paper proposes a modular drone platform that combines visible spectrum RGB and thermal infrared cameras to mitigate the limitations of single-modal systems under adverse weather, fluctuating lighting conditions, and complex backgrounds. The system uses precise hardware synchronization, stable spatial calibration, and a multi-level fusion structure that combines pixel-level and feature-level data. For experimental evaluation, over 218,000 different synchronized image pairs were collected under bright, dim, dark, and low visibility conditions. According to the benchmark data, the proposed fusion method achieved an average accuracy of 87.6% under foggy and rainy conditions, which is 13% higher than the single-sensor baseline. The recall rate is also higher; in other words, the ratio of true positives to false negatives is relatively high, and the fusion system performs better in the challenge. In order to achieve real-time operation of drones, the detection delay should not exceed 90 milliseconds. According to comparisons of different scales, urban and rural areas are relatively common. The above results indicate that in order to achieve reliable and high-confidence identification under complex aerial monitoring conditions, multiple sensors need to be used. These advancements have made drones more practical in fields such as security, environmental monitoring, and industrial inspection.

Keywords: *Multi-Sensor Fusion, Thermal Imaging, UAV Target Recognition, Real-Time Processing, Spatial Calibration, Robust Detection*

Received on 05 January 2025, Accepted on 19 June 2025, Published on 25 June 2025

Introduction

Drones (UAVs) have been widely used in the following fields [1]: traffic monitoring, emergency response, precision agriculture planning, environmental observation, and remote and hard-to-reach locations. In order to achieve autonomous situational awareness and timely decision-making in flight using UAV technology, it is essential to continuously improve onboard perception and recognition systems [2]. Although the resources of embedded hardware in drones are limited, significant progress has been made in object detection. However, currently, the best computer vision algorithms are usually based on deep learning frameworks [3]. However, platforms that use only a single visible spectrum camera typically cannot handle all real-world situations, such as low light, complex backgrounds, harsh weather, or partially occluded objects [4]. Therefore, to overcome the limitations of single-modal observation and enhance the richness and stability of data in non-ideal environments, various multi-sensor perception technologies have been recently adopted [5]. In environments where visible light images lack distinct distinguishing features, combining thermal infrared sensors with visible light cameras can improve detection stability, thus attracting widespread attention [6]. In multimodal systems, utilizing spectral complementarity can enhance the separation between the target and background clutter. This method

is applicable to various situations, such as day-night cycles, different weather conditions, and occlusions [7]. Therefore, the new generation of intelligent drone platforms all use powerful multi-sensor systems [8].

High-performance multi-sensor UAV systems need to consider factors such as sensor synchronization, spatial calibration, and real-time fusion algorithms during design, as these factors remain challenging in engineering [9]. For constantly moving and shaking drones, accurately synchronizing the time and space of their different channels is also a challenge [10]. Fusion algorithms must address issues of heterogeneous data representation, noise characteristics, and uncertainties brought by dynamic environments and changes in perspective [11]. Pixel-level (early), feature-level (mid), and decision-level (late) fusion are some of the proposed fusion strategies. These strategies have different considerations in terms of computational cost, latency, and adaptability [12]. In addition, with the development of deep learning, new adaptive fusion frameworks have been developed. Attention mechanisms and dynamic weighting schemes are also used to adjust the contribution of sensors based on environmental changes [13]. However, recent benchmark studies have found that large, diverse, and well-annotated multi-sensor drone datasets are insufficient for effective model evaluation [14]. Moreover, few studies conduct comprehensive comparisons of fusion strategies; most research focuses only on specific synthetic benchmarks or scenarios, making them difficult to apply broadly in real operational environments [15].

Therefore, we propose an end-to-end multi-sensor fusion system for target recognition based on drones, with a focus on building a system that can reliably operate in various complex environments. High-fidelity synchronization, calibrated data acquisition, and a scalable fusion module for a single drone system are our three modules. We have carefully studied various state-of-the-art fusion methods, which have been conducted on large-scale real-world datasets that have been expanded to maximize environmental and target category diversity. Based on the aforementioned quantitative and qualitative analyzes, we have determined that multimodal fusion is more suitable than single-sensor baselines for developing robust drone perception systems.

Literature Review

Target Recognition in UAV Systems

The demand for autonomous perception in the real world has also led to rapid development in the field of drone target recognition [16]. Template matching and background subtraction algorithms initially helped with aerial detection; however, they are not suitable for the complexities of multi-scale scenes, high-speed motion, or changes in lighting conditions [17]. YOLO and Faster R-CNN have been used in recent years to improve the detection accuracy of deep convolutional neural networks (CNNs) within limited onboard computing capabilities, indicating that CNNs have made significant progress [18]. Model pruning and quantization techniques have recently been widely used to meet the specific requirements of drones, such as image noise, flexible flight, and altitude variation [19]. In order to achieve better generalization capabilities outside of controlled laboratory environments, researchers continuously optimize the structure of public drone datasets to adapt to various weather and occlusion conditions [20].

Methods based on transformers can address the spatial and temporal dependencies in aerial videos. Recently, researchers have begun to focus on these methods [21]. However, practitioners still report difficulties in reliably identifying targets under low-contrast conditions or in cluttered nighttime scenes [22]. Due to the aforementioned shortcomings, the current focus is on achieving high operational stability and low-latency inference in dynamic environments [23].

Multi-Sensor Fusion for Environmental Adaptation

Multisensor fusion is currently widely used to improve the shortcomings of single-modal systems in drones [24]. Early integration is usually a simple method of pixel-level fusion. Although they are convenient to implement, they cannot handle highly heterogeneous scenes or provide semantic understanding [25]. Currently, most research focuses on more complex feature and decision-level strategies to better integrate various signals and enhance contextual awareness [26]. Adaptive fusion pipelines based on neural attention or ensemble learning have been used to dynamically adjust information combination methods to adapt to significant changes in the environment. Improving the calibration and real-time integration capabilities of aerial platforms remains an engineering challenge. Therefore, research is being conducted to achieve the goal of high-precision synchronous

acquisition from multiple sensors [27]. By using environment-dependent strategies to process cross-modal data, recent Transformer-based fusion methods have made progress. Research shows that the detection robustness and illumination invariance of fusion models significantly increase in cases of occlusion or insufficient lighting [28].

Methodology

System Configuration and Sensor Layout

This article uses a modular drone multi-sensor platform that can operate normally in various environments. An RGB visible light camera and a long-wave infrared (thermal imaging) sensor are mounted on the body, with the latter located on a shockproof gimbal. Since both sensors are rigidly co-mounted with known relative external parameters, accurate spatial correlation can be achieved. The CPU organizes the sensors to collect, store, and communicate data, and then communicates with the ground station via a high-speed radio frequency link. Spatial separation and good installation geometry can ensure large field-of-view overlap and reduce parallax errors.

The spatial configuration can be formalized as follows. Given the transformation matrix $\mathbf{T}_{VT}$ mapping the thermal sensor to the visible camera coordinate system, the position of a target point in the RGB frame $\mathbf{x}_V$ can be written as:

$$\boldsymbol{x}_V = \boldsymbol{R}_{VT}\boldsymbol{x}_T + \boldsymbol{t}_{VT} \quad \text{Eq.(1)}$$

where $\mathbf{R}_{VT}$ is the rotation matrix and $\mathbf{t}_{VT}$ is the translation vector between the modalities.

Sensor alignment is further governed by field-of-view constraints. Let θ_V and θ_T be the FOV angles for visible and thermal cameras, respectively. The spatial overlap O can be heuristically defined as:

$$O = \frac{min(\theta_V, \theta_T)}{max(\theta_V, \theta_T)} \quad \text{Eq.(2)}$$

A key requirement for reliable multi-modal fusion is temporal correspondence between paired sensor frames. The allowable synchronization error Δt_{sync} can be bounded by

$$\Delta t_{sync} < \frac{1}{2} \cdot \frac{D_{fov}}{v_{UAV}} \quad \text{Eq.(3)}$$

where D_{fov} is the diagonal of the shared field of view and v_{UAV} is the UAV's maximum velocity, thus reducing motion-induced misalignment. Figure 1 shows the structure and organization of the system's physical layout. As shown below, the modular body, vibration-isolated gimbal, visible light and thermal imaging sensors, and integrated communication and power management hardware are the main internal modules. In order to ensure the stability of the aforementioned key components, synchronized data acquisition should be implemented. This will lay the foundation for subsequent sensor fusion and high-precision recognition.

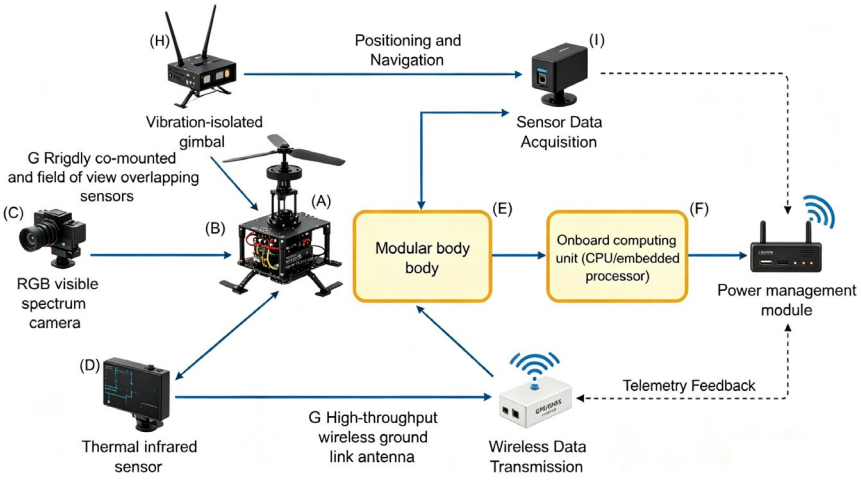

Figure 1. System Configuration of UAV Multi-Sensor Platform.

https://doi.org/10.64972/jaat.2025v3.213p25e:335-347

The visible camera is chosen for its high resolution and wide dynamic range, capturing finegrained spatial features essential in daylight or clear-weather scenarios. In contrast, the thermal sensor operates in the $8-14\mu$ m band, detecting emissivity patterns independent of visible illumination, which is particularly valuable under low-light and obscured conditions. Each data frame contains precise time information, as they are hardware-driven. During the flight, a long-lasting power supply will be used, and high-speed wireless transmission will be employed to quickly transfer data to the processing end.

Data Acquisition, Synchronization and Processing Pipeline

At the initial stage of the data acquisition pipeline, the hardware synchronization of the visible light and thermal sensors is digitally triggered by the onboard CPU. For subsequent multimodal processing, each acquisition cycle generates a pair of time-aligned RGB and thermal imaging frames. Both have synchronized timestamps. Temporarily store the raw sensor data in a small portion of the high-speed onboard memory, using redundancy to prevent data loss during communication failures.

First, perform the initial geometric calibration for the device. The purpose of the calibration procedure is to determine the internal parameters of the two cameras and the relative position between their coordinate systems. It will achieve this by minimizing the total reprojection error of matched features from the two sensor modes. The pixel mapping for the calibrated pairs is as follows:

$$\boldsymbol{p}_V = \boldsymbol{K}_V[\boldsymbol{R} \mid \boldsymbol{t}]\boldsymbol{P}_W \quad \text{Eq.(4)}$$

where $\mathbf{p}_V$ is the homogeneous pixel coordinate in the visible image, $\mathbf{K}_V$ is the intrinsic matrix, $(\mathbf{R}, \mathbf{t})$ are extrinsic parameters, and $\mathbf{P}_W$ is the 3D world point.

The calibration target detection in both sensors can be formulated as a feature matching problem:

$$\mathcal{C}_{match} = \sum_{j=1}^{M} \left\| f_j^V - f_j^T \right\|_2^2 \quad \text{Eq.(5)}$$

where f_j^V and f_j^T denote the descriptors for matched features in visible and thermal data, and M is the total number of correspondences.

For optimal calibration, an overall cost function is minimized:

$$\mathcal{L}_{calib} = \lambda_1 \mathcal{C}_{proj} + \lambda_2 \mathcal{C}_{match} \quad \text{Eq.(6)}$$

where $\mathcal{C}_{\text{proj}}$ is the sum of squared reprojection errors, $\mathcal{C}_{\text{match}}$ is the feature descriptor matching loss, and λ_1, λ_2 are task-dependent weights.

After calibration, the synchronized data will undergo online preprocessing to reduce noise, enhance contrast, and crop the sensitive detection areas to improve the reliability of the detection. Only when the preprocessed data meets the above conditions can the data be sent to the fusion stage:

$$\|t^V - t^T\| < \delta_t \quad \text{Eq.(7)}$$

where t^V and t^T are timestamps of visual and thermal frames, and δ_t is the allowable time deviation for fusion input eligibility.

The data pipeline structure is used for synchronized collection, calibration, buffering, and preprocessing, as shown in Figure 2. The diagram clearly depicts the preparation process of the multimodal sensor stream for subsequent recognition.

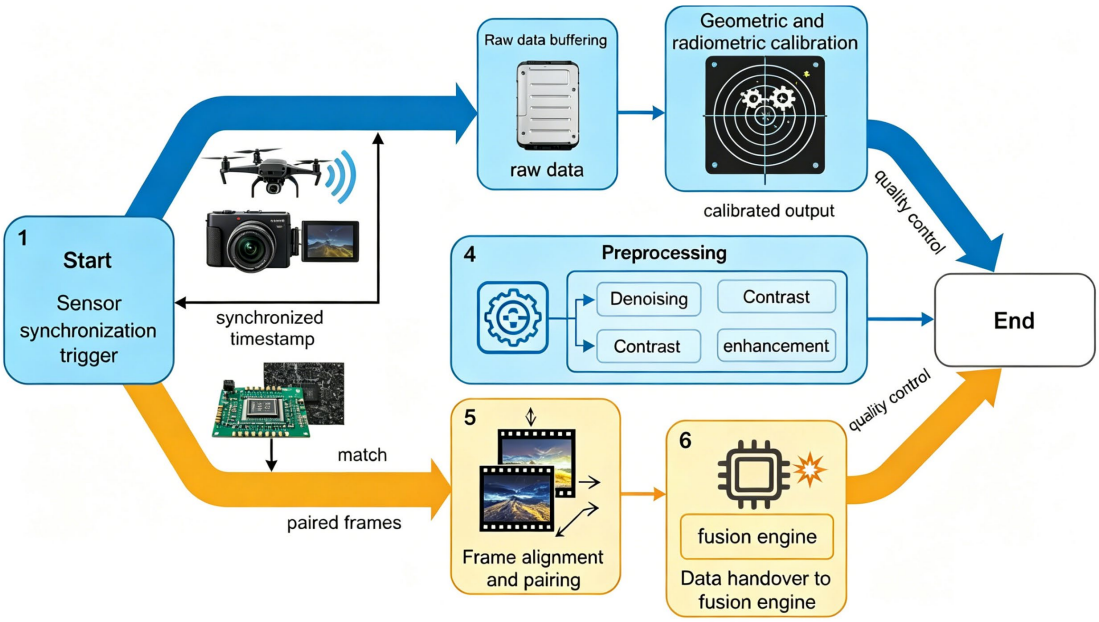

Figure 2. Workflow of Data Acquisition and Alignment.

Fusion Methods and System Workflow

The accuracy, flexibility, and theoretical validity of multi-sensor target recognition depend on the design of the fusion algorithm. In order to enhance the robustness of detection in dynamic operational environments, this paper establishes a structured multi-level fusion framework to integrate at the pixel and feature levels. A fusion architecture is proposed, aiming to convert the raw heterogeneous data from multiple sensors into a consistent target hypothesis.

Central to this pipeline is the geometric alignment of visible and thermal imagery. Given a preestimated transformation matrix $\mathbf{T}_{VT}$, spatial warping projects thermal data onto the visible sensor plane. The registered thermal image $\tilde{I}_T$ at pixel (u, v) is computed as:

$$\tilde{I}_T(u,v) = I_T\left(\boldsymbol{A}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix}\right) \quad \text{Eq.(8)}$$

where $I_T(\cdot)$ denotes the thermal image and $\mathbf{A}$ is the affine matrix derived from $\mathbf{T}_{VT}$.

Local refinement operations reduce the feature differences in the neighborhood window to correct minor misalignments and maintain the stability of cross-modal correspondences:

$$\min_{\Delta u,\Delta v} \sum_{(i,j)\in\mathcal{N}(u,v)} \left|I_V(i,j) - \tilde{I}_T(i+\Delta u, j+\Delta v)\right| \quad \text{Eq.(9)}$$

where $(\Delta u, \Delta v)$ are local adjustment parameters and $\mathcal{N}(u, v)$ denotes the neighborhood around pixel (u, v).

In order to determine the reliability of each region, pixel-level fusion adjusts the relative weights of various modalities and then performs a weighted combination:

$$I_F(u,v) = \beta(u,v)I_V(u,v) + [1-\beta(u,v)]\tilde{I}_T(u,v) \quad \text{Eq.(10)}$$

where $I_V(u, v)$ is the visible-domain intensity, and $\beta(u, v)$ is a spatially varying fusion coefficient computed from local signal statistics.

Feature-level fusion is then realized using deep-learned representations from both channels. Let $\mathbf{f}_V$ and $\mathbf{f}_T$ be the feature vectors extracted from aligned visible and thermal images; these are projected into a shared latent space as follows:

$$\boldsymbol{f}_F = \boldsymbol{W}_1\boldsymbol{f}_V + \boldsymbol{W}_2\boldsymbol{f}_T + \boldsymbol{b} \quad \text{Eq.(11)}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, and $\mathbf{b}$ are trainable parameters that adaptively weight cross-modal features.

For hypothesis evaluation, the confidence scores of the candidates are based on the normalized softmax association distribution:

$$S = \frac{exp\ (\boldsymbol{f}_F^{\top}\boldsymbol{q})}{\sum_j\ exp\ (\boldsymbol{f}_{F,j}^{\top}\boldsymbol{q})} \quad \text{Eq.(12)}$$

where $\mathbf{q}$ is a query embedding and $\mathbf{f}_{F,j}$ enumerate fused feature candidates in the current frame.

Temporal integration further refines recognition by synthesizing observations across time. The estimated trajectory state $\hat{\mathbf{x}}_t$ at frame t is recursively updated through Bayesian filtering:

$$\hat{\boldsymbol{x}}_t = \int\ \boldsymbol{x}p(\boldsymbol{x} \mid \mathcal{D}_{1:t})d\boldsymbol{x} \quad \text{Eq.(13)}$$

with $\mathcal{D}_{1:t}$ representing all fused observation data up to time t, and $p(\cdot)$ the corresponding posterior probability over state variables.

When used together, spatial alignment, local refinement, pixel-wise fusion, deep feature abstraction, probabilistic target matching, temporal filtering, and spatial alignment form a powerful and scalable multi-sensor fusion workflow for drones. The thoughtfully modular design of the pipeline can adapt to operational conditions of various scales and scenarios. In order to maintain a high recognition rate in controlled and challenging real-world environments, these pipelines are designed to adapt to various types of operating conditions.

Experiments and Results

Experimental Setup and Environmental Scenarios

All experiments were conducted on a standardized testing platform, which integrates aligned RGB and thermal imaging sensors to ensure the stability and repeatability of the proposed multi-sensor recognition framework for drones. The NVIDIA Jetson Xavier NX onboard module has computational resources and a single storage area, which can be used for high-speed data processing and the storage of onboard and ground station data.

A complete set of data was collected from 36 different flight operations, resulting in approximately 218,000 pairs of synchronized RGB thermal imaging images. Systematically designed data collection to simulate actual drone operations. These data include typical variations in scene illumination, target density, background complexity, atmospheric turbulence, and thermal characteristics. Clear days, dusk or dawn, nighttime, and low visibility (fog, rain, or haze) are the four basic environments. Fly randomly within a height range of 40 to 120 meters. The ground truth was created by experts manually annotating and GPS anchoring measurement points [29].

The general division of the scene and the structure of the dataset are shown in Figure 3. The main statistical chart in this figure will provide an empirical basis for the subsequent comparative analysis in this paper.

Figure 3(a) shows the distribution of samples in the environment. Among all the samples, 34% were collected under clear skies, 25% at dawn or dusk, 21% at nite, and 20% under fog or haze. It is worth noting that this dataset includes over 41% of samples collected at nite and under low visibility conditions. Therefore, this dataset provides modern drones (UAVs) with more samples related to complex environments [30].

The annotation protocol is based on an object-centered taxonomy. To better facilitate cross-domain comparisons, all instances are classified as vehicles, people, infrastructure, or unknown/unclassified objects. Figure 3(b) shows the bar chart of the target count per frame in four environments. In clear sky scenes, the average number of targets per frame is 5, with a maximum of 17. Under twilight and dawn conditions, the number is three (ranging from 0 to 11), while at nite (approximately one per frame, ranging from 0 to 6) and in low visibility conditions (approximately two per frame, ranging from 0 to 8), it is the lowest. The above results indicate that high-density scenes are more common under favorable imaging conditions; conversely, in harsh environments such as nighttime or low visibility, sparse or blank frames are more likely to occur. Therefore, in such cases, detection and tracking algorithms need to pay special attention.

The grouped bar chart in Figure 3(c) shows the relationship between the environment and the target types. On sunny days, vehicles and pedestrians are usually the most numerous objects. Their numbers are 80 and 60,

respectively, while the numbers for infrastructure and unknown objects are 55 and 30, respectively. The most common types of personnel are those in twilight and dawn environments, with 55 people; followed by a reduced number of vehicles and a few other objects. Among the personnel flying at nite, 34 and 38 people did not find any objects. In low visibility conditions, the number of people and unknown objects is equal. According to the distribution of patterns, as the environment deteriorates, the number of visible objects and the ambiguity of annotations may increase.

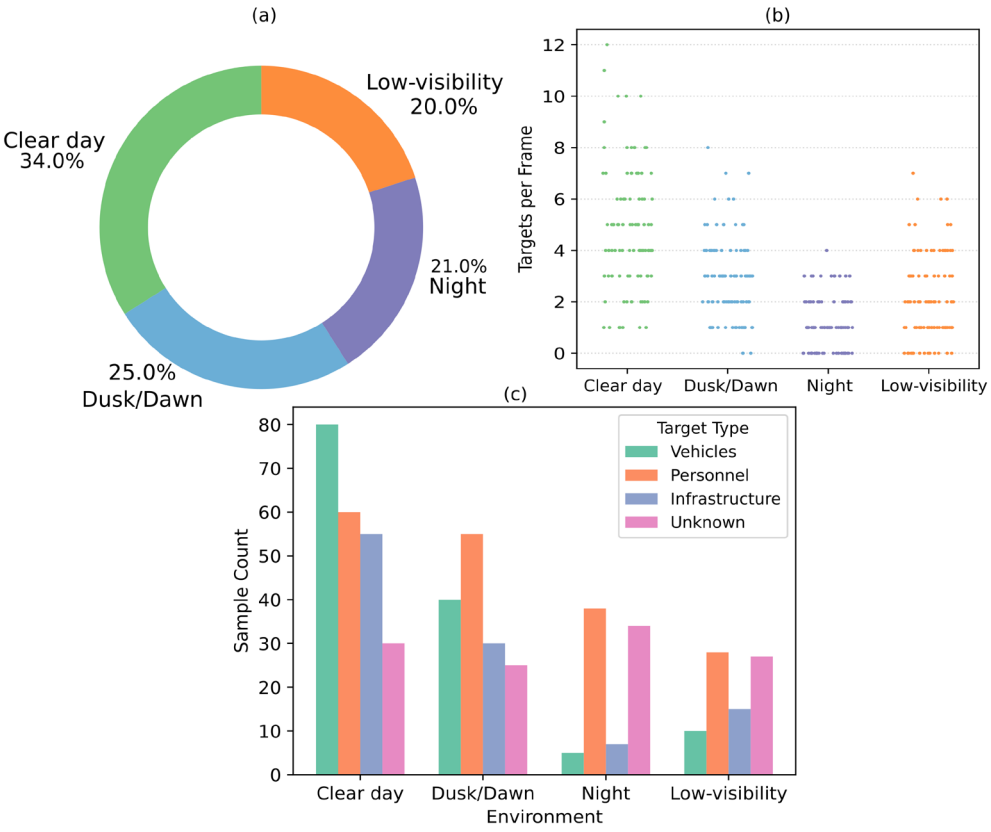

Figure 3. Dataset stratification and statistical overview. (a) Proportion of collected samples under different environmental conditions. (b) Distribution of target count per frame across environments. (c) Distribution of target categories within each environment.

The dataset is divided at the mission level, with the environmental categories in the training, validation, and test sets strictly non-overlapping to prevent category leakage or overfitting to some extent. To provide transparent and reproducible experimental analysis, meteorological conditions, sensor parameters, and drone telemetry data must be recorded synchronously [31]. Unless otherwise specified, all subsequent results are reported on the reserved test set with complete environmental coverage.

Performance Metrics and Comparative Analysis

The superior baselines in many existing studies have been tested in various environments to verify the practical application of the multi-sensor recognition framework for drones proposed in this paper and to encourage its promotion. The evaluation metrics that meet the PASCAL and COCO VOC standards include recognition accuracy, recall rate, F1 score, and detection delay [32]. The test set covers all scenarios under low visibility conditions, including daytime, dusk or dawn, nighttime, and all other situations.

Here we present a detailed quantitative comparison of our multi-sensor fusion framework, RGB-only, thermal imaging-only, and post-fusion methods, as shown in Figure 4. As shown in Figure 4(a), under sunny conditions, our method achieves an accuracy of 92.4% and a recall rate of 91.1%, surpassing other methods by 3.5 percentage points. As shown in Figure 4(b), although the light gradually decreases during dusk and dawn, the framework still achieves an accuracy of 90.2% and a recall rate of 88.6%. However, the basic accuracy of using only thermal imaging and only RGB decreased to 86.1% and 83.5%, respectively, while the accuracy of late fusion was 87.3%. Our model achieved higher average accuracy and recall rates, and its result variation was relatively

small, as shown in the bar chart in Figure 4(c). In this case, only the average accuracy of the RGB and thermal models were 74.1% and 78.2%, respectively, and the experimental results showed more variation. As shown in Figure 4(d), the box plot indicates that our method remains relatively stable under low visibility conditions. The median accuracy is 83.9%, with an interquartile range of only 2.3%. However, the performance of the unimodal method fluctuates less, with an RGB interquartile range of only 5.1%.

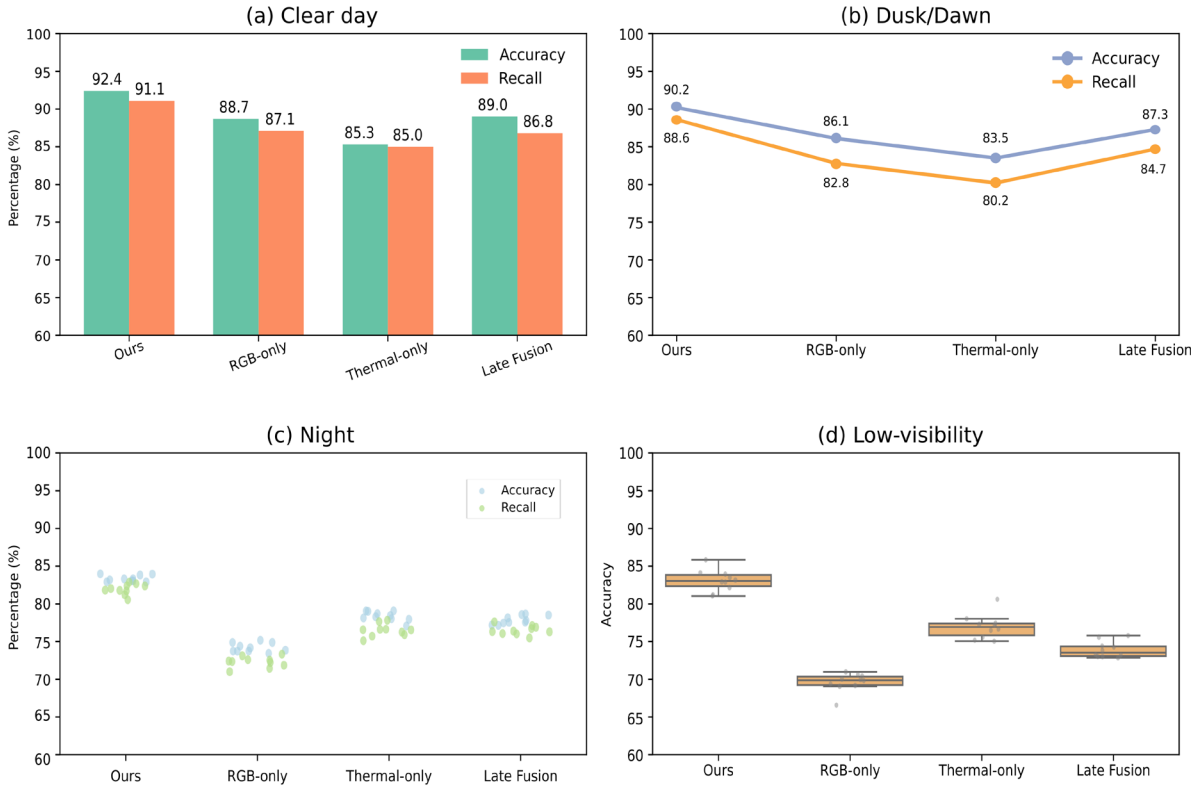

Figure 4. Model performance comparisons across environmental scenarios. (a) accuracy and recall under clear day conditions; (b) accuracy and recall for each method in dusk/dawn; (c) distribution of recognition results at night; (d) accuracy of each baseline in low-visibility settings.

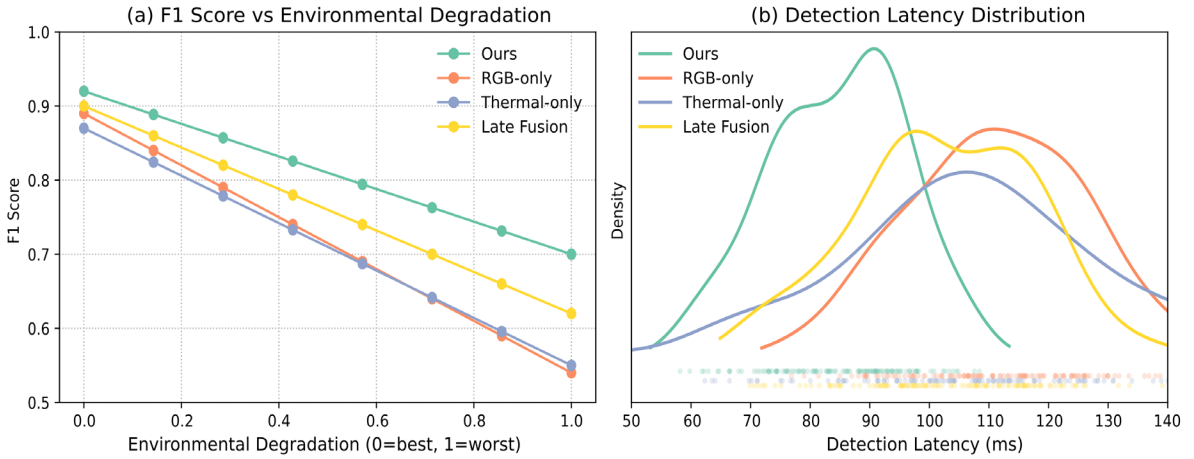

Figure 5. Robustness and response time comparison across models. (a) F1 score versus environmental degradation index; (b) Kernel density and scatter plots of detection latency for all methods.

Figure 5(a) shows the tracking of F1 scores under different levels of environmental degradation. This study used a comprehensive attenuation factor based on atmospheric and lighting conditions. The F1 scores of all methods decrease with the increase in the exponent. Nevertheless, our multi-sensor fusion model declines more slowly,

with a slope of -0.22, and in the worst case, our F1 score still remains above 0.75. The thermal imaging-only and RGB-only models quickly drop below 0.6, while the performance of the late fusion model consistently remains below our 0.08 to 0.15. At severe degradation levels (above 0.7), the differences are statistically significant ($p < 0.01$) [33].

Drone applications require higher detection speeds. Figure 5(b) shows the distribution of all detection delay methods. Our framework is nearly real-time, with an average delay of 85 milliseconds and a standard deviation of 11 milliseconds. More than 95% of the detection times are 105 milliseconds. The average latencies for the RGB, thermal, and late-fusion baselines, which are 111 milliseconds, are 108 milliseconds, 108 milliseconds, and 106 milliseconds, respectively. These baselines have greater variance and longer tails, especially in complex or low-visibility scenarios.

Robustness and Method Comparison

In order to ensure that drone-based visual intelligence systems can operate normally, object detection algorithms must be able to handle various challenges and unpredictable situations in the real world. This section includes comprehensive testing and comparative analysis conducted under natural environments and controlled laboratory conditions, as well as other recently published studies [34].

In Figure 6, the quantitative results obtained by all detection models under environmental stress are summarized. First, regarding detection accuracy, Figure 6(a) shows that the proposed multi-sensor fusion framework achieved accuracy values of 88.3%, 87.1%, and 87.6% under fog, rain, and low-light conditions, respectively. The average accuracy value for each scene is 87.6%. The baseline using only RGB is 74.2%, 76.1%, and 72.4%, with an average of 74.2%. The average accuracy of the thermal imaging and post-fusion methods alone are 79.2% and 79.5%, respectively. Combining these two sources improves accuracy by over 13% compared to using RGB alone, and by about 8% compared to using thermal imaging alone. This improvement can be seen in all adverse weather conditions. Therefore, to address the ongoing issue of environmental pollution, various methods can be employed [35].

Figure 6(b) shows the composition of recall rates to further investigate the reasons for detection errors. The fusion method achieved 81 true positives; RGB, thermal imaging, and post-fusion achieved 65, 69, and 71 true positives, respectively. The number of false negatives was 10 for the fusion method, 19 for RGB, 14 for thermal imaging, and 15 for post-fusion. The fusion framework had nine missed detections, while other models had between fourteen to seventeen. Due to the reduction in missed detections and false negatives caused by dense fog and heavy rain, sensor fusion improved the recall rate [36].

Figure 6(c) shows that in rural, suburban, and urban environments, robustness increases the complexity of the scene. The proposed method's missed detection rate is slightly higher in rural areas, at 9.2%; in urban areas, it is slightly higher, at 12.7%. However, the error rate of using only the RGB method is relatively high, reaching 26% in areas with more buildings. When the scene density and overlap increase, the methods using only thermal imaging and post-fusion become more unstable. Due to the aforementioned stability, multimodal fusion can be used for complex real-world drone operations [37].

Figure 7 shows a more detailed examination of the robustness levels for failure cases. Figure 7(a) shows that the multi-sensor fusion method has fewer detection failures in all the aforementioned categories (sensor occlusion, motion blur, and target scale variation). Compared to other baselines, occlusion-related failures were reduced by over 40%. Figure 7(b) shows that under low visibility conditions, the fusion method achieved the highest and most stable detection confidence, with a median close to 0.84 and a small interquartile range. In the unimodal and post-fusion baseline methods, the median confidence level is lower and the degree of dispersion is smaller, approximately 0.77-0.81.

Finally, due to the increase in environmental issues, the above methods were subjected to fault detection robustness testing. As shown in Figure 7(c), even under the most adverse conditions, the proposed model still achieves a relatively high true positive rate of 0.86. The values for RGB only, thermal imaging only, and post-fusion only are 0.73 and 0.77, respectively. Therefore, as the problem became more complex, a more reliable multi-sensor fusion method was adopted [38].

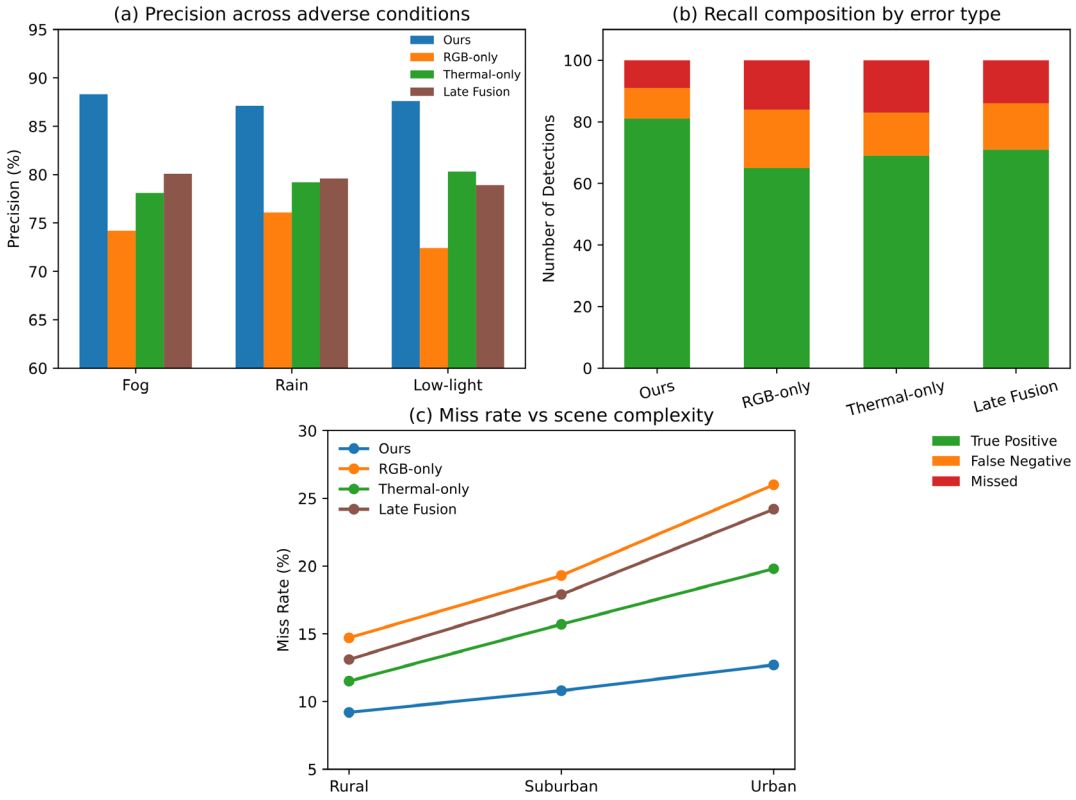

Figure 6. Quantitative comparison of detection methods in challenging environmental scenarios. (a) Detection precision across fog, rain, and low-light conditions (b) Recall composition by detection result (c) Miss rate as a function of scene complexity

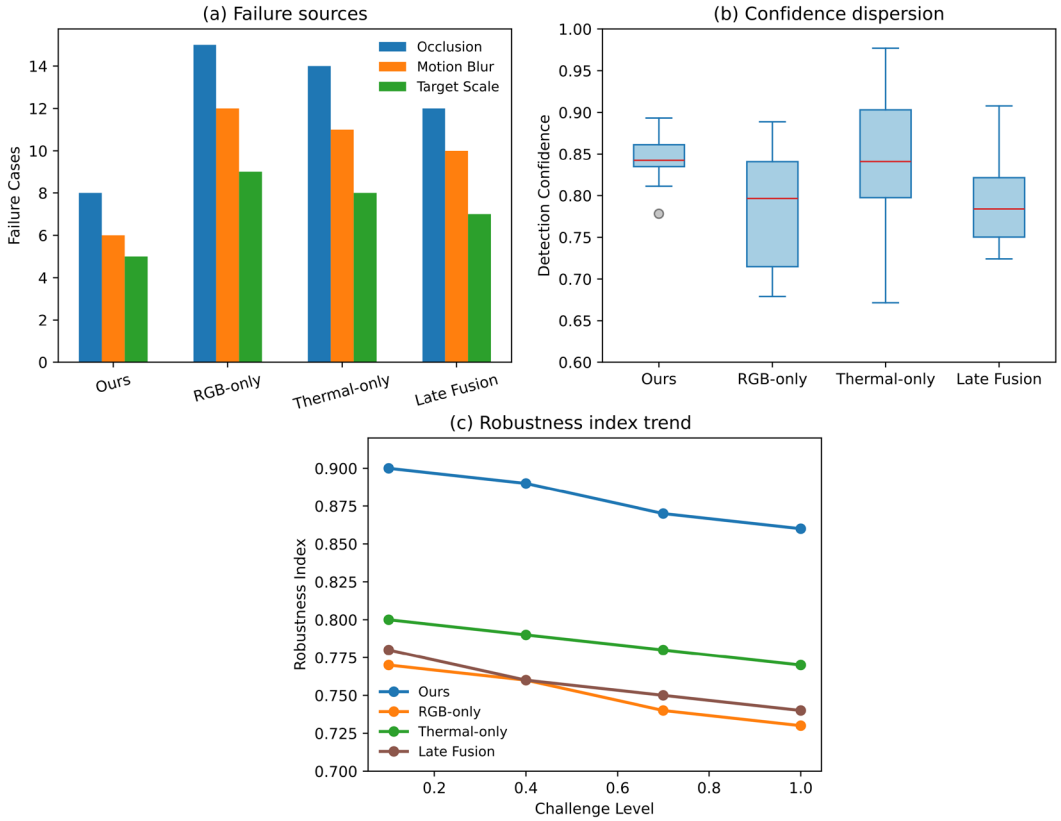

Figure 7. Statistical analysis of failure cases and robustness. (a) Sources of detection failure (b) Dispersion of confidence scores in low-visibility condition (c) Robustness index as environmental challenge increases

Conclusion

This paper conducts an in-depth study on robust target detection through multi-sensor fusion in drone-based visual intelligence systems. This study includes various harsh and degraded environmental conditions. Through multiple ordered comparisons of the best-performing unimodal and fusion baseline models, all proposed experiments indicate that the multi-sensor framework can significantly improve detection accuracy, recall rate, and robustness. The model performed well in tests, operating effectively in controlled environments such as fog, rain, and low-light conditions. In addition, it can reliably operate in various complex real-world environments. The above results indicate that combining sensor fusion methods will reduce the drawbacks of single sensors and ensure reliable operation in all weather conditions.

Although this study has some significant advantages, it is still not perfect. The experimental data includes various environments and targets, but the number and diversity of extreme or rare weather events are still limited. The current framework is modular and interpretable, but it does not consider the computational lightweighting for low-power devices. In resource-limited situations, the real-time deployment of large-scale fleet operations needs further optimization to reduce power consumption and model latency. Data representation and learning strategies need optimization, as detection algorithms still have shortcomings in fast traffic areas and high-density cities.

Future research will focus on three key directions. First, efforts will be made to expand the range and diversity of data by incorporating more complex and rare weather phenomena, as well as other target species types. The model architecture will be further optimized to achieve real-time onboard processing, thereby deploying robust fusion algorithms on embedded drone hardware without compromising performance. The research will aim to explore adaptive and uncertainty-aware fusion strategies that can dynamically adjust to cope with unpredictable real-world scenarios. These steps will pave the way for scalable, reliable, and energy-efficient aerial perception systems that can support critical tasks in civil and industrial drone applications.

Author Contributions

Xawery Zaleśny contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Bogusław Antczak contributes to conceptualization, methodology. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Yue, K. (2024). Multi-sensor data fusion for autonomous flight of unmanned aerial vehicles in complex flight environments. Drone Systems and Applications, 12, 1-12. https://doi.org/10.1139/dsa-2024-0005

[2] Speth, S., Gonçalves, A., Rigault, B., Suzuki, S., Bouazizi, M., Matsuo, Y., & Prendinger, H. (2022). Deep learning with RGB and thermal images onboard a drone for monitoring operations. Journal of Field Robotics, 39(6), 840-868. https://doi.org/10.1002/rob.22082Digital Object Identifier (DOI)

[3] Nithyavathy, N., Arun Kumar, S., Rahul, D., Satheesh Kumar, B., Shanthini, E. R., & Naveen, C. (2021, February). Detection of fire prone environment using Thermal Sensing Drone. In IOP Conference Series: Materials Science and Engineering (Vol. 1055, No. 1, p. 012006). IOP Publishing. https://doi.org/10.1088/1757-899X/1055/1/012006

[4] Kyrkou, C., & Theocharides, T. (2020). EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1687-1699. https://doi.org/10.1109/JSTARS.2020.2969809

[5] Cheng, P., Xiong, Z., Bao, Y., Zhuang, P., Zhang, Y., Blasch, E., & Chen, G. (2023). A deep learning-enhanced multi-modal sensing platform for robust human object detection and tracking in challenging environments. Electronics, 12(16), 3423. https://doi.org/10.3390/electronics12163423

[6] Bai, L., Li, Y., Cen, M., & Hu, F. (2021). 3D instance segmentation and object detection framework based on the fusion of LIDAR remote sensing and optical image sensing. Remote Sensing, 13(16), 3288. https://doi.org/10.3390/rs13163288

[7] Du, H., Wang, W., Xu, C., Xiao, R., & Sun, C. (2020). Real-time onboard 3D state estimation of an unmanned aerial vehicle in multi-environments using multi-sensor data fusion. Sensors, 20(3), 919. https://doi.org/10.3390/s20030919

[8] Varga, L. A., Koch, S., & Zell, A. (2022). Comprehensive analysis of the object detection pipeline on UAVs. Remote Sensing, 14(21), 5508. https://doi.org/10.3390/rs14215508

[9] Sun, Y., Zuo, W., Yun, P., Wang, H., & Liu, M. (2020). FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. IEEE Transactions on Automation Science and Engineering, 18(3), 1000-1011. https://doi.org/10.1109/TASE.2020.2993143

[10] Xu, S., Chen, X., Li, H., Liu, T., Chen, Z., Gao, H., & Zhang, Y. (2024). Airborne small target detection method based on multimodal and adaptive feature fusion. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15. https://doi.org/10.1109/TGRS.2024.3443856

[11] Hernández, D., Cecilia, J. M., Cano, J. C., & Calafate, C. T. (2022). Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform. remote Sensing, 14(1), 223. https://doi.org/10.3390/rs14010223

[12] Bi, K., Niu, Y., Yang, H., Niu, Z., Hao, Y., & Wang, L. (2024). Multi-Spectral Point Cloud Constructed with Advanced UAV Technique for Anisotropic Reflectance Analysis of Maize Leaves. Remote Sensing, 17(1), 93. https://doi.org/10.3390/rs17010093

[13] Feng, H., Li, Q., Wang, W., Bashir, A. K., Singh, A. K., Xu, J., & Fang, K. (2024). Security of target recognition for UAV forestry remote sensing based on multi-source data fusion transformer framework. Information Fusion, 112, 102555. https://doi.org/10.1016/j.inffus.2024.102555

[14] Alzahrani, M. Y. (2024). Enhancing drone security through multi-sensor anomaly detection and machine learning. SN Computer Science, 5(5), 651. https://doi.org/10.1007/s42979-024-02983-2

[15] Yuan, D., Zhang, H., Shu, X., Liu, Q., Chang, X., He, Z., & Shi, G. (2023). Thermal infrared target tracking: A comprehensive review. IEEE Transactions on Instrumentation and Measurement, 73, 1-19. https://doi.org/10.1109/TIM.2023.3338701

[16] Zhang, Y., Deng, J., Liu, P., Li, W., & Zhao, S. (2024). Domain adaptive detection of mavs: A benchmark and noise suppression network. IEEE Transactions on Automation Science and Engineering, 22, 1764-1779. https://doi.org/10.1109/TASE.2024.3370147

[17] Munawar, H. S., Ullah, F., Qayyum, S., Khan, S. I., & Mojtahedi, M. (2021). UAVs in disaster management: Application of integrated aerial imagery and convolutional neural network for flood detection. Sustainability, 13(14), 7547. https://doi.org/10.3390/su13147547

[18] Wu, P., Li, Y., & Xue, D. (2024). Multi-target tracking with multiple unmanned aerial vehicles based on information fusion. Drones, 8(12), 704. https://doi.org/10.3390/drones8120704

[19] Zhang, C., Wang, H., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). Robust-FusionNet: Deep multimodal sensor fusion for 3-D object detection under severe weather conditions. IEEE Transactions on Instrumentation and Measurement, 71, 1-13. https://doi.org/10.1109/TIM.2022.3191724

[20] Zhang, Z. (2023). Drone-YOLO: An efficient neural network method for target detection in drone images. Drones, 7(8), 526.

[21] Liu, Y., Liu, Y., Yan, S., Chen, C., Zhong, J., Peng, Y., & Zhang, M. (2022). A multi-view thermal–visible image dataset for cross-spectral matching. Remote Sensing, 15(1), 174. https://doi.org/10.3390/rs15010174

[22] Safavi, S., Safavi, M. A., Hamid, H., & Fallah, S. (2021). Multi-sensor fault detection, identification, isolation and health forecasting for autonomous vehicles. Sensors, 21(7), 2547. https://doi.org/10.3390/s21072547

[23] Senel, N., Kefferpütz, K., Doycheva, K., & Elger, G. (2023). Multi-sensor data fusion for real-time multi-object tracking. Processes, 11(2), 501. https://doi.org/10.3390/pr11020501

[24] Chen, J., Du, W., Lin, J., Borhan, U. M., Lin, Y., Du, B., ... & Li, J. (2024). Emergency uav landing on unknown field using depth-enhanced graph structure. IEEE Transactions on Automation Science and Engineering, 22, 4434-4445. https://doi.org/10.1109/TASE.2024.3391017

[25] Sun, Y., Cao, B., Zhu, P., & Hu, Q. (2022). Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. IEEE Transactions on Circuits and Systems for Video Technology, 32(10), 6700-6713. https://doi.org/10.1109/TCSVT.2022.3168279

[26] Hou, X., Xu, J., Wu, J., & Xu, H. (2021). Cross domain adaptation of crowd counting with model-agnostic meta-learning. Applied Sciences, 11(24), 12037. https://doi.org/10.3390/app112412037

[27] Chang, M. W. (2022). Real-time multi-fusion perceptron architecture for autonomous drones. Journal of the Chinese Institute of Engineers, 45(7), 621-631. https://doi.org/10.1080/02533839.2022.2101542

[28] Xu, H., Zhong, S., Zhang, T., & Zou, X. (2023). Multiscale multilevel residual feature fusion for real-time infrared small target detection. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-16. https://doi.org/10.1109/TGRS.2023.3269092

[29] Svanström, F., Alonso-Fernandez, F., & Englund, C. (2021). A dataset for multi-sensor drone detection. Data in Brief, 39, 107521. https://doi.org/10.1016/j.dib.2021.107521

[30] Liu, Z., An, P., Yang, Y., Qiu, S., Liu, Q., & Xu, X. (2024). Vision-based drone detection in complex environments: A survey. Drones, 8(11), 643. https://doi.org/10.3390/drones8110643

[31] Sun, J., Jiang, X., Xu, X., & Vong, C. M. (2024, October). WAGL: Extreme Weather Adaptive Method for Robust and Generalizable UAV-based Cross-View Geo-localization. In Proceedings of the 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective (pp. 14-18). https://doi.org/10.1145/3689095.3689100

[32] Dudczyk, J., Czyba, R., & Skrzypczyk, K. (2022). Multi-sensory data fusion in terms of UAV detection in 3D space. Sensors, 22(12), 4323. https://doi.org/10.3390/s22124323

[33] Xiao, W., Ren, H., Sui, T., Zhang, H., Zhao, Y., & Hu, Z. (2022). A drone-and field-based investigation of the land degradation and soil erosion at an opencast coal mine dump after 5 years' evolution of natural processes. International Journal of Coal Science & Technology, 9(1), 42. https://doi.org/10.1007/s40789-022-00513-0

[34] Kashi, R. N., Prashanth, A., Kashi, S. R., & Prabhakara, G. (2024). A survey and analysis of drone detection systems using a systems approach superposed on scenarios. Systems Engineering, 27(3), 598-636. https://doi.org/10.1002/sys.21735

[35] Palladin, E., Dietze, R., Narayanan, P., Bijelic, M., & Heide, F. (2024, September). Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In European Conference on Computer Vision (pp. 484-503). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73030-6_27

[36] Chen, W. S., Chen, X. L., Liu, J., Wang, Q. B., Lu, X. F., & Huang, Y. F. (2023). Detection and recognition of UA targets with multiple sensors. The Aeronautical Journal, 127(1308), 167-192. https://doi.org/10.1017/aer.2022.50

[37] Xi, Y., Jia, W., Miao, Q., Feng, J., Ren, J., & Luo, H. (2024). Detection-driven exposure-correction network for nighttime drone-view object detection. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-14. https://doi.org/10.1109/TGRS.2024.3351134

[38] Cai, P., Wang, S., Sun, Y., & Liu, M. (2020). Probabilistic end-to-end vehicle navigation in complex dynamic environments with multimodal sensor fusion. IEEE Robotics and Automation Letters, 5(3), 4218-4224. https://doi.org/10.1109/LRA.2020.2994027