EMPIRICAL ANALYSIS OF RGB-IR FEATURE FUSION FOR UAV-BASED OBJECT DETECTION

Thi Lan Nguyen[1], *Cao Truong Tran*[2,*]

Abstract

Object detection based on Unmanned Aerial Vehicles (UAVs) plays a crucial role in applications such as surveillance, disaster management, and military operations. However, traditional methods relying solely on visible Red-Green-Blue (RGB) imagery often perform poorly under low-light conditions and occlusions. To overcome these challenges, recent studies have explored the fusion of RGB and infrared (IR) images, leveraging their complementary properties. Among various fusion strategies, feature-level fusion has gained increasing attention due to its flexibility and superior performance compared to pixel-level and decision-level approaches. Despite its potential, the impact of the specific stage within the network where modality-specific features are integrated remains insufficiently investigated. This study focuses on feature-level fusion and conducts a comprehensive empirical analysis within a unified dual-stream detection framework to examine how fusion at different depths-early, middle, and late-affects detection performance. Additionally, we evaluate multi-position fusion schemes by combining features from multiple levels. Experimental results on the DroneVehicle dataset reveal that middle fusion achieves the best balance between detection accuracy and efficiency among single-layer fusion configurations. Furthermore, early-middle multi-position fusion further improves localization precision, albeit with moderate computational overhead. These findings offer practical insights into designing more effective and efficient RGB-IR fusion networks for UAV-based object detection systems.

Index terms

UAV-based object detection; feature-level fusion; deep learning; RGB-IR fusion.

1. Introduction

UAVs have emerged as a powerful platform for real-time object detection in a wide range of applications, including surveillance, disaster response, traffic monitoring, and and military reconnaissance [1]–[3]. However, accurate object detection in UAV imagery remains a challenging task due to factors such as varying altitudes, dynamic lighting

[1] Institute of Simulation Technology, Le Quy Don Technical University

[2] Institute of Information and Communication Technology, Le Quy Don Technical University

[*] Corresponding author, email: truongct@lqdtu.edu.vn

DOI: 10.56651/lqdtu.jst.v14.n01.1047.ict

conditions, and complex backgrounds. Traditional methods that rely solely on RGB images often suffer from performance degradation in low-visibility scenarios, such as nighttime or foggy environments [4], [5].

To address these challenges, recent research has explored the fusion of RGB and IR imagery, leveraging the complementary strengths of both modalities. While RGB images provide fine-grained texture and color information, IR images offer thermal cues that are resilient to lighting variations and can reveal obscured objects. Combining these modalities has demonstrated significant potential for improving detection robustness under low-light conditions [6], [7]. RGB-IR fusion can be applied at various levels of the processing pipeline, typically categorized into three main strategies: pixel-level fusion, feature-level fusion, and decision-level fusion. Among them, feature-level fusion is increasingly favored for its compatibility with deep learning and its effectiveness in balancing modality-specific and complementary information [8].

Despite the progress achieved, one important aspect remains underexplored: the position at which feature fusion occurs within the network architecture. Specifically, the impact of fusing features at early, middle, or late stages of the backbone on detection performance has not been systematically and comprehensively evaluated. Moreover, while some works have considered multi-scale or hierarchical fusion strategies, there is still a lack of empirical analysis regarding the trade-offs between fusion position, model complexity, and detection accuracy.

In this paper, we aim to fill this gap by conducting a structured and systematic empirical study on feature-level fusion positions for UAV-based RGB-IR object detection. Specifically, we investigate the impact of early, middle, and late feature fusion strategies within a unified dual-stream YOLOv11-based framework [9]. Furthermore, we extend our analysis to multi-position fusion schemes to explore the benefits of hierarchical feature integration. All experiments are conducted on the DroneVehicle dataset, which contains paired RGB-IR UAV images captured under diverse and challenging conditions. Our findings reveal that the fusion position significantly impacts detection accuracy and computational efficiency. Middle fusion achieves the best performance among single fusion points, while early-middle multi-position fusion further enhances localization precision. These insights offer practical guidance for designing more effective and efficient RGB-IR fusion networks in UAV-based object detection applications.

The main contributions of the paper are as follows:

- Dual-stream architecture design: We extend the YOLO framework to support dual-stream RGB-IR input, enabling modality-specific feature extraction and flexible fusion control across the network.

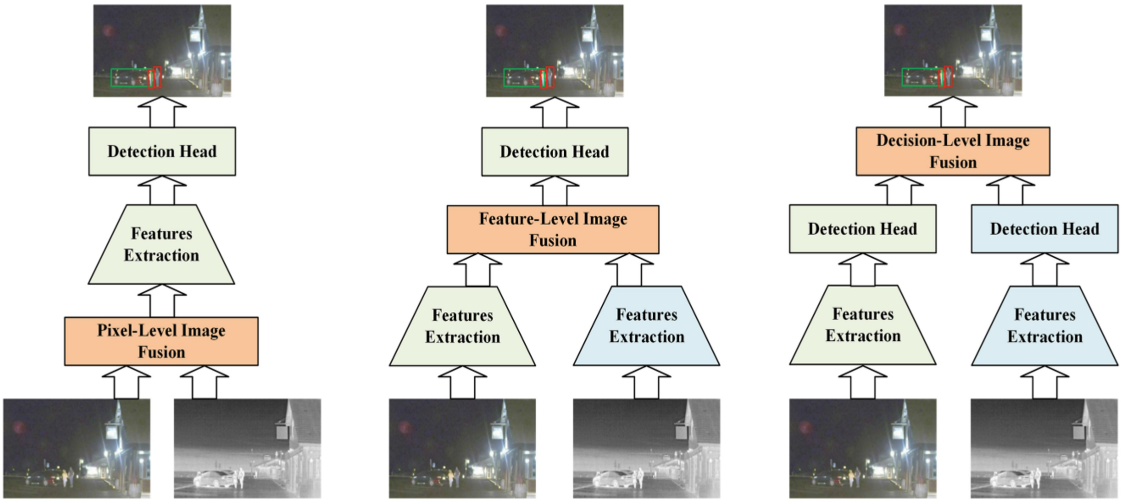
- Comprehensive feature fusion analysis: We conduct a systematic empirical study comparing early-, middle-, late-, and multi-position feature fusion strategies under a unified detection framework.

- Insights and guidance: Our experimental results identify the most effective fusion positions and offer practical guidance for designing robust RGB-IR object detectors for UAV applications.

2. Related work

2.1. RGB-IR image fusion strategies

The fusion of RGB and IR imagery has emerged as a powerful approach to enhance object detection performance, particularly under challenging conditions such as poor lighting, and occlusion. By leveraging the complementary properties of the two modalities-where RGB images provide rich texture and color information while IR images capture thermal signatures and are less sensitive to lighting variations-fusion techniques aim to build more robust and discriminative feature representations. Depending on the stage at which the integration occurs within the processing pipeline, RGB-IR image fusion strategies are generally categorized into three main types: pixel-level fusion, feature-level fusion, and decision-level fusion (Figure 1). Each approach offers distinct advantages and challenges, influencing the final detection performance in different ways. In the following sections, we provide a detailed overview of these fusion strategies [8].



a) Pixel-level fusion b) Feature-level fusion c) Decision-level fusion

Fig. 1. Fusion levels [8].

2.1.1. Pixel-level fusion

Pixel-level fusion is a straightforward strategy that has been explored for integrating RGB-IR images. In this approach, the two modalities are combined directly at the input stage using techniques such as channel concatenation to form a four-channel image [10], weighted averaging [11], or channel substitution [12]. The resulting fused image is then processed by an object detection network (Figure 1a). Although easy to implement, these fusion techniques have been shown to compromise detection performance due to the loss of modality-specific features, particularly in complex scenarios where preserving the complementary characteristics of both RGB and IR data is essential [13].

To address these shortcomings, more advanced fusion approaches have been developed. For instance, Fu *et al.* [14] proposed a neural network that decomposes RGB and IR images into high- and low-frequency components, which are then selectively fused. Zhao *et al.* [15] introduced an auto-encoder-based fusion framework, while Liu *et al.* [16] utilized a generative adversarial network with dedicated discriminators to enhance object information from IR images and texture details from RGB images. Despite their sophistication, these methods are not fully end-to-end, and the generated fused images do not always lead to improved detection accuracy. Moreover, they may suffer from challenges such as image misalignment, incorporation of redundant information, and the introduction of noise, all of which can further degrade performance [8], [10], [17].

2.1.2. Decision-level fusion

Decision-level fusion refers to a strategy in which RGB and IR images are processed independently by separate object detection networks, and their outputs are subsequently combined to produce the final detection results (Figure 1c). This type of fusion takes place after feature extraction and classification, relying on outputs such as bounding boxes and confidence scores. Common fusion techniques include weighted averaging [18], [19], score-based selection [20], and ensemble-based non-maximum suppression [21], [22]. Notably, probabilistic ensembling methods like ProbEn [23] enhance robustness by allowing low-confidence detections from one modality to support and strengthen predictions from the other. By aggregating high-level decisions rather than raw features, decision-level fusion is generally more tolerant of cross-modal misalignment and preserves complementary information across modalities. However, its effectiveness is highly dependent on the accuracy of the individual unimodal detectors, which can be compromised in cases of degraded or poorly aligned inputs. Additionally, despite its simplicity, decision-level fusion may potentially introduce higher inference latency due to the need to run multiple networks in parallel [8], [24], [25].

2.1.3. Feature-level fusion

In feature fusion, a dual-backbone architecture is commonly adopted, where RGB and IR images are independently processed by separate feature extractors. Each branch captures modality-specific representations, such as texture, shape, and edges, to retain the unique characteristics of each input. The extracted features are then merged through a specialized fusion module, allowing the model to leverage complementary information across modalities for improved detection performance (Figure 1b). A key advantage of feature-level fusion is its flexibility: it provides ample room to design customized fusion mechanisms tailored to modality characteristics. These can range from basic operations such as feature concatenation (channel-wise merging) and feature weighting (e.g., weighted summation), to more advanced mechanisms such as attention-based modules or transformer-based cross-modal fusion, which enable both intra- and inter-modal interactions. For instance, guided attention methods can utilize predicted semantic masks to enhance discriminative features, while transformer-based modules employ self-attention to align and merge features across modalities more effectively [8].

Moreover, the fusion process can be implemented at a single layer or across multiple levels of the backbone, allowing hierarchical refinement of shared representations. Single-layer fusion refers to the strategy where modality-specific features are fused only once at a specific point in the network. Importantly, this fusion can occur at different stages of the architecture: early-feature fusion, where low-level features (closer to input) are merged early in the backbone. Mid-feature fusion, where features are fused in intermediate layers. And late-feature fusion, where high-level semantic features are combined after full extraction. In contrast, multi-layer fusion-performs feature integration at multiple levels throughout the network. This approach allows modality interaction at different semantic depths and spatial resolutions [8], [26].

However, this flexibility and increased representational capacity come at a cost. The incorporation of dual-branch backbones and complex fusion modules leads to greater model complexity, increased memory consumption. Despite this, feature-level fusion remains the most widely adopted strategy in RGB-IR object detection, due to its superior ability to learn deep cross-modal correlations and enhance robustness in real-world applications. This strategy has gained considerable attention in multimodal object detection research, primarily due to its balance between flexibility and performance. By preserving modality-specific features during the early stages of processing and enabling interaction at the feature level, feature-level fusion allows for the development of more sophisticated fusion mechanisms. These mechanisms can be tailored to capture complex inter-modal relationships, thereby improving the model's ability to handle diverse environmental conditions such as occlusions and variable lighting-common challenges in UAV-based scenarios [27], [28].

2.2. Deep learning for RGB-IR feature-level fusion in UAV-based object detection

The fusion of RGB and IR modalities has emerged as a powerful strategy in UAV-based object detection, enabling systems to remain robust under challenging conditions such as low-light, occlusion, and environmental noise. Among various fusion strategies, feature-level fusion stands out for its ability to exploit complementary information from both modalities while preserving spatial and semantic coherence [8].

To enhance feature-level fusion in multimodal UAV detection, a variety of deep learning approaches have been developed. Transformer-based models, such as C2Former [29] and CALNet [30], leverage self-attention mechanisms to refine cross-modal alignment and mitigate modality conflicts. Complementarily, methods like AFFCM [31] integrate Softpooling Channel Attention with a Multimodal Adaptive Feature Fusion module to reduce feature redundancy, while CMDistill [32] adopts a cross-modal distillation framework that incorporates semantic relation encoding and IoU-driven optimization to effectively transfer knowledge from infrared to RGB-based detectors. To address spatial misalignment challenges, TSRA [33] aligns translation-sensitive features, while CAGTDet [34] compensates for scale and orientation discrepancies across modalities. In addition, models like IG-GAN [35], dual-attention frameworks [36], and E2E-MFD [37] contribute complementary solutions via generative modeling, hierarchical attention strategies, and streamlined end-to-end fusion architectures, collectively advancing the robustness and precision of RGB-IR object detection systems.

Recent studies have advanced multimodal UAV detection toward more efficient and robust designs. Fusion-DETR [38] integrates an attention-based interaction module with a residual CNN block to enhance intra- and inter-modal alignment, while maintaining real-time performance through an efficient decoding strategy. SAMS-YOLO [39], built on the YOLO framework, combines a Group Shuffled Multi-receptive Attention for better multi-scale representation with a Multi-modal Supervision mechanism to handle annotation misalignment. Both approaches improve detection under challenging conditions such as occlusion and low-light environments. Other methods focus on addressing data imbalance and semantic alignment. EMCFormer [40] uses a Gumbel-softmax-based attention module and Equalized-Adaptive Focal Loss to improve learning from long-tailed UAV datasets. Meanwhile, LPANet [41] leverages large language models to guide progressive semantic and spatial alignment using textual descriptions generated by ChatGPT. Together, these approaches highlight ongoing efforts to improve fusion quality, robustness to real-world challenges, and performance in complex aerial scenarios.

However, while feature-level fusion has shown great potential, most existing studies do not systematically explore the impact of fusion depth or position within the network architecture. Prior works typically apply fusion at fixed stages without thoroughly evaluating how the position of fusion affects final detection performance. This oversight leaves open a key research question regarding the optimal fusion stage for effective multimodal integration in UAV scenarios. The paper aims to address this gap by conducting a systematic empirical analysis of feature-level fusion at different network depths. Specifically, we investigate early, middle, and late feature fusion configurations using a consistent architectural backbone, and analyze how the fusion position influences cross-modal learning and detection accuracy in RGB-IR UAV imagery. Our findings offer practical insights into optimizing fusion design for improved accuracy, robustness, and generalizability in real-world UAV applications.

3. The proposed comparison methods

To investigate the role of feature-level fusion in multispectral UAV-based object detection, we design a set of controlled fusion strategies based on a unified dual-stream YOLOv11 backbone. All configurations integrate modality-specific features from the RGB and IR branches at designated positions within the feature extraction hierarchy. To ensure architectural neutrality and isolate the impact of fusion position, we employ a simple channel-wise concatenation scheme as the sole fusion operation across all configurations. This avoids confounding effects introduced by more complex mechanisms such as weighting, or attention.

We categorize our fusion designs into two groups: Single-level feature fusion, where the integration occurs once at a specific semantic depth, and Multi-position feature fusion, where feature merging is conducted at multiple levels throughout the backbone. This design enables us to systematically examine how the timing and granularity of fusion influence cross-modal representation learning.

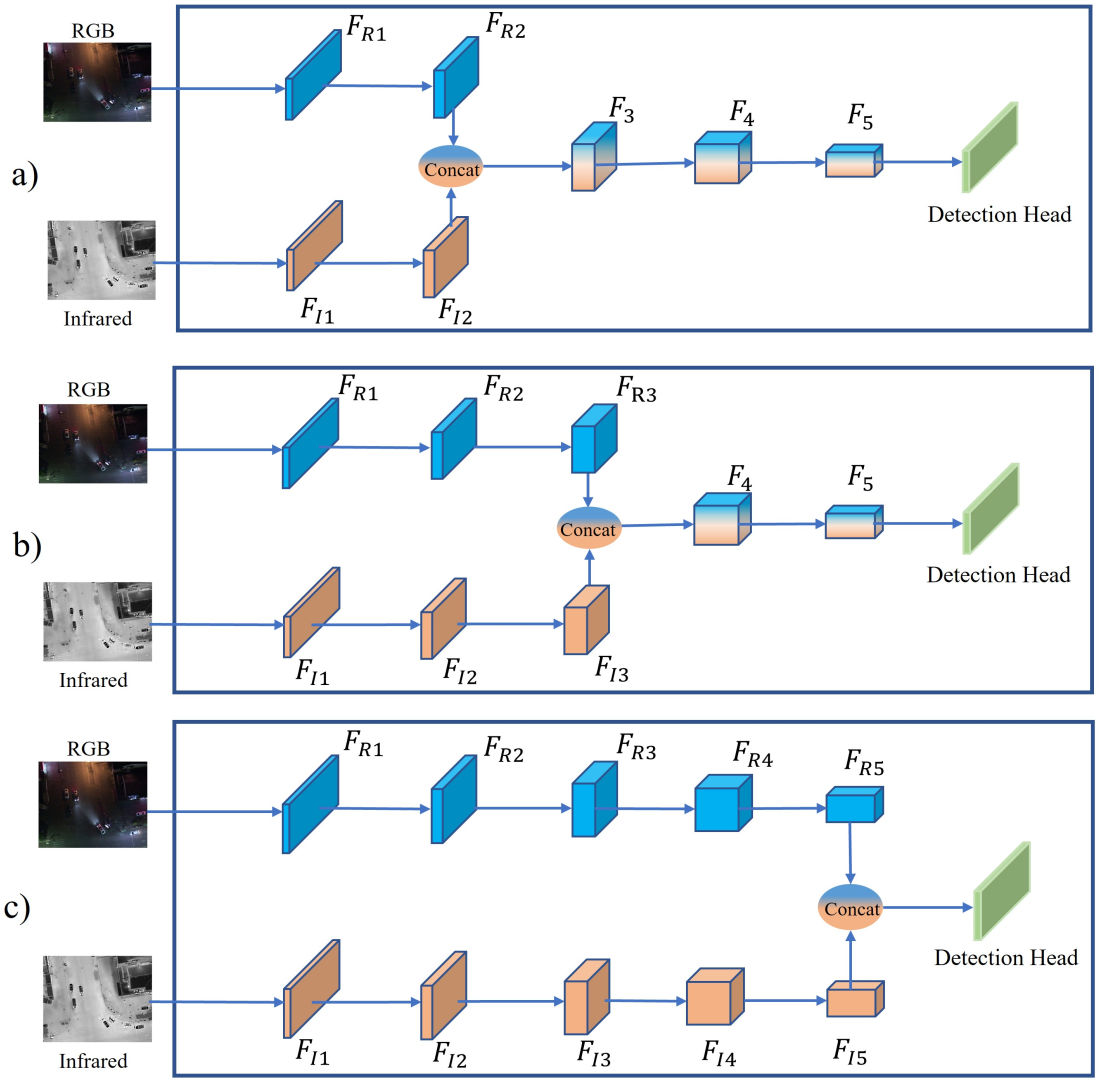
3.1. Single-level feature fusion

In the single-level fusion paradigm, features from the RGB and IR streams are integrated once at a specific stage of the backbone. This controlled setup enables isolation of the effects of fusion depth on detection performance, while maintaining a consistent overall architecture and training protocol. To this end, we define three representative configurations based on the semantic depth at which fusion occurs, as illustrated in Figure 2:

Early fusion: In this configuration, the RGB and IR feature maps are fused shortly after passing through the initial convolutional layers of the backbone, specifically after the second stage. This timing ensures that each modality has already captured fundamental low-level features-such as edges, textures, and local intensity variations-prior to integration. Following the fusion point, the network continues with a unified feature stream, without maintaining separate modality-specific branches.

From a structural perspective, early fusion enables cross-modal interaction to begin at the shallowest level, allowing subsequent layers to learn from jointly processed features throughout the network. This early integration may facilitate spatial alignment across modalities and promote shared representation learning from the outset. However, since fusion occurs prior to the formation of higher-level semantic abstractions, the combined features may still be influenced by modality-specific noise or local inconsistencies. This configuration provides a basis for evaluating how early fusion impacts downstream representation learning, and whether initiating fusion at low abstraction levels supports or hinders the exploitation of cross-modal complementarity.

Middle fusion: In this configuration, RGB and IR features are independently processed through the early layers of the backbone up to a middle stage. Fusion is then performed at this intermediate depth, after which the network transitions to a unified processing stream. The first two stages thus remain modality-specific, allowing the model to extract and retain low-level spectral characteristics before integration. At the fusion point, the feature maps are designed to capture both spatial details and semantic patterns such as object boundaries and contextual information. This representation facilitates effective cross-modal interaction while retaining essential characteristics unique to each modality.



a) Early-feature level fusion b) Middle-feature level fusion c) Late-feature level fusion

Fig. 2. Single-level feature fusion.

From a design standpoint, middle fusion enables a delayed integration strategy that preserves early-stage modality specialization while enabling shared representation learning in deeper layers. This configuration allows investigation into whether fusion at an intermediate depth offers a favorable trade-off between maintaining spectral distinctiveness and enabling joint reasoning. Empirical evaluation of this setup is expected to provide insights into how fusion timing influences the network's capacity to align and leverage complementary information across modalities.

Late fusion: In the late fusion configuration, the RGB and IR branches are processed independently throughout the entire backbone. Fusion is performed only at the final stage of feature extraction, where the modality-specific features are concatenated and passed directly to the detection head. This architecture allows each modality to develop and refine its own high-level representations without mutual interference. At this depth, the resulting feature maps typically encode semantic information such as object categories, contextual relationships, and structural patterns. The purpose of this design is to integrate complementary modality cues at the semantic level, while preserving the distinct processing pipelines in earlier stages. This can be especially beneficial in cases where spatial misalignment or modality-specific noise is present.

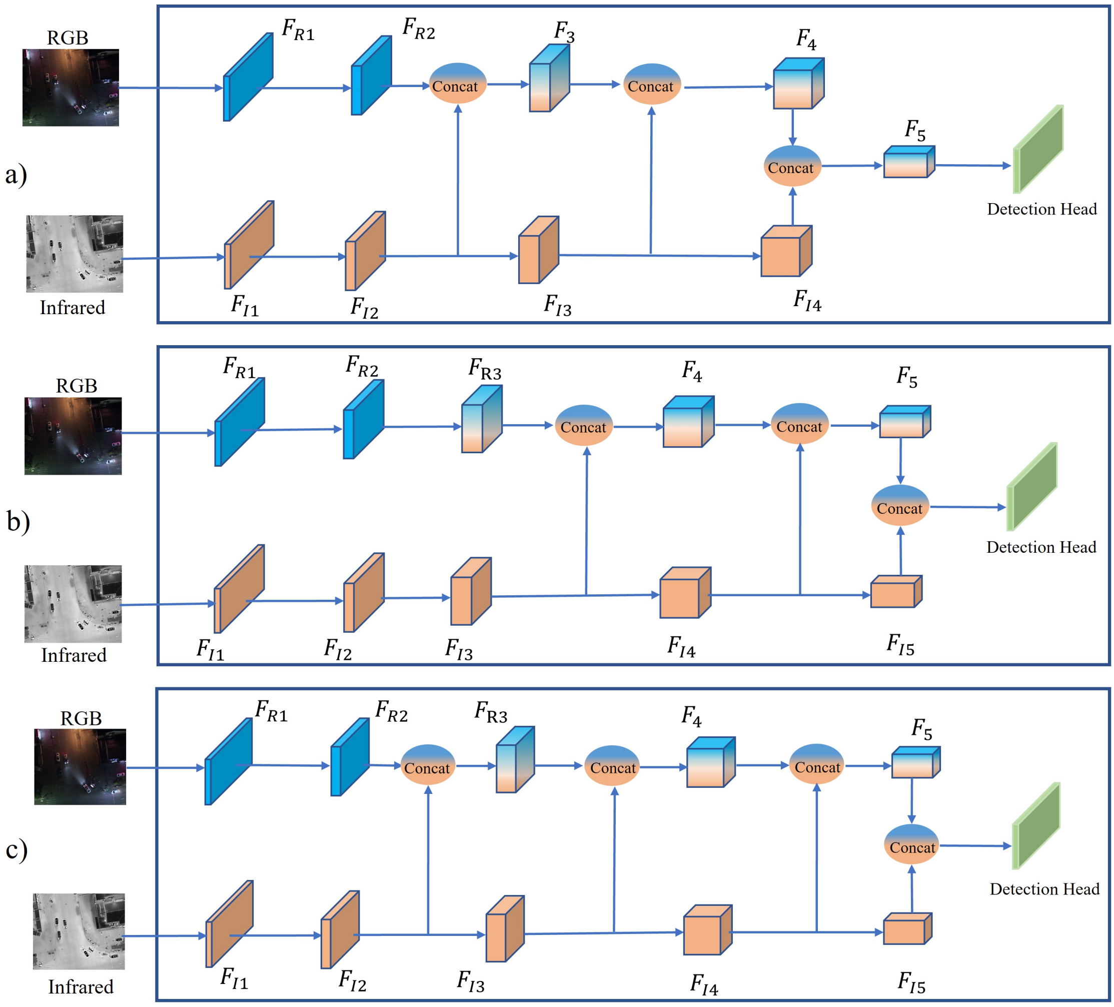
Because the fusion is applied only at the final stage, the network lacks additional layers to jointly refine or align the merged features. As a result, the ability to exploit cross-modal interactions is limited to the detection head. This configuration is therefore useful for examining whether semantic-level fusion alone is sufficient to support effective multispectral object detection. It also provides a reference point for comparing the impact of fusion timing on detection accuracy and robustness in UAV-based RGB-IR scenarios.

3.2. Multi-position feature fusion

To explore the effects of distributed fusion across the network hierarchy, we design a series of multi-position feature fusion configurations. These configurations allow cross-modal interaction to occur at multiple semantic depths, leveraging the hierarchical structure of deep convolutional backbones. By integrating RGB and IR features at more than one stage within the feature extraction pipeline, the network may capture complementary information at varying levels of abstraction and spatial resolution. All multi-position fusion schemes use channel-wise concatenation as the fusion operation, without introducing additional parameters. The architecture, training settings, and detection head remain fixed across all configurations to ensure that differences in performance can be attributed solely to fusion positioning.

Three representative configurations are considered, as illustrated in Figure 3:

Early-middle fusion: In this configuration, RGB and IR features are integrated at both shallow and intermediate stages of the feature extraction process. Following the final fusion point, the two modality-specific branches are unified into a single feature stream, and all subsequent processing is carried out jointly. This setup enables progressive interaction between modalities, starting from low-level spatial alignment and extending toward more semantically meaningful representations. By concentrating fusion in the earlier half of the network and unifying the streams before high-level processing, the configuration establishes a clear transition point for shared feature reasoning in the deeper layers.



a) Early and middle feature level fusion b) Middle and late feature level fusion
c) Early, middle and late feature level fusion

Fig. 3. Multi-position feature fusion.

Middle-late fusion: In this configuration, cross-modal fusion is introduced only at intermediate and deep stages of the network. The RGB and IR streams are processed independently through the early layers, allowing each modality to retain its unique characteristics before any interaction occurs. Fusion is then applied progressively in the latter part of the backbone, with features from both modalities concatenated at each selected stage while still maintaining dual-stream processing until the final fusion point. This design emphasizes semantic-level integration, enabling joint reasoning over more abstracted representations. By deferring fusion, the configuration may reduce the risk of introducing low-level noise or misalignment that can arise when modalities are combined too early. It also provides more time for each stream to develop modality-specific features before being aligned.

Early-middle-late fusion: This configuration introduces cross-modal fusion at multiple levels across the entire feature extraction hierarchy, including shallow, intermediate, and deep stages. At each selected point, features from the RGB and IR branches are concatenated, while the dual-stream structure is preserved between fusion steps and only unified near the end of the backbone. By distributing fusion throughout the network, this design maximizes the frequency and depth of cross-modal interaction, potentially supporting gradual alignment and refinement of shared representations. It allows the network to leverage complementary information across varying semantic levels, from low-level spatial patterns to high-level contextual features. However, the increased number of fusion operations also adds to the architectural complexity and computational cost.

4. Experimental settings

4.1. Dual-stream unified backbone

In this study, we adopt YOLOv11 [9] as a dual-stream unified backbone to investigate the effectiveness of different fusion positions in the feature-level fusion pipeline for dual-modal object detection using RGB and IR images. Although a newer version, YOLOv12, has recently been released, YOLOv11 remains the preferred choice due to its greater stability and superior performance across various scenarios-particularly where real-time inference and model reliability are crucial.

YOLOv11 represents a significant advancement in the YOLO series, incorporating architectural innovations such as the C3k2 block (an efficient variant of the CSP bottleneck), SPPF (Spatial Pyramid Pooling - Fast), and the C2PSA (Cross Stage Partial with Spatial Attention) module. These components collectively enhance the model's capacity to extract salient features and focus attention on critical regions of the input image. Despite its lightweight design, YOLOv11 maintains high performance and supports diverse computer vision tasks, including object detection, instance segmentation, pose estimation, and oriented object detection [42].

To adapt YOLOv11 for the dual-modal scenario, we extend the original single-modality design into a dual-stream structure, processing RGB and IR inputs in parallel. This unified backbone enables a systematic investigation of fusion positions within the feature extraction pipeline. As our primary objective is to understand where fusion should occur-rather than how-we employ a straightforward feature concatenation strategy for combining modality-specific representations. This experimental setup ensures a fair comparison across different configurations while preserving the original model's efficiency and representational strength. To isolate the effect of fusion depth, we keep all training settings and architectural components identical across experiments, altering only the fusion position. This control ensures that any observed performance differences stem solely from the fusion strategy rather than confounding factors such as network capacity, or data preprocessing.

In summary, the YOLOv11-based dual-stream unified backbone offers a robust and efficient foundation for our experimental framework. Its strong balance of accuracy, speed, and architectural clarity makes it particularly well-suited for analyzing the impact of fusion location in dual-modal object detection systems.

4.2. Dataset

To evaluate the performance of fusion strategies in a unified and realistic experimental setup, we employ the DroneVehicle dataset [43], a comprehensive UAV-based benchmark specifically designed for multimodal object detection using RGB and IR imagery. Released in 2022, the dataset consists of 56,878 images captured by UAVs, equally divided between RGB and IR modalities. It covers a wide range of real-world conditions, including diverse illumination scenarios, viewing angles, and flight altitudes, making it highly suitable for testing the robustness of fusion-based detection methods in practical applications. More specifically, the dataset comprises 15,475 vertical-view and 12,964 oblique-view RGB-IR image pairs. Oblique images are acquired from angles of 15°, 30°, and 45°, at flight altitudes of 80m, 100m, and 120m, offering a variety of geometric perspectives. Each image is annotated with bounding boxes across five vehicle categories-car, truck, bus, van, and freight car-which enables detailed and category-specific detection evaluation.

Prior to annotation, the raw image data collected by UAVs undergoes several preprocessing steps to ensure quality and consistency. Initially, low-quality images-such as those with severe motion blur-are discarded. The remaining images are manually reviewed, and their resolutions are uniformly resized to 840 × 712 pixels. After this cleaning step, distortion correction is applied to all retained images. Due to the inherent instability in UAV flight posture during data collection, cross-modal image pairs captured by the RGB and IR cameras often exhibit pixel-level misalignment. To address this, an affine transformation is applied, followed by region cropping, during the calibration phase. These operations help minimize spatial discrepancies and ensure that the majority of RGB-IR image pairs are geometrically aligned, facilitating reliable multimodal object detection in subsequent stages.

4.3. Metrics

In this study, we evaluate the performance of multi-modal object detection using the COCO-style mean Average Precision (mAP) metrics. Specifically, we report mAP50 and mAP50:95, which represent detection accuracy at a single IoU threshold and across a range of IoU thresholds, respectively. To evaluate model complexity and deployment feasibility, we also report the number of parameters (Params), and the inference speed in frames per second (FPS) using 640 × 640 input resolution. FPS is computed by averaging the per-image inference time over multiple runs after warm-up on a single NVIDIA GPU. These combined metrics provide a comprehensive view of each model's accuracy and real-time applicability. They also allow fair comparison of architectures by capturing both accuracy and efficiency-crucial for UAV applications where accuracy-speed trade-offs matter.

4.4. Implementation details

The experiments were conducted on a workstation equipped with an NVIDIA Tesla P40 GPU, using CUDA version 12.2 and driver version 535.183.01 to ensure efficient GPU acceleration. All models were implemented based on a customized version of YOLOv11. During training, we adopted a dual-branch feature extraction framework, with all input images resized to 640 × 640 pixels for consistency. Mosaic augmentation was applied during preprocessing to enhance data diversity and improve detection performance. The models were trained for 50 epochs using the stochastic gradient descent optimizer with a momentum of 0.937 and a weight decay of 0.0005. The learning rate was set to 0.01 with a fixed batch size of 4.

5. Results

5.1. Comparison of single-level feature fusion strategies

Table 1 summarizes the results of three single-fusion configurations corresponding to early-, middle-, and late-feature fusion positions. Among these, the middle fusion configuration achieved the highest overall detection performance, with an mAP50 of 82.9% and an mAP50:95 of 61.6%. It also recorded the highest precision (79.3%) and recall (78.8%), suggesting an effective balance between reducing false positives and capturing true detections. The early fusion setup yielded competitive results, with an mAP50 of 81.6%, mAP50:95 of 61.4%, precision of 78.2%, and recall of 77.7%. Notably, it achieved the fastest inference speed (58.5 FPS) and the smallest model size (19.2 M parameters), indicating computational efficiency despite slightly lower accuracy compared to middle fusion. In contrast, the late fusion configuration resulted in the lowest detection accuracy (mAP50: 78.7%, mAP50:95: 58.4%) and precision (75.6%), while recall (78.7%) remained comparable to other setups. It also incurred the highest computational cost, with the largest parameter count (26.4 M) and the slowest inference speed (51.0 FPS).

Table 1. Comparison of single-level fusion positions in the feature extraction pipeline

Position	**Params**	**Precision**	**Recall**	**mAP50 (%)**	**mAP50:95 (%)**	**FPS**
Early	19.2M	78.2	77.7	81.6	61.4	**58.5**
Middle	20.8M	79.3	78.8	**82.9**	**61.6**	53.8
Late	26.4M	75.6	78.7	78.7	58.4	51.0

An analysis of these results reveals several key observations:

- Early fusion integrates RGB and IR features at a stage dominated by low-level spatial and structural information. This early cross-modal interaction enables the network to progressively refine shared representations through deeper layers, enhancing feature alignment across modalities. Furthermore, the compact architecture resulting from early fusion contributes to higher computational efficiency and faster inference. However, because low-level features are less semantically rich, early fusion may limit the exploitation of higher-order inter-modal relationships, potentially explaining its slightly lower accuracy compared to middle fusion.

- Middle fusion introduces cross-modal interaction at an intermediate stage where modality-specific representations have developed sufficiently complex semantic abstractions but have not yet become fully specialized. This stage appears to offer a favorable trade-off between intra-modal feature learning and inter-modal integration, enabling the network to better exploit complementary information from both modalities. Consequently, middle fusion produced the highest overall detection accuracy in our experiments.

- Late fusion delays the integration of RGB and IR features until the final stages of the backbone, at which point each modality has already undergone deep, independent feature extraction. As a result, the learned representations from the two streams are highly modality-specific and semantically divergent. This separation limits the network's ability to exploit cross-modal complementarity, as much of the lower- and mid-level structural and contextual information-where RGB and IR could support and enhance one another-has already been processed in isolation.

Moreover, because fusion occurs near the detection head, there is little to no opportunity for the network to refine or align the combined features through shared downstream layers. This late-stage merging leads to suboptimal feature interaction and hinders the model's ability to resolve inconsistencies between modalities, such as spatial misalignments or differing contrast patterns. Consequently, despite having the highest model complexity among the three configurations, late fusion fails to deliver proportional improvements in accuracy. Instead, it exhibits the lowest detection performance (mAP50 of 78.7%), suggesting that deferring fusion too far in the network sacrifices early integration benefits while introducing unnecessary computational overhead.

Overall, these findings suggest that the timing of cross-modal fusion plays a crucial role in multispectral object detection. Introducing fusion at earlier or intermediate feature stages is more beneficial than fusing at deeper stages, both in terms of detection accuracy and computational efficiency. Future network designs for RGB-IR fusion could further benefit from carefully selecting the integration depth based on the specific application requirements and resource constraints.

5.2. Comparison of multi-position feature fusion strategies

Table 2 shows that the early-middle fusion strategy achieved the highest mAP50 (82.7%) and mAP50:95 (62.3%), with a moderate model size (25.4 M parameters) and a relatively high inference speed (48.5 FPS). The middle-late fusion configuration demonstrated a comparable mAP50 (82.4%) and the highest recall (78.9%), while achieving the fastest inference speed (48.8 FPS), albeit with a larger parameter count (30.0 M). Meanwhile, the early-middle-late fusion, despite involving multiple fusion points, resulted in no additional performance gains. Instead, it slightly reduced the mAP50 to 82.0%, incurred the highest model complexity (30.3 M parameters), and led to the lowest inference speed (47.2 FPS).

Table 2. Comparison of multi-position fusion in the feature extraction pipeline

Position	**Params**	**Precision**	**Recall**	**mAP50 (%)**	**mAP50:95 (%)**	**FPS**
Early-middle	25.4M	79.6	78.6	**82.7**	**62.3**	48.5
Middle-late	30.0M	78.6	78.9	82.4	62.1	**48.8**
Early-middle-late	30.3M	78.5	78.4	82.0	62.0	47.2

These findings suggest that while multi-level fusion can benefit detection performance by integrating features across different semantic stages, an excessive number of fusion points may introduce redundancy and computational overhead without yielding further accuracy improvements. In particular, early-middle fusion strikes a favorable balance between detection accuracy, model size, and inference efficiency, making it a practical choice for real-time UAV object detection applications.

5.3. Comparison between single-level and multi-position feature fusion

By comparing the results in Table 1 and Table 2, we can observe distinct trade-offs between single-level and multi-level fusion strategies. Single-level fusion at the middle stage achieved the highest mAP50 (82.9%) overall, indicating superior localization performance for objects when evaluated with a loose IoU threshold. In contrast, multi-level fusion configurations produced slightly lower mAP50 scores (82.7% for early-middle fusion and 82.4% for middle-late fusion), suggesting a minor reduction in coarse detection accuracy.

However, when considering the stricter mAP50:95 metric, which emphasizes precise localization across a range of IoU thresholds, multi-level fusion methods consistently outperformed single-level fusion. Specifically, early-middle fusion achieved the highest mAP50:95 (62.3%), compared to 61.6% in the best single-level setting (middle fusion). This indicates that multi-level fusion enhances the network's ability to accurately align and delineate object boundaries, likely due to the richer and more complementary feature interactions across different semantic stages. In terms of computational efficiency, early single-level fusion offered the highest inference speed (58.5 FPS), while multi-level fusion strategies, although slower (around 48.5-48.8 FPS), still maintained acceptable real-time performance for UAV applications.

These findings suggest that the choice between single-level and multi-level fusion should be guided by the specific requirements of the target application. For scenarios prioritizing high-speed detection with moderate precision, such as real-time UAV surveillance, single-level early or middle fusion would be preferable. Conversely, for applications requiring higher localization precision under strict evaluation conditions, such as UAV-based search and rescue or infrastructure inspection, multi-level fusion may provide better detection reliability, despite a slight sacrifice in inference speed.

5.4. Comparison with single-modality baselines

To further demonstrate the effectiveness of RGB-IR feature fusion, we compare our dual-modal model with two single-modality baselines that process either RGB or IR inputs independently, both using the same YOLOv11 backbone. All models are trained under identical conditions to ensure a fair and meaningful comparison. As presented in Table 3, the proposed dual-stream model achieves superior detection performance, with mAP50 reaching 82.9% and mAP50:95 reaching 61.6%, outperforming both unimodal variants by a noticeable margin.

Table 3. Comparison with single-modality baselines

Base	Input Type	Precision	Recall	mAP50 (%)	mAP50:95 (%)
YOLOv11	RGB Only	74.4	72.0	76.2	54.9
YOLOv11	IR Only	75.6	76.6	79.4	59.2
Dual-YOLO11	RGB+IR (middle)	79.3	78.8	82.9	61.6

The RGB-only baseline benefits from rich texture and color details, which are advantageous under well-lit conditions. However, its performance tends to degrade in challenging lighting environments such as nighttime. Conversely, the IR-only baseline maintains greater robustness in low-visibility scenarios but suffers from limited spatial detail and object texture, which affects overall precision. By effectively combining complementary features from both modalities, the fusion-based approach significantly enhances the model's ability to detect objects accurately and consistently across diverse UAV imagery conditions. These results highlight the robustness and practicality of our proposed RGB-IR fusion strategy for real-world UAV-based object detection tasks.

5.5. Comparison with state-of-the-art fusion techniques

Recent studies have proposed various sophisticated fusion strategies to address challenges in RGB-IR object detection, including modality inconsistency, semantic conflict, and misalignment. For example, He *et al*. proposed CALNet [30] with Cross-Modal Conflict Rectification and Selected Cross-Modal Fusion modules, achieving 75.4% mAP50 on the DroneVehicle dataset. Yuan and Wei introduced C2Former [29], which leverages Inter-modality Cross-Attention and Adaptive Feature Sampling modules, reaching 74.2% mAP50. Similarly, Yuan *et al*. developed CAGTDet [34] featuring Translation-Scale-Rotation Alignment and Complementary Fusion Transformer modules, resulting in 74.6% mAP50. DDCINet [44] achieved 78.4% mAP50 through dynamic cross-modal feature interaction.

Although these methods utilize complex fusion modules, our experiments reveal that a simple feature concatenation strategy combined with an adapted strong backbone achieves higher performance (Table 4). Specifically, all fusion levels in our method, even when relying solely on straightforward concatenation, consistently surpass the mAP50 results of the aforementioned methods, achieving over 82% mAP50 on the DroneVehicle dataset. This remarkable performance can be attributed to several factors. First, although we use simple concatenation, it is integrated into the network at an appropriately selected stage, enabling effective combination of modality-specific information while preserving critical spatial details. Second, the adapted YOLOv11 backbone, which we modify to simultaneously process dual input streams (RGB and IR), provides strong feature extraction capabilities that facilitate effective multimodal learning even without complex fusion operations. Third, while complex fusion modules aim to model intricate cross-modal interactions, they may inadvertently introduce noise, redundancy, or optimization difficulties, especially in UAV scenarios with weakly aligned modalities. In contrast, a simpler concatenation approach avoids unnecessary complexity, allowing the network to learn modality interactions more naturally through end-to-end training.

Table 4. Comparison of state-of-the-art fusion techniques on the DroneVehicle dataset

Models	**Modality**	**Car**	**Truck**	**Bus**	**Freight**	**Van**	**mAP50(%)**
YOLO11 (RGB only) [9]	RGB	97.0	75.8	95.3	54.3	58.5	76.2
YOLO11 (IR only) [9]	IR	98.4	78.2	95.0	66.0	59.2	79.4
CALNet (2023) [30]	RGB+IR	90.3	76.2	89.1	63.0	58.5	75.4
C2Former (2024) [29]	RGB+IR	90.2	68.3	89.8	64.4	58.5	74.2
CAGTDet (2024) [34]	RGB+IR	90.8	69.7	90.5	66.3	55.6	74.6
DDCINet (2025) [44]	RGB+IR	91.0	78.9	90.7	66.1	**65.5**	78.4
Early fusion (Ours)	RGB+IR	**98.6**	81.8	96.4	69.2	61.8	81.6
Middle fusion (Ours)	RGB+IR	**98.6**	**83.3**	96.6	**72.3**	63.6	**82.9**
Early-middle fusion (Ours)	RGB+IR	**98.6**	82.9	**96.7**	71.2	64.1	82.7

Beyond achieving the highest mAP50 across all evaluated categories, the proposed dual-stream RGB-IR fusion models exhibit strong robustness under diverse environmental conditions. As shown in Table 4, models utilizing both RGB and IR modalities consistently outperform their single-modality counterparts. In favorable lighting conditions, RGB-only models benefit from rich texture and color details, but their performance drops in low-visibility scenes. Conversely, IR-only models retain reasonable detection capability in poorly lit environments by leveraging thermal signatures, yet lack spatial detail in complex scenes. The proposed fusion models effectively combine these complementary strengths.

Although the proposed fusion model achieves high overall performance with an mAP50 of up to 82.9%, its accuracy across object categories is not entirely uniform. Notably, the Van class records a relatively low mAP50 of 63.6%, which is significantly lower than other categories such as Car or Truck. This suggests that the model has not yet fully leveraged the complementary strengths of RGB and IR modalities in detecting objects that are small, have varying shapes, or are frequently occluded. These characteristics make such objects more challenging to identify, especially in complex aerial scenes.

It is important to highlight that the main objective of the study is to investigate where feature fusion should occur within the network, rather than to develop a new fusion mechanism. Therefore, while the current results validate the effectiveness of the chosen fusion position, future research will aim to improve the fusion strategy itself. Specifically, enhancing the model's ability to capture difficult object classes-like Van-will be a key focus, especially for real-world UAV applications where these targets are operationally important but easily missed.

6. Conclusion and future work

The study analyzed the impact of fusion depth and configuration on UAV-based object detection using RGB and IR images. Our experiments showed that middle fusion outperforms early and late fusion in detection accuracy, precision, and recall. Early fusion, while slightly less accurate, provided faster inference and smaller model size, making it suitable for real-time applications. Multi-position fusion strategies, particularly early-middle fusion, struck the best balance between accuracy, model size, and inference speed. However, the early-middle-late fusion configuration offered no additional performance gains and increased computational costs. These findings highlight the importance of selecting an appropriate fusion depth to balance detection performance and computational efficiency in UAV-based object detection systems.

Future research could explore the integration of advanced fusion techniques, such as attention mechanisms or multi-modal transformers, to further improve feature alignment and performance. Additionally, investigating adaptive fusion strategies that dynamically adjust based on input conditions or task requirements could provide more efficient solutions for UAV-based detection systems.

References

[1] N. Merkle, R. Bahmanyar, C. Henry, S. M. Azimi, X. Yuan, S. Schopferer, V. Gstaiger, S. Auer, A. Schneibel, M. Wieland, and T. Kraft, "Drones4Good: Supporting disaster relief through remote sensing and AI," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023, pp. 3772–3776. DOI: 10.48 550/arXiv.2308.05 074

[2] P. W. Patil, A. Dudhane, S. Chaudhary, and S. Murala, "Multi-frame based adversarial learning approach for video surveillance," *Pattern Recognition*, vol. 122, 2022. DOI: 10.1016/j.patcog.2021.108350

[3] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2021. DOI: 10.1109/MGRS.2021.3115137

[4] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, "HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection," *Scientific Data*, vol. 10, no. 1, p. 227, 2023. DOI: 10.1038/s41597-023-02066-6

[5] H. Wang, C. Wang, Q. Fu, D. Zhang, R. Kou, Y. Yu, and J. Song, "Cross-modal oriented object detection of UAV aerial images based on image feature," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–21, 2024. DOI: 10.1109/TGRS.2024.3367934

[6] Y. Xiao, F. Meng, Q. Wu, L. Xu, M. He, and H. Li, “GM-DETR: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 5541–5549.

[7] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, “A review of multimodal medical image fusion techniques,” *Computational and Mathematical Methods in Medicine*, vol. 2020, no. 1, 2020. DOI: 10.1155/2020/8279342

[8] Y. Sun, Y. Meng, Q. Wang, M. Tang, T. Shen, and Q. Wang, “Visible and infrared image fusion for object detection: a survey,” in *International Conference on Image, Vision and Intelligent*, pp. 236–248, 2023. DOI: 10.1117/12.3011457

[9] Ultralytics, “YOLO11 official release,” 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[10] J. Wagner, V. Fischer, M. Herman, and S. Behnke, “Multispectral pedestrian detection using deep fusion convolutional neural networks.” in *European Symposium on Artificial Neural Networks*, vol. 587, 2016, pp. 509–514.

[11] G. French, G. Finlayson, and M. Mackiewicz, “Multi-spectral pedestrian detection via image fusion and deep neural networks,” *Journal of Imaging Science and Technology*, pp. 176–181, 2018. DOI: 10.2352/J.lmagingSci.Technol.2018.62.5.050406

[12] M. Vandersteegen, K. Van Beeck, and T. Goedemé, “Real-time multispectral pedestrian detection with a single-pass deep neural network,” in *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15.* Springer, pp. 419–426, 2018. DOI: 10.1007/978-3-319-93000-8_47

[13] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, “Multispectral deep neural networks for pedestrian detection,” *arXiv preprint arXiv:1611.02644*, 2016. DOI: 10.48550/arXiv.1611.02644

[14] Y. Fu, X.-J. Wu, and J. Kittler, “A deep decomposition network for image processing: A case study for visible and infrared image fusion,” *arXiv preprint arXiv:2102.10526*, 2021. DOI: 10.48550/arXiv.2102.10526

[15] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, “DIDFuse: Deep image decomposition for infrared and visible image fusion,” *arXiv preprint arXiv:2003.09210*, 2020. DOI: 10.24963/ijcai.2020/135

[16] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, “Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811. DOI: 10.48550/arXiv.2203.16220

[17] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, “Weakly aligned cross-modal learning for multispectral pedestrian detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5127–5137. DOI: 10.1109/ICCV.2019.00523

[18] Y. Zhuang, Z. Pu, J. Hu, and Y. Wang, “Illumination and temperature-aware multispectral networks for edge-computing-enabled pedestrian detection,” *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1282–1295, 2022. DOI: 10.1109/TNSE.2021.3139335

[19] Q. Li, C. Zhang, Q. Hu, H. Fu, and P. Zhu, “Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection,” *IEEE Transactions on Multimedia*, vol. 25, pp. 3420–3431, 2022. DOI: 10.1109/TMM.2022.3160589

[20] Z. Hu, Y. Jing, and G. Wu, “Decision-level fusion detection method of visible and infrared images under low light conditions,” *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, 2023. DOI: 10.1186/s13634-023-01002-5

[21] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, “Soft-NMS–improving object detection with one line of code,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5561–5569. DOI: 10.48550/arXiv.1704.04503

[22] R. Solovyev, W. Wang, and T. Gabruseva, “Weighted boxes fusion: Ensembling boxes from different object detection models,” *Image and Vision Computing*, vol. 107, 2021. DOI: 10.1016/j.imavis.2021.104117

[23] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, “Multimodal object detection via probabilistic ensembling,” in *European Conference on Computer Vision.* Springer, 2022, pp. 139–158. DOI: 10.1007/978-3-031-20077-9_9

[24] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, “Multimodal object detection by channel switching and spatial attention,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 403–411. DOI: 10.1109/CVPRW59228.2023.00046

[25] M. He, Q. Wu, K. N. Ngan, F. Jiang, F. Meng, and L. Xu, “Misaligned RGB-infrared object detection via adaptive dual-discrepancy calibration,” *Remote Sensing*, vol. 15, no. 19, 2023. DOI: 10.3390/rs15194887

[26] P. V. Borges, T. Peynot, Liang *et al.*, “A survey on terrain traversability analysis for autonomous ground vehicles: Methods, sensors, and challenges,” *Field Robotics*, vol. 2, pp. 1567–1627, 2022. DOI: 10.55417/fr.2022049

[27] Y. Luo and Z. Luo, "Infrared and visible image fusion: Methods, datasets, applications, and prospects," *Applied Sciences*, vol. 13, no. 19, 2023. DOI: 10.3390/app131910891

[28] K. Song, Y. Zhao, L. Huang, Y. Yan, and Q. Meng, "RGB-T image analysis technology and application: A survey," *Engineering Applications of Artificial Intelligence*, vol. 120, 2023. DOI: 10.1016/j.engappai.2023.105919

[29] M. Yuan and X. Wei, "C2former: Calibrated and complementary transformer for RGB-infrared object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024. DOI: 10.1109/TGRS.2024.3376819

[30] X. He, C. Tang, X. Zou, and W. Zhang, "Multispectral object detection via cross-modal conflict-aware learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1465–1474. DOI: 10.1145/3581783.3612651

[31] Y. Wu, X. Guan, B. Zhao, L. Ni, and M. Huang, "Vehicle detection based on adaptive multimodal feature fusion and cross-modal vehicle index using RGB-T images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 8166–8177, 2023. DOI: 10.1109/JSTARS.2023.3294624

[32] X. Tong, X. Guo, X. Sun, R. Guo, S. Su, and Z. Zuo, "CMDistill: Cross-modal distillation framework for UAV image object detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 1395–1409, 2024. DOI: 10.1109/JSTARS.2024.3479717

[33] M. Yuan, Y. Wang, and X. Wei, "Translation, scale and rotation: cross-modal alignment meets RGB-infrared vehicle detection," in *European Conference on Computer Vision*. Springer, pp. 509–525, 2022. DOI: 10.48550/arXiv.2209.13801

[34] M. Yuan, X. Shi, N. Wang, Y. Wang, and X. Wei, "Improving RGB-infrared object detection with cascade alignment-guided transformer," *Information Fusion*, vol. 105, 2024. DOI: 10.1016/j.inffus.2024.102246

[35] C. Sui, G. Yang, D. Hong, H. Wang, J. Yao, P. M. Atkinson, and P. Ghamisi, "IG-GAN: Interactive guided generative adversarial networks for multimodal image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–19, 2024. DOI: 10.1109/TGRS.2024.3433619

[36] Y. Hu, L. Shi, L. Yao, and L. Weng, "Dual attention feature fusion for visible-infrared object detection," in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 53–65. DOI: 10.1007/978-3-031-44195-0_5

[37] J. Zhang, M. Cao, W. Xie, J. Lei, D. Li, W. Huang, Y. Li, and X. Yang, "E2E-MFD: Towards end-to-end synchronous multimodal fusion detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 52 296–52 322, 2024. DOI: 10.48550/arXiv.2403.09323

[38] X. Huang and G. Ma, "Cross-modality object detection based on DETR," *IEEE Access*, vol. 13, pp. 51 220–51 230, 2025. DOI: 10.1109/ACCESS.2025.3551947

[39] J. Wang, X. Tian, S. Dai, T. Zhuo, H. Zeng, H. Liu, J. Liu, X. Zhang, and Y. Zhang, "RGB-T object detection via group shuffled multi-receptive attention and multi-modal supervision," in *International Conference on Pattern Recognition*. Springer, 2025, pp. 284–298. DOI: 10.1007/978-3-031-78447-7 19

[40] Z. Wang, X. Liao, J. Yuan, C. Lu, and Z. Li, "EMCFormer: Equalized multi-modal cues fusion transformer for remote sensing visible-infrared object detection under long-tailed distribution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 9533–9545, 2025. DOI: 10.1109/JSTARS.2025.3553747

[41] W. Wu, C. Li, X. Wang, B. Luo, and Q. Liu, "Large language model guided progressive feature alignment for multimodal UAV object detection," *arXiv preprint arXiv:2503.06948*, 2025. DOI: 10.48550/arXiv.2503.06948

[42] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024. DOI: 10.48550/arXiv.2410.17725

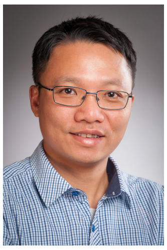
[43] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022. DOI: 10.1109/TCSVT.2022.3168279

[44] W. Bao, M. Huang, J. Hu, and X. Xiang, "Dual dynamic cross-modal interaction network for multimodal remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025. DOI: 10.1109/TGRS.2025.3530085

Manuscript received 19-03-2025; Accepted 15-06-2025.

■

Thi Lan Nguyen received her Bachelor’s degree in Computer Science in 2013 from Le Quy Don Technical University. In 2021, she completed her Master’s degree in Information Systems at the same university, gaining further experience in software systems, data processing, and information management. She is currently pursuing a Ph.D. in Mathematical Foundations for Information Technology at the Institute of Information and Communication Technology, Le Quy Don Technical University. Her research interests include artificial intelligence and computer vision, particularly focusing on deep learning methods and their potential applications in image understanding and intelligent automation.
Email: lannt.simtech@lqdtu.edu.vn



Cao Truong Tran received the PhD degree in computer science from Victoria University of Wellington, New Zealand. He also did postdoc at Victoria University of Wellington. He is researching in the field of machine learning and evolutionary computation, specialized with evolutionary machine learning for data mining with missing data. He serves as a reviewer of international journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Cybernetics, Pattern Recognition, Knowledge-Based Systems, Applied Soft Computing and Engineering Application of Artificial Intelligence. He is also a PC member of international conferences, including IEEE Congress on Evolutionary Computation, IEEE Symposium Series on Computational Intelligence, the Australasian Joint Conference on Artificial Intelligence, and the AAAI Conference on Artificial Intelligence.
Email: truongct@lqdtu.edu.vn

PHÂN TÍCH THỰC NGHIỆM VIỆC KẾT HỢP ĐẶC TRƯNG CỦA ẢNH RGB-IR TRONG PHÁT HIỆN ĐỐI TƯỢNG TRÊN ẢNH UAV

Nguyễn Thị Lan, Trần Cao Trưởng

Tóm tắt

Phát hiện đối tượng dựa trên UAV đóng vai trò quan trọng trong nhiều ứng dụng như giám sát, quản lý thiên tai và hoạt động quân sự. Tuy nhiên, các phương pháp truyền thống thường gặp khó khăn trong điều kiện ánh sáng kém hoặc khi đối tượng bị che khuất. Để khắc phục những hạn chế này, các nghiên cứu gần đây đã tập trung vào việc kết hợp giữa ảnh nhìn thấy (RGB) và ảnh hồng ngoại (IR) nhằm khai thác các ưu điểm bổ sung của hai loại dữ liệu này. Trong số các chiến lược kết hợp ảnh hiện có, kết hợp ở cấp độ đặc trưng (*feature-level fusion*) nhận được nhiều sự quan tâm nhờ tính linh hoạt trong việc lựa chọn vị trí kết hợp và khả năng tận dụng các mô hình học sâu, dẫn đến hiệu suất thường vượt trội so với phương pháp kết hợp ở cấp độ pixel hoặc cấp độ quyết định. Trong bài báo này, chúng tôi tiến hành phân tích thực nghiệm có hệ thống dưới cùng một framework nhằm đánh giá ảnh hưởng của vị trí kết hợp đặc trưng đến hiệu suất phát hiện. Kết quả thực nghiệm trên bộ dữ liệu DroneVehicle cho thấy: khi thực hiện kết hợp tại một vị trí duy nhất, việc tích hợp ở tầng giữa (*middle fusion*) mang lại sự cân bằng tối ưu giữa độ chính xác phát hiện và tốc độ suy luận, trong khi chiến lược kết hợp đa vị trí ở giai đoạn early-middle tiếp tục nâng cao độ chính xác định vị đối tượng, dù đi kèm với mức tăng nhẹ về độ phức tạp tính toán.

Từ khóa

Phát hiện đối tượng trên ảnh UAV; kết hợp ảnh mức đặc trưng; học sâu; kết hợp ảnh RGB-IR.