IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 July 2026, accepted 8 September 2026, date of publication 21 September 2026, date of current version 2 October 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3735919

RESEARCH ARTICLE

Early Versus Late RGB–Infrared Fusion for Counter-UAV Detection: A Comparative Study

ALBERTO DE ZAN, DENIS TAVARIS, GIAN LUCA FORESTI, (Senior Member, IEEE), AND IVAN SCAGNETTO

Department of Mathematics, Computer Science, and Physics, University of Udine, 33100 Udine, Italy

Corresponding author: Ivan Scagnetto (ivan.scagnetto@uniud.it)

ABSTRACT This paper presents a novel strategy for UAV detection through visible (RGB) and infrared (IR) fusion, in order to improve accuracy and precision in harsh environments and weather conditions. In this research, two opposite approaches have been studied and implemented: the first concerns image-level (early) fusion, which merges two different spectra into a single image to which one detector is applied. On the other hand, the second approach, called decision-level (late) fusion, focuses on blending together the output bounding boxes of two different detection pipelines, aiming to create a common result that can be applied to both spectra simultaneously. Both strategies have been extensively tested in a dedicated environment, which included alternating low and high brightness scenes, different backgrounds (sky and ground), distances (5, 25 and 60 m), altitudes and zoom levels (1X, 2X and 4X). These tests allowed for the assessment of actual detection capabilities and system reliability in real-world applications. Furthermore, the two fusion strategies were evaluated and compared over a set of aligned images in order to highlight the better approach of the two. The results indicate that early fusion achieves better raw performance, since it incorporates the thermal information from the beginning of the training process. However, late fusion presents an innovative and useful detection workflow that can be applied in cases of sensor or frame misalignment. Both positive and negative aspects have also been pointed out and discussed, in order to give a more comprehensive overview. Finally, the paper outlines implementation considerations and practical recommendations based on operational priorities. Overall, the findings contribute to the research in RGB–IR fusion modalities, with the goal of enhancing UAV detection capabilities.

INDEX TERMS Uncrewed aerial vehicles (UAVs), object detection, YOLO, early fusion, late fusion, multimodality, infrared imaging.

I. INTRODUCTION

Uncrewed Aerial Vehicles (UAVs) are becoming increasingly common in military and civilian applications. They can be used in surveillance, mapping and transporting materials. However, there are safety and security concerns which have to be taken into consideration, since modern UAVs might be able to operate in varied environments and evade traditional surveillance methods. In order to reduce these weaknesses, many advanced sensors are used in current scenarios. For example, visible light (RGB) cameras are widely adopted, as they offer excellent spatial resolutions and reduced costs [1]. However, they have problems in low light, adverse weather conditions [2] and crowded backgrounds. Concerning this particular aspect, the terrain texture may be difficult to separate from the drone's surface, since UAVs are usually made from composite and camouflaged materials. On the other hand, infrared (IR) cameras work effectively in low light situations and night scenes. Nevertheless, they have reduced thermal and spatial resolution and are expensive. Therefore, this research focuses on fused solutions for UAV detection, aiming to combine the strengths of both sensors [3].

There are some studies in Computer Vision which have explored the capabilities of sensor fusion strategies in object

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

detection systems [4]. However, the Counter-UAV setting still remains an underexplored research area for fusion. Most of the existing approaches require significant adjustments and fixes to the sensors before fusion can be applied. In this paper, the following innovative fusion pipelines for YOLO-based Counter-UAV detection have been adopted and evaluated:

- An image-level (early) fusion that composes synchronized and aligned RGB and IR inputs into a unified representation for a single detector.
- A decision-level (late) fusion that processes RGB and IR streams independently and merges their detections post-inference.

This research presents two logical strategies and the framework designed and applied for each fusion type.

Finally, in order to present a complete evaluation, extensive performance analyses have been conducted, including quantitative metrics (such as average detection precision, recall and F1-score) and qualitative visualizations through image samples.

Synopsis: This work presents two RGB–IR fusion architectures for drone-to-drone detection using YOLO: an image-level (early) fusion combining synchronized inputs into a single stream, and a decision-level (late) fusion merging outputs from separate detectors. Evaluated across varied conditions, the study compares accuracy, efficiency, and robustness, offering practical guidance for Counter-UAV deployment.

II. RELATED WORK

A. DETECTION TECHNIQUES

The detection of small uncrewed aerial vehicles (UAVs) has attracted increasing attention due to their rapid proliferation and the associated security implications [5]. Early vision-based UAV detection methods relied primarily on classical image processing and handcrafted feature extraction techniques. Background subtraction, motion-based segmentation and orientation detection were commonly employed in static-camera surveillance scenarios, leveraging temporal consistency to isolate moving aerial targets [6]. Feature-based approaches utilizing Histogram of Oriented Gradients (HOG) [7], Haar-like features and Scale-Invariant Feature Transform (SIFT) [8] descriptors were also explored for detecting aerial objects against sky backgrounds. While computationally efficient, such methods often suffer from limited robustness under dynamic backgrounds, illumination changes and cluttered scenes.

The emergence of deep learning has significantly advanced object detection performance [9]. Two-stage detectors, such as Faster R-CNN [10], achieve high detection accuracy through region proposal mechanisms followed by classification and regression refinement. In addition, they offer strong capabilities in feature extraction tasks. However, their computational complexity can limit real-time applicability in resource-constrained operational environments. In contrast, one-stage detectors, including Single Shot MultiBox Detector (SSD) [11] and the You Only Look Once (YOLO) [12] family, directly predict bounding boxes and class probabilities in a unified framework, enabling significantly higher inference speeds [13].

Moreover, models based on transformers, such as Vision Transformers (ViT) [14], [15], DEtection TRansformer (DETR) [16] and Spatio-Temporal Transformer [17], have increased overall robustness by capturing long-range dependencies within images. Furthermore, advanced multimodal deep learning models with hybrid CNN-transformer architectures enhance both feature extraction and decision making through data obtained from numerous sensors. Additionally, the integration of multi-sensor systems into embedded platforms has substantially increased their value in practical applications, including drone surveillance and search-and-rescue operations, demonstrating the effectiveness of these multi-sensor approaches.

Classification-and-regression-based detectors, such as YOLO, have become particularly attractive for UAV detection due to their favorable trade-off between accuracy and latency and wide compatibility with many devices. Several studies have adapted YOLO architectures for small-object detection, aerial surveillance and drone monitoring tasks [18]. Modifications such as multi-scale feature aggregation and anchor box optimization have been proposed to improve sensitivity to small aerial targets. Nevertheless, most existing UAV detection approaches rely solely on visible-spectrum imagery, which inherently limits performance under low-light, night-time or adverse weather conditions [19]. These limitations motivate the exploration of complementary sensing modalities to enhance robustness in real-world Counter-UAV applications, potentially involving other drones in the detection process [20].

B. FUSION TECHNIQUES

Overall, RGB and infrared modalities exhibit complementary characteristics. Therefore, their integration represents a promising direction for enhancing UAV detection performance in operational Counter-UAV scenarios [21], [22]. Methods such as deep fusion networks and attention mechanisms have shown significant improvements in accuracy, particularly in scenarios such as low-visibility conditions and occlusions [23], but also in challenging cases such as the resemblance between drones and birds. Furthermore, several studies have shown interest in RGB and IR image fusion methods, with the goal of creating highly informative multi-modal images that integrate features from both spectra. These techniques employ two-scale image decomposition [24], contrast enhancement [25] and swin transformer [26]. However, they usually need aligned image pairs for achieving fusion, which are not always available due to the sensors' physical limitations. For this reason, some fusion methods have been developed to also integrate image registration and alignment [27], [28]. However, recent works have explored approaches for fusing spatially misaligned yet temporally

synchronized images [29]. Furthermore, multi-modal object detection has already been explored by Cao et al. [30] and Xu et al. [31]. Some studies also involve Bayesian fusion for enhancing perception [32]. Modern approaches also integrate additional data sources such as LiDAR and radar, in order to enhance depth perception and robustness, although these methods are mainly oriented toward pedestrian detection [33]. In the UAV domain, RGB-T fusion detectors such as SF-YOLO [34] have also been proposed for drone scenes. Therefore, these reasons contribute to the innovation level of this research, since publicly available UAV detection methods operating on fused data are still at an early stage.

After this brief introduction, in the following subsections we delve deeper into the three most important fusion methods families: early-fusion, mid-fusion and late-fusion.

1) EARLY-FUSION (PIXEL-LEVEL)

It combines the information of the two modalities at the pixel level. It is the most intuitive fusion approach, where IR and RGB images are concatenated to generate a 4-channel image, which is then fed into a regular object detection architecture. Several image-level fusion networks have been proposed in the literature, ranging from autoencoder-based methods such as DenseFuse [35] to illumination-aware architectures such as PIAFusion [36], IAIFNet [37] and FCDFusion [38]. These techniques rely on dedicated, relatively complex fusion architectures to better merge the images in the two spectra. However, as Wagner et al. show [39], pixel-level fusion was outperformed by deeper fusion variants in their pedestrian detection experiments, even though there are ways to improve it, such as the ones presented by Zhao et al. [40]. Since aligned RGB-Thermal datasets are not available, the method developed by Yang et al. [41] has also been evaluated, which uses a CycleGAN to create aligned synthetic RGB/IR data [42], [43]. In fact, it is important that the image pairs are perfectly aligned before performing early-fusion. In a more simplified way, this architecture can be described as a neural network composed of two different branches, dedicated respectively to RGB and IR image processing. In addition, this network presents a decision layer, which chooses how much importance to give to which of the two sub-networks.

In order to merge RGB and IR images, it is possible to use Python Imaging Library (PIL or Pillow) or dstack. The approach that these two techniques follow is based on the raw channel information overlapping. In fact, they take into consideration the bit depth of RGB images, which refers to the number of bits used to represent the color of a single pixel. It is usually 24 bits (8 bits for each color channel) plus other 8 bits for transparency (alpha channel). However, the transparency channel is not always present and it depends on the extension in which the image file is saved. In fact, if more than 3 channels are needed, the `.png` format needs to be considered, since `.jpg` supports a maximum of 3 channels. For IR images, only their thermal information (8 bits) should be considered, which will be added to the final fused image. Therefore, the resulting merged image will be of 32 bits.

2) LATE-FUSION (DECISION-LEVEL)

This fusion type merges detection results from separate models, improving robustness by leveraging complementary sensor characteristics. Furthermore, the aligned inputs (RGB and the respective IR counterpart) of the two modalities are fed separately into two unimodal object detection models to generate bounding boxes and the predicted boxes are fused using statistical methods, such as weighted averaging of the bounding box coordinates, which is based on the confidence scores from each modality. This method gives more importance to the bounding boxes with higher confidence scores, and this is a suitable option for the purpose of UAV detection.

This method works with both aligned and misaligned images. For this reason, it could be considered more resilient to sensor misalignment with respect to early fusion. However, when the same object appears with different orientations or in different positions across the two images, the resulting bounding boxes may be less precise in terms of localization. In practice, the bounding box tends to expand, and especially in the case of unaligned images, the object is still correctly identified while the box remains largely consistent across both views. As a result, the final bounding box may show slight variations in size and position.

In addition to the adopted strategy, there exist also ensemble methods such as probabilistic ensembling [32], in which a bounding box is considered valid if it is detected by both modalities.

3) MID-FUSION (FEATURE-LEVEL)

This strategy, which falls between early and late fusion techniques [44], might be the best and most balanced option to choose when trying to develop a computer vision system that exploits sensor fusion. Essentially, this fusion integrates feature maps extracted separately from RGB and IR modalities before applying detection networks. More specifically, features are captured from each of the two spectra using specific convolutional pipelines in the first stages. Subsequently, feature maps are merged using methods like concatenation or sum, which prepare information to be used for the final layers of detection. Taking into account these aspects, the mid-fusion idea has the ability to improve the robustness of computer vision and object detection systems in harsh environments with low light [45], [46].

This type of fusion has also many advantages compared with the other two. First of all, images do not need to be merged into a single one as early-fusion does, but all their channels are kept in the original form, preserving modality-specific information. Nor are the final detection results combined using statistical methods. Therefore, this strategy appears to be more flexible with respect to the input data type and the final detection goals that we may have. In fact, it has already been studied and implemented in some feature alignment methods for RGB and IR images [47]. However, despite these positive aspects, only the other two strategies have been implemented, as mid-fusion implies creating a

custom detection architecture, markedly different from the state-of-the-art YOLO one used for the tests. In addition, early-fusion and late-fusion allowed us to obtain useful metrics for comparing the strategies in several scenarios, which were not possible with mid-fusion.

III. LOGICAL PROPOSED ARCHITECTURES

The architecture proposed is based on the YOLO framework, since it represents the state-of-the-art in the real-time object detection field. Moreover, two separate fusion pipelines have been developed for enhancing detection performance and accuracy. Finally, a multimodal dataset [48] of real-world data on still and moving drones was built, for testing the validity of the following fusion methods. This approach has been preferred to the synthetic one because real acquired data better reflect operational conditions and lead to superior performance. In addition, the actual behaviour of a drone, that is, its ability to rotate, shift and change pose over the three axes (roll, pitch and yaw), has also been considered while collecting data.

A. EARLY FUSION

The early fusion approach combines RGB and IR images at the pixel level to create a single multimodal input. This approach requires frames to be aligned. This process was performed by applying transformation steps to each image, such as interpolation, perspective distortion and rotation. After this, an overlap of the channels was performed, while removing the transparency level (also called "A" in RGBA image format), which was present in `.png` images. However, since the focus of this research strictly concerns the analysis of fusion strategies, the reader is referred to the paper on autonomous UAV detection architecture development [49], in which all alignment steps are described.

Once spatial alignment between RGB and infrared frames was achieved, the fusion was performed through overlap, by stacking the image channels into a single array of 4 spectra. The RGB image was split into three channels: Red, Green and Blue. Each of them is represented by a two-dimensional intensity map, where every pixel contains intensity values for that particular channel. A visual interpretation is shown in Figure 1. On the other hand, the IR image was converted into a grayscale single-channel image. These channels were then combined into a 4-channel image (RGBT), where the fourth channel corresponded to the thermal information. The operation result is shown in Figure 2.



FIGURE 1. From left to right: red, green and blue channel's intensity maps.

For implementation simplicity, the Python Imaging Library (PIL) was employed, as it allows independent access to individual image channels. Initial experiments involved the creation of four-channel RGBT images obtained by stacking the red, green, blue, and thermal components into a single tensor. In addition to direct channel stacking, pretrained fusion models such as PIAFusion in Figure 3, FCDFusion in Figure 4 and DenseFuse in Figure 5 were evaluated to assess alternative image-level fusion strategies. These approaches generated fused outputs without requiring manual channel manipulation. However, Ultralytics introduced support for multi-channel training only recently [50], subsequent to the experiments reported here. Hence, although OpenCV2 library, commonly used within YOLO pipelines, allows image loading with an arbitrary number of channels using the `cv2.IMREAD_UNCHANGED` flag, in previous versions of Ultralytics, in order to use 4-channel inputs, extensive modifications to the YOLO architecture would be required, including adaptation of the model's input layer and associated preprocessing routines. In order to preserve architectural consistency and ensure fair comparison with baseline experiments, modifications to the core framework and alteration to the network structure were avoided.

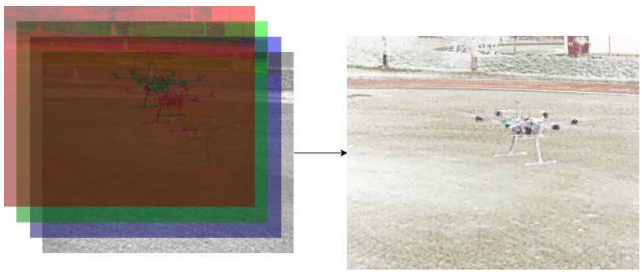


FIGURE 2. Graphical representation of how layers are overlapped by Pillow library in order to get the resulting image on the right.



FIGURE 3: PIAFusion



FIGURE 4: FCDFusion



FIGURE 5: DenseFuse

FIGURE 6. Visible comparison between state-of-the-art fusion strategies, exploiting IAN framework for multimodal image generation.

Instead, the characteristics of RGB channels were taken into consideration. RGB usually exhibits a high degree of spectral correlation, since most objects in the environment reflect light similarly across the visible spectrum. Both limitations and contributions of each channel have been summarized in Table 1. Due to these highlights reported,

TABLE 1. Summary of the characteristics of the RGB channels [51].

Channel	Main contribution	Limitations / Remarks
Green	Contributes the most to the perception of luminance and fine details.	Provides high structural information.
Red	Crucial for discriminating between minerals and soils.	Complementary for material recognition.
Blue	Captures wavelengths in the 450–495 nm range, useful for identifying water bodies and atmospheric effects.	Lower contrast, limited thermal and low-light contributions, sensitive to Rayleigh scattering, noisier in long-distance imagery.

a modified 3-channel tensor was created by discarding the blue channel, which turned out to be redundant in several outdoor UAV scenes where red and green spectral content are often predominant. The blue channel was then replaced with the thermal component, mitigating some of the atmospheric noise. This resulted in an RGT image (Red, Green, Thermal), maintaining compatibility with YOLOv8 architecture while integrating complementary spectral information. Consequently, a 3-channel fusion representation was adopted. Therefore, the same channel-stacking principles used in the 4-channel implementation were applied to construct the 3-channel fused images. The methodological workflow is summarized in the dedicated flowchart in Figure 7. The process shown in Figure 8 can be summarized as a shift on the left of all the image channels, in order to include the thermal one inside the final image. Therefore, the resulting fused image consisted of red, green and thermal channels, forming an RGT representation. Some examples are shown in Figure 9. The distinction between a CIR image [52] and the proposed RGT one relies on the infrared type chosen. In fact, the camera sensor used for the tests operates in far-infrared (FIR) range, whereas CIR relies on near-infrared (NIR) imagery.

This design enabled YOLOv8s model training without architectural modification and allowed operation with the traditional YOLO detection pipeline while reducing potentially redundant spectral information.

Formally, each spectral channel is defined as a function over pixel coordinates (i, j):

$$R(i,j): \text{ Red intensity at pixel } (i,j) \tag{1}$$
$$G(i,j): \text{ Green intensity at pixel } (i,j) \tag{2}$$
$$T(i,j): \text{ Thermal intensity at pixel } (i,j) \tag{3}$$

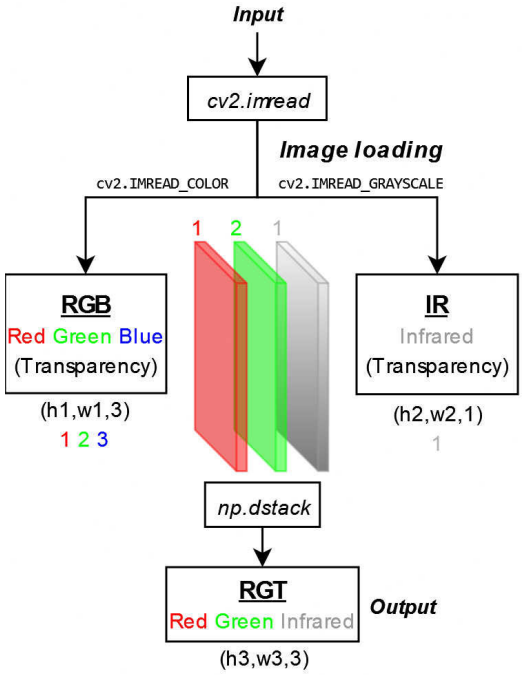


FIGURE 7. Early Fusion framework.

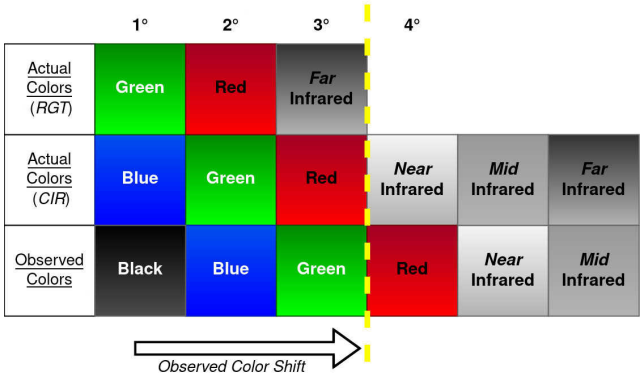


FIGURE 8. Channel shifting process. As it can be seen, there is a clear difference between CIR and RGT representations. Color-infrared imagery uses a portion of the electromagnetic spectrum known as near infrared, which ranges from 0.70μm to 1.0μm. RGT uses instead far infrared. In both cases, blue wavelengths are shifted out of the visible portion of the spectrum. However, only 3 channels are considered part of the RGT representation, without considering the non-visible part of the spectrum.

The fused image tensor is obtained by stacking these channels:

$$I_{RGT}(i,j) = \begin{bmatrix} R(i,j) \\ G(i,j) \\ T(i,j) \end{bmatrix} \tag{4}$$

Each pixel location $(i, j) \in \mathbb{Z}^2$ is therefore mapped to a three-dimensional spectral vector in $\mathbb{R}^3$:

$$I_{RGT} : \mathbb{Z}^2 \to \mathbb{R}^3, \quad (i,j) \mapsto (R(i,j), G(i,j), T(i,j)). \tag{5}$$

The complete fused image can thus be expressed as a tensor:

$$I_{RGT} \in \mathbb{R}^{H \times W \times 3}, \tag{6}$$

where H and W denote image height and width, respectively.

The same pipeline was also adopted for the other two channel combinations, respectively RBT and GBT, each obtained by omitting one of the three RGB channels. This enabled an ablation study on channel selection, providing a formal justification for the choice of RGT in the proposed testing scenario.



FIGURE 9. From the top: RGT, RBT and GBT fusion strategies. In all the cases only 3 of the 4 channels are considered for creating the fusions. For RGT the blue channel has been omitted in favour of the thermal one. A similar type of image is usually called CIR or false-color image [53]. The other types have only been considered for a technical comparison.

After generating and augmenting a substantial number of fused images, the YOLOv8s detector was trained using the standard (visible and infrared) and the fused datasets. All of them contain a total of 300 images for the training part and 100 for the validation one, corresponding to a split of 75% and 25% respectively. The results obtained after training and validation processes were compared against models trained exclusively on RGB and IR data.

These findings can be interpreted in light of the Data Fusion Principle.[1] Accordingly, a fusion-based system is expected to outperform its unimodal counterparts if complementary information is effectively exploited.

The comparative performance metrics obtained on unimodal and multimodal datasets are summarized in Table 2. While the standard RGB configuration achieved higher precision, the overall detection performance did not uniformly surpass the fused models across all evaluation metrics. In fact, the superiority of the three proposed fusion strategies first emerges on mAP@50-95. The highest one is reached by RGT with 56.2%. A similar improvement can be observed on mAP@50, where the proposed fused approach achieves a result of 97.8%, compared to 89.8% of Infrared. These are indexes of good model performance, since mAP evaluates it over all confidence thresholds, rather than at a single operating point.

[1]The integration of multiple information sources should yield performance superior to that of each individual modality considered separately.

TABLE 2. Performance comparison of YOLOv8s model over standard spectra (RGB, IR), literature fusions (FCD, PIA, Dense) and proposed early fusion approaches (RGT, RBT, GBT). The model has been trained for 300 epochs over the same dataset, only changing the color channels and the fusion strategies for the images inside it.

Dataset	Precision	Recall	mAP@50	mAP@50-95	F1-Score
RGT	**95.9%**	**94.6%**	**97.8%**	**56.2%**	**95.2%**
RBT	**96.9%**	**93.8%**	**96.9%**	**55.6%**	**95.3%**
GBT	**93.2%**	**85%**	**92.4%**	**53.3%**	**88.9%**
FCDFusion [38]	98.9%	91.8%	96.8%	56.9%	95.2%
PIAFusion [36]	96.2%	76.4%	84.6%	42.2%	85.2%
DenseFuse [35]	95.5%	84.7%	91.5%	50.3%	89.8%
IR	89%	80.8%	89.8%	51.5%	84.7%
RGB	97.9%	91.7%	97%	48.8%	94.7%

Another important point regards F1-Score, which penalizes models with unbalanced precision and recall values.

All these results highlight the guaranteed improvements of RGT fusion in comparison to the other strategies reported. In fact, the metrics suggest that the proposed multimodal approach has gained the capability to enrich the scope and improve the stability of the model in object detection tasks. In addition, the comparison between fusion strategies enriches the study, highlighting the importance of a dedicated fusion approach with respect to the use-case scenario.

Another interesting point regards the results obtained from literature fusion strategies. All of them cannot be considered for real-time purposes, since they rely on dedicated fusion networks (illumination-aware or autoencoder-based) whose computational cost is unsuitable for real-time processing. Thus, even though the FCD approach gives promising results, it is not usable in real-time scenarios.

According to the results obtained, the RGT fusion strategy was also better than the other two presented (RBT and GBT). For this reason, it was selected for tests in real-world scenarios. Nevertheless, it should be pointed out that this is a preliminary evaluation step for the topic. It could surely be enriched and extended with a larger fused dataset and other object detection models, in order to carry on the research in the field.

B. LATE FUSION

In the late fusion approach aligned RGB and IR images are processed separately by two object detection models. The resulting bounding boxes are then fused using a weighted average method. For each pair of boxes, the center coordinates are averaged based on their confidence scores, while the maximum width and height are kept to avoid box shrinking. A distance threshold of 50 pixels is used to decide whether two boxes should be fused. In particular, if the centers of two boxes are within this range, they are fused, otherwise they are left as they are. More in detail, the Euclidean distance between the center points of two bounding boxes in the two spectra is computed. If this distance falls in the threshold set, weighted average is performed. Otherwise, the original box created by the model is preserved. This approach has been designed to avoid discarding detections that may only exist in one modality and to work in case of

misaligned frames. In fact, if the centers of the two bounding boxes are near, the system fuses them. Alternatively, it preserves what each model has decided during detection.

In the decision-level fusion strategy, two spatially aligned images, one RGB and its corresponding infrared (IR) counterpart, are processed independently by their respective detection models. Each detector produces a set of bounding boxes with associated confidence scores. Fusion is subsequently performed at the bounding box level using a weighted averaging method.

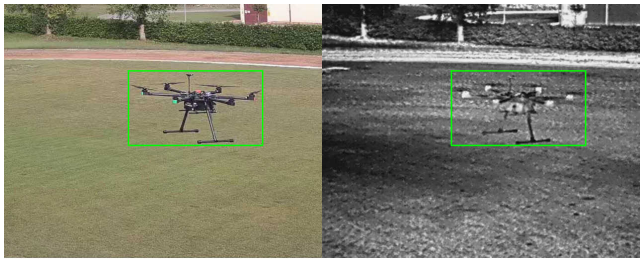


FIGURE 10. The bounding box is unique for both images. As it was seen in the experiments, it could work even with misaligned frames.

The weighted average approach computes the fused bounding box coordinates based on the confidence scores of the individual detections. This method can be applied to both perfectly aligned and moderately misaligned image pairs. However, if the same object appears with slight positional or orientational discrepancies across the two modalities, the resulting fused bounding box may exhibit reduced localization precision. In such cases, the fused bounding box may become enlarged, encompassing the object regions identified in both modalities within a single unified box.

The fusion procedure begins with the computation of the center coordinates of the new bounding box using confidence-weighted averaging. The width and height of the fused bounding box are then defined as the maximum of the corresponding dimensions from the RGB and IR detections. This design choice prevents the fused bounding box from shrinking excessively and ensures that the spatial extent of the detected object is fully preserved.

A dedicated fusion function was implemented to manage all detection scenarios. Since decision-level fusion requires detections from both modalities, situations may arise in which a UAV is detected by only one of the two models. In such cases, the bounding boxes generated by the detecting model are retained without modification.

In order to regulate the fusion process, a heuristic distance threshold (δ) was introduced. In particular, the sensor we used offers two non-coaxial lenses, which are physically separated and present a parallax error. Therefore, the offset between the two images varies depending on the distance of the sensor from the drone and on frame resolution.

For each pair of RGB and IR bounding boxes, the Euclidean distance between their center coordinates has also been computed. If this distance is below the threshold,

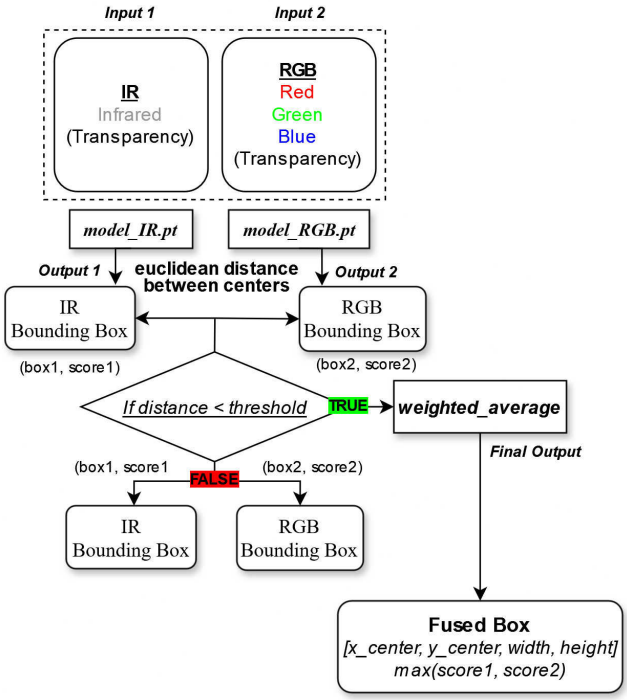


FIGURE 11. Late Fusion logic.

weighted averaging is applied. Otherwise, the bounding boxes are preserved independently. This strategy enables the method to operate effectively under moderate sensor misalignment. When detections are spatially consistent across modalities, fusion is performed; when they differ significantly, each model's decision is maintained.

Formally, consider two bounding boxes

$$B_1 = (x_1, y_1, w_1, h_1), \quad B_2 = (x_2, y_2, w_2, h_2),$$

with associated confidence scores s_1 and s_2. The center coordinates of the fused bounding box are computed as confidence-weighted averages:

$$x_c = \frac{s_1 x_1 + s_2 x_2}{s_1 + s_2} \tag{7}$$

$$y_c = \frac{s_1 y_1 + s_2 y_2}{s_1 + s_2} \tag{8}$$

The width w and height h of the fused bounding box are defined as the maximum between the values of the two original boxes.

The resulting fused bounding box is therefore represented as:

$$B_f = (x_c, y_c, w, h) \tag{9}$$

To determine whether fusion should be applied, the Euclidean distance between the centers $C_1 = (x_1, y_1)$ and $C_2 = (x_2, y_2)$ is computed as:

$$D(C_1, C_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{10}$$

Fusion is performed if:

$$D(C_1, C_2) < \delta \tag{11}$$

where δ is the empirical distance threshold. The final confidence score of the fused bounding box is the maximum of the two score values.

If the distance condition is not satisfied, the original bounding boxes are preserved. This formulation ensures robustness to moderate spatial discrepancies between modalities while maintaining reliable detection outputs in the presence of partial or asymmetric detections.

IV. EXPERIMENTAL TESTS

After computing the performance metrics of each model, the results gained from tests in real-world scenarios are analyzed and compared, both visually and empirically. The analysis will cover all the phases with both strategies: Early and Late Fusion.

A. ACQUISITION SCENARIOS

After recording all data in the two modalities, some results are presented as frames of the collected videos, discussing both benefits and limitations of the two systems in all the conditions tested.



FIGURE 12. The outdoor testing field · Google Earth Pro. The takeoff field is highlighted in green. Inside it, in the red area, the master drone used for detection takes off. In the rest of the yellow scenario, a target drone flies at different altitudes, in order to test the capabilities of the detection models.

The metric considered in external tests is the confidence score calculated by the model during detection. This value indicates the probability of an unidentified flying object belonging to the UAV class and it has allowed us to perform an early qualitative review of the system in different environmental conditions. In the final part of the section, some quantitative results are also presented in a dedicated table, in which other metrics are considered and analyzed, in order to give a review of the two frameworks as complete as possible. Different lighting scenarios were faced during the experiments. For safety reasons, a maximum distance of 60 meters and an altitude of 20 m were set for all the tests. Zooming in was also used in borderline cases, where the background was wooded and the drone was camouflaged among the leaves, thus difficult to detect. Confidence scores were increasing when zooming in and the drone was still detected with all degrees of zoom and at great distances. However, detection at 1X is sometimes intermittent, whereas 2X and 4X work reliably, making the system more stable. Furthermore, the score drops when the target flies in front of an obstacle. At those hours, temperatures were lower, which prevented disturbances from light and heat. On the other hand, the sky background greatly improved the confidence score, even with clouds. In fact, the cloud formation let light pass through and the environment was at least sufficiently illuminated, enhancing the contrast between the target UAV and the sky color.

Late-fusion technique was more challenging due to the management of two independent models. As can be seen, the stream was aligned during detection, in order to have both frames in the same plane with a horizontal sync. This method proves particularly useful in environments in which one of the two modalities is less reliable. In fact, the best of the two helps the other to perform detection, thanks to the weighted average strategy developed. As before, sky offers higher confidence scores, due to the better contrast between the UAV and the background. However, some small issues have been discovered with trees and terrain setup, as they generate clutter, a phenomenon that challenges the optical sensors. In addition, leaves make detection even more difficult due to the openings inside the trees, which let light in, reducing the possibility of detection. Moreover, since we recorded on summer days, the heat and brightness affected the infrared sensors, generating noise. Another aspect which should be taken into consideration regards UAV movements. If the drone is stationary or changes direction quickly, it can be lost. However, if the drone is in the air, it is much better identified, even at great distances.

TABLE 3. Summary of the experimental evaluation conducted at different hours of the day, under different environmental conditions. The target drone was moving at different altitudes and distances from the master drone, while covering two main backgrounds: trees and sky. In the same way, three zoom levels have been tested.

Time	Fusion	Zoom	Background	Distance	Altitude
09:00	Early	1×, 2×, 4×	Trees	5, 25, 60 m	5, 10, 20 m
09:00	Early	1×, 2×, 4×	Sky	5, 25, 60 m	5, 10, 20 m
09:00	Late	1×, 2×, 4×	Trees	5, 25, 60 m	5, 10, 20 m
09:00	Late	1×, 2×, 4×	Sky	5, 25, 60 m	5, 10, 20 m
12:00	Early	1×, 2×, 4×	Trees	5, 25, 60 m	5, 10, 20 m
12:00	Early	1×, 2×, 4×	Sky	5, 25, 60 m	5, 10, 20 m
12:00	Late	1×, 2×, 4×	Trees	5, 25, 60 m	5, 10, 20 m
12:00	Late	1×, 2×, 4×	Sky	5, 25, 60 m	5, 10, 20 m
15:00	Early	1×, 2×, 4×	Trees	5, 25, 60 m	5, 10, 20 m
15:00	Early	1×, 2×, 4×	Sky	5, 25, 60 m	5, 10, 20 m
15:00	Late	1×, 2×, 4×	Trees	5, 25, 60 m	5, 10, 20 m
15:00	Late	1×, 2×, 4×	Sky	5, 25, 60 m	5, 10, 20 m

V. RESULTS

This section presents a qualitative and quantitative comparison between the two proposed fusion approaches. Table 4 summarizes the validation metrics computed from both

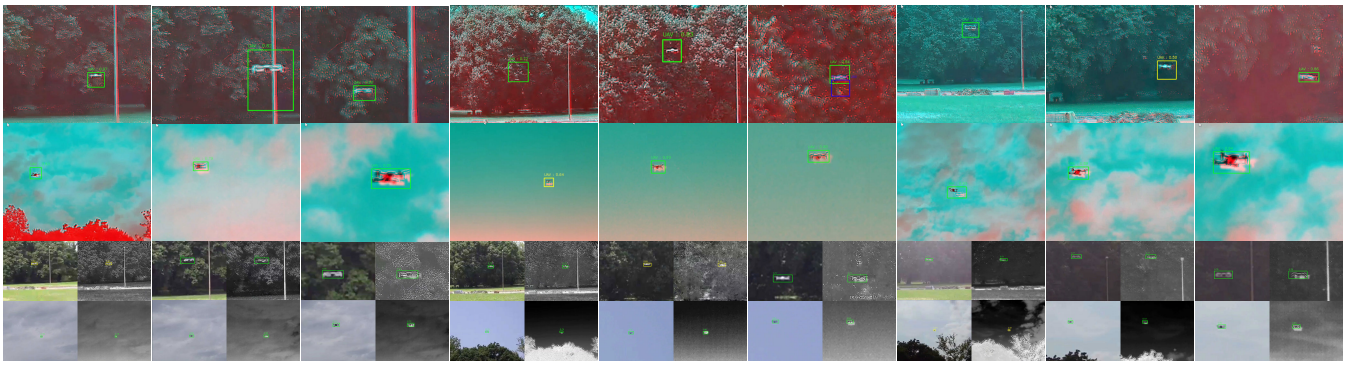


FIGURE 13. Ground and sky comparisons at different times and resolutions. From the top left corner of the image, it is possible to see the first line, regarding Early Fusion RGT process on the ground at different zoom levels (1X, 2X, 4X) in different times of day (9:00, 12:00, 15:00). Following it, there is a second line showing the same fusion techniques applied for detecting over a sky background. Finally, the third and fourth lines contain images of tests conducted in the same conditions presented before while using Late Fusion instead.

modalities. The evaluation process for both modalities follows a common series of steps. In particular, validation was conducted on a pool of $N = 100$ spatially aligned images, while inference was performed using a default Intersection-over-Union (IoU) threshold of 0.5.

TABLE 4. Comparison of mean metrics between Late-Fusion (RGB-IR) and Early-Fusion (RGT) strategies (where mR is mean Recall and mF1 is mean F1).

Fusion Method	mAP@50	mR	mF1-score
Late-Fusion (RGB-IR)	72%	77%	73.7%
Early-Fusion (RGT)	**75.7%**	**83%**	**77.8%**

For the Early-Fusion (RGT) approach, predictions generated by the detector for each image were compared with the corresponding ground-truth annotations using an overlap mechanism. This step allowed the computation of global True Positive, False Positive and False Negatives in the entire pool of images.

$$TP = \sum_{i=1}^{N} TP_i \qquad FP = \sum_{i=1}^{N} FP_i \qquad FN = \sum_{i=1}^{N} FN_i \quad (12)$$

These quantities were then used to compute average precision, recall and F1-score.

In contrast, the Late-Fusion approach performs inference independently over each RGB-IR image pair. Then, a set of fused bounding boxes is produced by the weighted average algorithm described in Equation 9. These fused detections are defined as:

$$\hat{\mathcal{D}}_i = \left\{ B_f^{(1)}, B_f^{(2)}, \ldots, B_f^{(M_i)} \right\}, \quad (13)$$

This set was then compared with the set of ground-truth annotations $\mathcal{G}_i$.

$$\mathcal{G}_i = \left\{ G_i^{(1)}, G_i^{(2)}, \ldots, G_i^{(N_i)} \right\}, \quad (14)$$

Since the image pair is composed of geometrically aligned RGB and IR frames, they share a single annotation file. Therefore, the label file is unique across both spectra. This allows the comparison between fused boxes and ground-truth.

From the results shown in Table 4, Early-Fusion turned out to be the better of the two, achieving the best overall performance. This point also reinforces the results obtained from previous trainings in Table 2. Despite the final outcome, Late-Fusion strategy still provides competitive results, demonstrating the effectiveness of combining complementary information for aerial target detection.

Finally, some detection results showing the overlapping step performed by the validation process are visible in Figure 14.

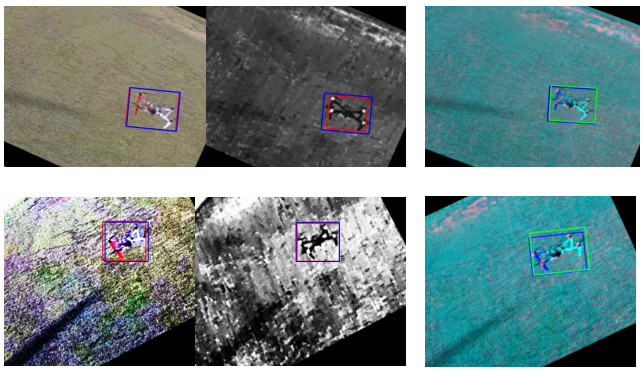


FIGURE 14. Comparison between blue ground truth boxes and models predicted boxes (red for Late-Fusion and green for Early-Fusion).

VI. CONCLUSION AND FUTURE WORK

It has been quite difficult to identify the best strategy of the two fusion techniques. In fact, both early and late fusions gave promising outputs. However, they could be exploited in different contexts. For example, late fusion might be advantageous at night, when the RGB counterpart is completely lost, whereas IR offers better performance and could support it. On the other hand, early fusion is the best for general use, as it improves detection by adopting channel blending, considering the tests conducted and the results

acquired. Despite this advantage, early fusion is harder to implement and run, due to the need to train a dedicated model over a fused dataset and its dependence on an accurate pre-alignment process.

Although the system reached good overall results, both mathematically and empirically, the need for more multi-modal data for future research remains crucial. In particular, the proposed solution was specifically designed for drone-to-drone detection architectures. For this reason, it is necessary to integrate this work with more data, in order to adapt this strategy to different detection pipelines.

Some future improvements to the project will surely involve an upgrade to the drone platform, enabling the use of newer and more accurate sensors for detection. This is indeed linked to the exploitation of different and complex image types, such as multispectral or hyperspectral images. However, the use of other sensor data should also be taken into consideration, since radar platforms are widely used for developing sensor fusion paradigms in modern aviation.

REFERENCES

[1] J. Zhao, J. Zhang, D. Li, and D. Wang, "Vision-based anti-UAV detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25323–25334, Dec. 2022.

[2] F. Samadzadegan, F. Dadrass Javan, F. Ashtari Mahini, and M. Gholamshahi, "Detection and recognition of drones based on a deep convolutional neural network using visible imagery," *Aerospace*, vol. 9, no. 1, p. 31, Jan. 2022, doi: 10.3390/aerospace9010031.

[3] O. Skorka and R. Ispasoiu, "Tradeoffs with RGB-IR image sensors," *IEEE Trans. Electron Devices*, vol. 69, no. 6, pp. 2915–2923, Jun. 2022.

[4] Y. Sun, Y. Meng, Q. Wang, M. Tang, T. Shen, and Q. Wang, "Visible and infrared image fusion for object detection: A survey," in *Proc. Int. Conf. Image, Vision Intell. Systems (ICIVIS)*, 2024, pp. 236–248.

[5] U. Seidaliyeva, L. Ilipbayeva, K. Taissariyeva, N. Smailov, and E. T. Matson, "Advances and challenges in drone detection and classification techniques: A state-of-the-art review," *Sensors*, vol. 24, no. 1, p. 125, Dec. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/24/1/125

[6] S. Scholes, A. Ruget, G. Mora-Martín, F. Zhu, I. Gyongy, and J. Leach, "DroneSense: The identification, segmentation, and orientation detection of drones via neural networks," *IEEE Access*, vol. 10, pp. 38154–38164, 2022.

[7] O. E. Díaz, A. S. López, M. B. B. Loranca, and R. G. Velázquez, "Real time drone object tracking using histograms of oriented gradients and particle filters," in *Proc. 15th Mex. Int. Conf. Artif. Intell. (MICAI)*, Oct. 2016, pp. 35–40.

[8] F. Liu, L. Yao, C. Zhang, T. Wu, X. Zhang, X. Jiang, and J. Zhou, "Boost UAV-based ojbect detection via scale-invariant feature disentanglement and adversarial learning," 2024, *arXiv:2405.15465*.

[9] F. N. M. Zamri, T. S. Gunawan, M. Kartiwi, A. Pratondo, S. H. Yusoff, and Y. M. Mustafah, "Deep learning techniques for advanced drone detection systems: A comprehensive review of techniques, challenges and future directions," *Indonesian J. Electr. Eng. Informat. (IJEEI)*, vol. 12, no. 4, pp. 931–944, Nov. 2024.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. ECCV*, 2016, pp. 21–37.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[13] P. Hidayatullah, N. Syakrani, M. R. Sholahuddin, T. Gelar, and R. Tubagus, "YOLOv8 to YOLO11: A comprehensive architecture in-depth comparative review," 2025, *arXiv:2501.13400*.

[14] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[15] S. V. Rebbapragada, P. Panda, and V. N. Balasubramanian, "C2FDrone: Coarse-to-fine drone-to-drone detection using vision transformer networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 6627–6633.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020, pp. 213–229.

[17] T. Sangam, I. R. Dave, W. Sultani, and M. Shah, "TransVisDrone: Spatio-temporal transformer for vision-based drone-to-drone detection in aerial videos," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 6006–6013.

[18] K. AlDosari, A. Osman, O. Elharrouss, S. AlMaadeed, and M. Z. Chaari, "Drone-type-set: Drone types detection benchmark for drone detection and tracking," 2024, *arXiv:2405.10398*.

[19] M. Bilal, S. Rehmat, N. Akhtar, and H. Gulzar, "Cost effective drone detection with deep neural networks for day and night surveillance," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2024, pp. 1–6.

[20] M. W. Ashraf, W. Sultani, and M. Shah, "Dogfight: Detecting drones from drones videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7063–7072.

[21] A. Sengupta, L. Cheng, and S. Cao, "Robust multiobject tracking using mmWave radar-camera sensor fusion," *IEEE Sensors Lett.*, vol. 6, no. 10, pp. 1–4, Oct. 2022.

[22] K. Yang, W. Xiang, Z. Chen, J. Zhang, and Y. Liu, "A review on infrared and visible image fusion algorithms based on neural networks," *J. Vis. Commun. Image Represent.*, vol. 101, May 2024, Art. no. 104179. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320324001342

[23] Y. Wang, Y. Chen, J. Choi, and C.-C.-J. Kuo, "Towards visible and thermal drone monitoring with convolutional neural networks," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, p. 5, 2019.

[24] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[25] Y. Liu, Z. Wu, X. Han, Q. Sun, J. Zhao, and J. Liu, "Infrared and visible image fusion based on visual saliency map and image contrast enhancement," *Sensors*, vol. 22, no. 17, p. 6390, Aug. 2022.

[26] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[27] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[28] S. Özer and A. P. Ndigande, "VisIRNet: Deep image alignment for UAV-taken visible and infrared image pairs," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024, doi: 10.1109/TGRS.2024.3367986.

[29] H. Hao, Y. Peng, Z. Ye, B. Han, W. Tang, W. Kang, X. Zhang, Q. Li, and W. Liu, "TMRGBT-D2D: A temporal misaligned RGB-thermal dataset for drone-to-drone target detection," *Drones*, vol. 9, no. 10, p. 694, Oct. 2025.

[30] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 403–411.

[31] F. Xu, Z. Lin, C. Chen, R. Yang, and H. Zhai, "Anti-drone detection in aerial multispectral images," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 9388–9391.

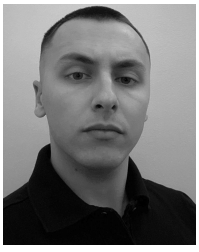
[32] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Proc. ECCV*, 2022, pp. 139–158.

[33] D. Guan, Y. Cao, J. Liang, Y. Cao, and M. Ying Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," 2018, *arXiv:1802.09972*.

[34] S. Zhang, T. Xie, Y. Wang, Y. Liu, Y. Dou, and S. Wang, "SF-YOLO: RGB-T fusion object detection in UAV scenes," in *Proc. 8th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2023, pp. 51–59.

[35] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[36] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vols. 83–84, pp. 79–92, Jul. 2022.
[37] Q. Yang, Y. Zhang, Z. Zhao, J. Zhang, and S. Zhang, "IAIFNet: An illumination-aware infrared and visible image fusion network," *IEEE Signal Process. Lett.*, vol. 31, pp. 1374–1378, 2024.
[38] H. Li and Y. Fu, "FCDFusion: A fast, low color deviation method for fusing visible and infrared image pairs," *Comput. Vis. Media*, vol. 11, no. 1, pp. 195–211, Feb. 2025.
[39] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proc. 24th Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn. (ESANN)*, vol. 587, 2016, pp. 509–514.
[40] T. Zhao, M. Yuan, F. Jiang, N. Wang, and X. Wei, "Removal then selection: A coarse-to-fine fusion perspective for RGB-infrared object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 27, no. 2, pp. 2504–2519, Feb. 2026.
[41] L. Yang, R. Ma, and A. Zakhor, "Drone object detection using rgb/ir fusion," in *Proc. Electron. Imag. Symp., Comput. Imag. Conf.*, 2022, pp. 1–6.
[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2017, pp. 1125–1134.
[44] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion (FUSION)*, Jul. 2020, pp. 1–6.
[45] X. Liu, Z. Rozsypálek, and T. Krajník, "Self-supervised learning for fusion of IR and RGB images in visual teach and repeat navigation," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Sep. 2023, pp. 1–7.
[46] T. Tian, J. Cai, Y. Xu, Z. Wu, Z. Wei, and J. Chanussot, "RGB-infrared multi-modal remote sensing object detection using CNN and transformer based feature fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 5728–5731.
[47] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3622–3631.
[48] D. Tavaris, A. de Zan, A. Toma, G. Luca Foresti, I. Scagnetto, and N. Martinel, "Multi-perspective RGB and infrared dataset for UAV detection," *IEEE Access*, vol. 13, pp. 168792–168803, 2025.
[49] A. De Zan, D. Tavaris, G. L. Foresti, I. Scagnetto, N. Martinel, and A. Toma, "Autonomous real-time drone-to-drone visual detection by onboard hardware platform," *Integr. Computer-Aided Eng.*, vol. 33, no. 3, pp. 368–381, Aug. 2026.
[50] G. Jocher, A. Chaurasia, and J. Qiu. (2025). *Ultralytics YOLOv8 (v8.3.112)*. [Online]. Available: https://github.com/ultralytics/ultralytics
[51] H.-C. Cheng, I. Ben-David, and S.-T. Wu, "Five-primary-color LCDs," *J. Display Technol.*, vol. 6, no. 1, pp. 3–7, Jan. 2010.
[52] Minnesota Geospatial Information Office. (2026). *Color-Infrared (CIR) Imagery*. [Online]. Available: https://www.mngeo.state.mn.us/chouse/airphoto/cir.html
[53] (2026). *What Do the Different Colors in a Color Infrared (CIR) Aerial Photograph Represent*. [Online]. Available: https://www.usgs.gov/faqs/what-do-different-colors-a-color-infrared-aerial-photograph-represent



ALBERTO DE ZAN received the bachelor's degree (cum laude) in computer science from the University of Udine, Italy, in 2023, on a thesis with title "drones cluster development for air surveillance and defense," the double degree program in artificial intelligence and cybersecurity from the University of Udine and Klagenfurt, Austria, and the Graduate degree (Hons.), in 2025, on a thesis with title "unmanned aerial vehicles detection adopting multiple sensor fusion." He is currently a Research Fellow on some projects regarding computer vision, radar platforms, drone detection, open source intelligence, and quantum computing with the Autonomous System Laboratory, University of Udine.



DENIS TAVARIS is currently pursuing the Ph.D. degree in artificial intelligence with the University of Udine, specializing in computer vision applied to robotics. His research interests include the development of autonomous and cooperative multi-drone systems for environmental monitoring, integrating vision-based technologies, and sensor fusion. He is involved in several research projects on autonomous systems, multi-domain robotics (aerial, surface, and underwater), and AI-driven environmental analysis. He has published papers in international journals and conferences, with a focus on efficient and scalable solutions for autonomous monitoring. He has a strong background in data analysis, system design, and software development, with expertise in Python, Linux, and sensor integration.



GIAN LUCA FORESTI (Senior Member, IEEE) received the Laurea degree (cum laude) in electronic engineering and the Ph.D. degree in computer science from the University of Genoa, Genoa, Italy, in 1990 and 1994, respectively. He is currently a Full Professor in computer science with the University of Udine, Udine, Italy, where he is also the Deputy Director of the Department of Mathematics, Computer Science and Physics. Since 2006, he has been a Visiting Professor in artificial vision with the University of Klagenfurt, Klagenfurt, Austria. From 2000 to 2009, he was the appointed Italian Member of NATO RTO Information System Technology Panel. He has authored over 300 papers published in international journals and international conferences. He has co-edited several books in the field of multimedia and video surveillance. His current research interests include computer vision and image processing, multisensor data and information fusion, pattern recognition, and neural networks. He is a member of IAPR and CVPL. He was a recipient of the Best IEEE Vehicular Electronics Paper, in 2002. He was the Finance Chair of the 11th IEEE Conference on Image Processing and the General Chair of the 16th International Conference on Image Analysis and Processing and the Eighth IEEE International Conference on Advanced Video and Signal Based Surveillance. He has been the Guest Editor of a Special Issue of the Proceedings of the IEEE on "Video Communications, Processing and Understanding for Third Generation Surveillance Systems."



IVAN SCAGNETTO received the Ph.D. degree in computer science from the University of Udine, in 2002. Since 2005, he has been an Assistant Professor with the University of Udine. His research interests include software certification by means of formal techniques based on computer aided formal reasoning to mobile computing, information retrieval with mobile devices, and the IoT systems. He is the author of several scientific papers with national and international peer review and he has been a scientific reviewer of international journals and conferences. He worked in several research projects funded by European Commission, Italian Ministry of University and Research (MUR), INTERREG Program, and National Military Research Plan (PNRM).