See the Unseen: Robust Drone and Payload Recognition Using RGB-IR Multimodal Fusion

Lahiru Cooray, Lasitha Amarasinghe, Mihiraja Kuruppu, Ravija Dulnath,
Shemal Perera, Shaveen Herath, Dinuka Madushan, Chandeepa Janith,
Dilsha Mihiranga, Kavishka Abeywardana, Muditha Fernando, Wageesha Manamperi
Department of Electronic and Telecommunication Engineering, University of Moratuwa, Sri Lanka

***Abstract*—Unmanned Aerial Vehicles (UAVs) have become increasingly prominent across various civilian and military domains, but their misuse poses growing threats to safety, security, and privacy. [1] This work focuses on developing a real-time, multimodal system for UAV detection, tracking, and payload classification using both RGB and infrared (IR) imagery. In this paper, we present a comprehensive deep learning framework that leverages RGB-IR fusion to enhance detection accuracy and robustness under diverse environmental conditions and distortions. Our pipeline comprises three parallel models: one trained on RGB, one on IR, and one on fused RGB-IR inputs. The system is designed to detect drones versus birds, track drone motion across video sequences, and classify carried payloads as harmful or benign. We demonstrate that multimodal fusion substantially improves performance in low visibility, occlusion, and noisy scenarios. Real-time inference with classification confidence and trajectory estimation confirms the effectiveness and deployability of our approach for practical surveillance environments. [2]**

I. Introduction

The proliferation of Unmanned Aerial Vehicles (UAVs), commonly known as drones, has transformed various domains ranging from logistics and agriculture to disaster relief and military surveillance. However, with the rise in drone usage comes increasing concern over unauthorized aerial activity, potential threats from malicious payloads, and significant challenges to public safety and privacy. This underscores the need for robust systems capable of detecting, tracking, and analyzing drones and their payloads in real-time, especially in sensitive or restricted airspace. [3]

Traditional vision-based approaches to drone detection primarily rely on RGB images. While effective in ideal lighting conditions, RGB-based systems degrade significantly under adverse environments such as low-light, fog, or camera instability. In contrast, infrared (IR) imaging captures thermal signatures, offering complementary advantages such as resilience in poor visibility or long-range scenarios. [4] Nonetheless, standalone use of either modality remains suboptimal due to the trade-off between spatial detail (RGB) and thermal contrast (IR). A fusion-based approach that integrates RGB and IR modalities provides a promising solution to leverage the strengths of both. [5], [6]

Moreover, detecting the presence of drones alone is insufficient in many operational contexts. It is equally important to identify the type of payload they carry, particularly when payloads may pose threats — such as explosives, surveillance tools, or contraband. The payloads themselves may differ in thermal properties, shapes, and textures, making their classification a challenging multimodal problem. [7] Fusion of RGB and IR data helps compensate for modality-specific blind spots, thus improving classification accuracy under challenging real-world conditions. To address these challenges, the overall problem statement is as follows:

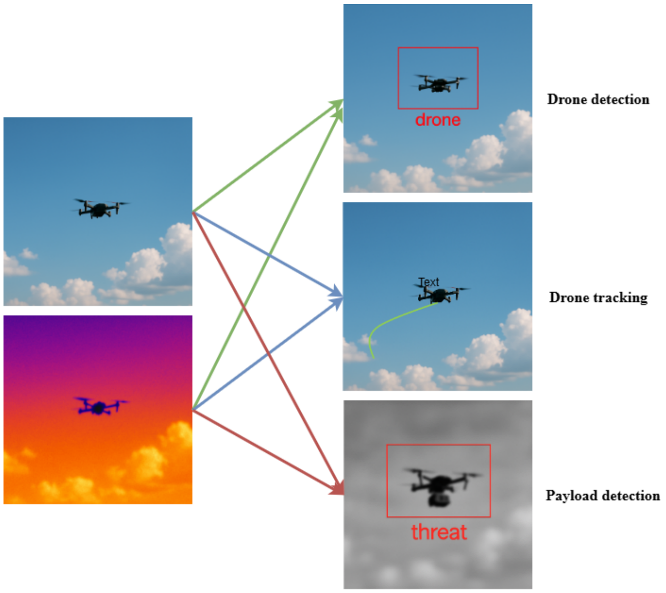

Fig. 1. Visual overview of the problem addressed in this work. RGB and infrared (IR) images are jointly used to detect aerial drones, track their trajectories over time, and identify the presence of potentially harmful payloads. The fusion of both modalities enables robust performance under adverse environmental conditions such as low light, occlusion, and motion blur.

- **Drone Detection:** Classifying aerial entities as drones or birds using RGB, IR, and fused data under various distortions and topographies.
- **Drone Tracking:** Estimating the drone trajectory and determining whether it is approaching or receding from the field of view (FoV).
- **Payload Identification:** Distinguishing between harmful and normal payloads based on RGB-IR fused imagery.

In this paper, we propose a deep learning framework tailored to this competition. Our contributions include:

- Design of a modular architecture supporting real-time detection, tracking, and classification on three input

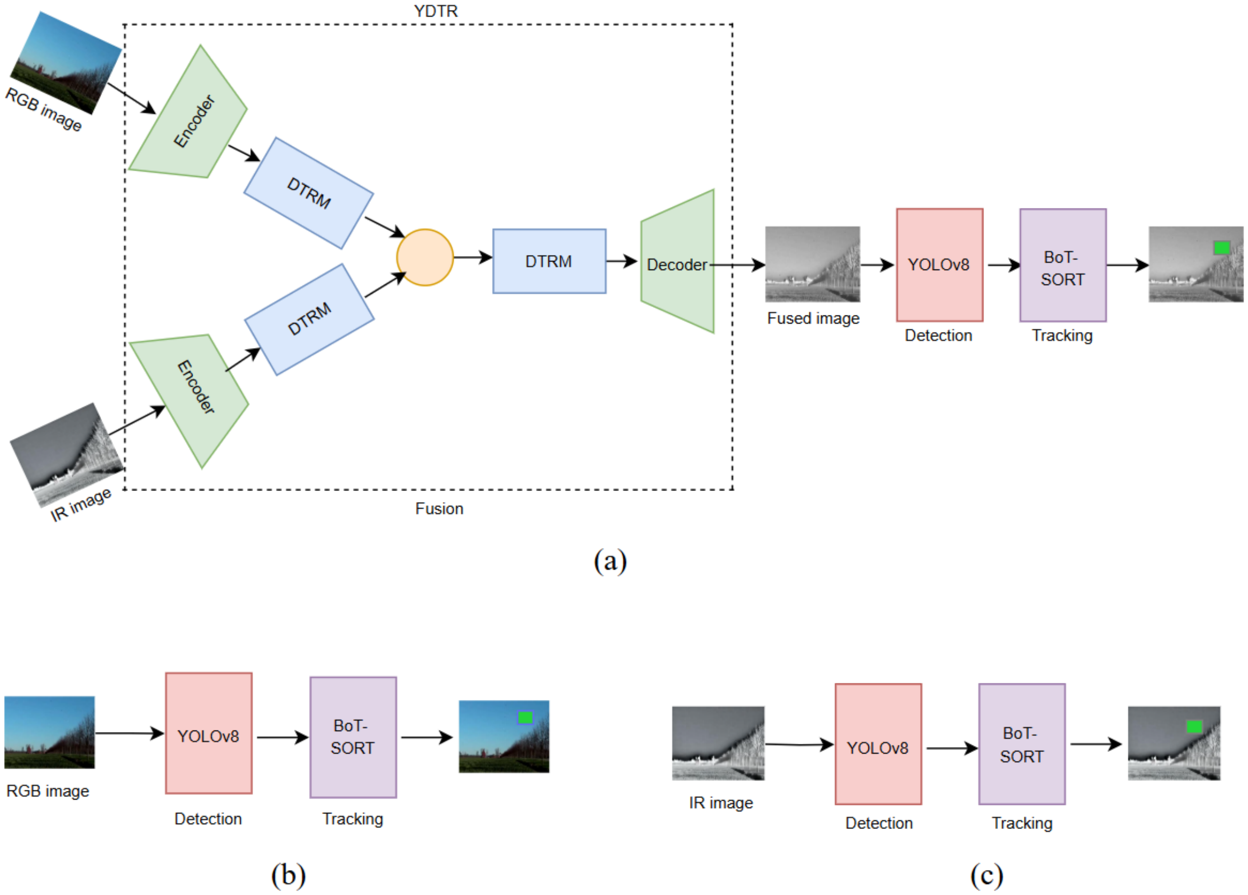

Fig. 2. Overview of the proposed drone detection and tracking pipeline. **(a)** RGB-IR fusion model: Y-shape Dynamic Transformer (YDTR) is used to fuse features from RGB and infrared images for enhanced detection and tracking. YOLOv8 serves as the object detector, and BoT-SORT is used for robust multi-object tracking. **(b)** RGB-only pipeline: uses YOLOv8 and BoT-SORT on visible spectrum images without fusion. **(c)** IR-only pipeline: detection and tracking are performed solely on thermal imagery. This setup enables a comparative analysis of single-modality and fusion-based models under varying environmental conditions.

modalities (RGB, IR, fusion).

- Implementation of robust multimodal data fusion to handle environmental distortions and motion artifacts.
- Evaluation of system performance across provided datasets, demonstrating high accuracy, inference speed, and practical deployability.

The rest of the paper is organized as follows: Section II reviews related work. Section III details our proposed methodology. Section IV presents experimental results and analysis. Section V concludes the paper with final insights and future directions.

II. Related Work

With the increasing deployment of UAVs in civilian and military environments, research on drone detection and classification has gained substantial momentum. Traditional RGB-based approaches are limited in low-light or cluttered backgrounds, prompting interest in multispectral fusion strategies.

Several works focus on the integration of RGB and IR modalities to improve robustness. Yang et al. [8] and Wang et al. [9] introduced transformer-based and illumination-aware methods to enhance UAV detection across spectral domains. These techniques improve robustness in fog, low-illumination, and camouflage settings. Additionally, Li et al. [6] presented a benchmark for multispectral pedestrian detection, providing foundational methodologies applicable to UAV detection.

Fusion-specific architectures such as TSIFNet [9], CFT [10], and YDTR [11] introduce dynamic and cross-modality attention mechanisms that extract salient features from both thermal and visual inputs. Zhao et al. [12] further extend this by proposing a generic Image Fusion Transformer, capable of handling heterogeneous modalities with learned attention gates.

In the context of object detection, the YOLO series has seen widespread adoption for real-time UAV tracking. Jocher et al. [13] and Li et al. [14] introduced YOLOv8 and YOLOv10 respectively, optimized for accuracy and latency. Gallagher and Oughton [15] comprehensively reviewed YOLO-based multispectral detection techniques, underscoring the growing trend of fusing multiple sensing modalities.

For robust multi-object tracking, Xu et al. [16] developed BoT-SORT, a tracking algorithm leveraging both appearance and motion cues. This framework has demonstrated effectiveness in high-speed, cluttered aerial video scenarios. Our system builds on this foundation by integrating BoT-SORT with fused RGB-IR detections to enable persistent tracking and motion behavior analysis.

On the security front, Nassi et al. [1] provide a systematized overview of privacy threats posed by drones, emphasizing the urgent need for real-time, adaptive surveillance solutions. Complementary surveys by Samaras et al. [4], Liu et al. [7], and the anonymous anti-UAV survey [3] highlight the evolution of multi-sensor and deep learning techniques, the challenges in payload detection, and the gaps in multi-agent UAV defense strategies.

Finally, Kim et al. [2] explore self-supervised learning in drone surveillance, offering pathways for unsupervised feature learning in complex environments—an avenue we consider valuable for future extensions of our multimodal approach.

III. Methodology

Our proposed system comprises three core components: (1) UAV Detection, (2) UAV Tracking, and (3) Payload Classification. Each module is trained and optimized independently across RGB, IR, and RGB-IR fused modalities. Central to our detection and tracking framework is the Y-shape Dynamic Transformer (YDTR) [11], which extracts and fuses features from both modalities. The system is designed for real-time operation under environmental conditions such as fog, motion blur, and low visibility. The whole pipeline is summarized as **Algorithm- 1**

A. UAV Detection with YDTR

The goal of this task is to detect UAVs using both RGB and infrared (IR) imagery, and distinguish them from other aerial objects such as birds. Accurate detection is critical for downstream tracking and payload analysis.

We employ a two-stage architecture: the Y-shape Dynamic Transformer (YDTR) for feature fusion and representation learning, followed by YOLO-based object detection [13]–[15]. The YDTR consists of two encoder branches—one for RGB and one for IR. Each branch includes:

- **Shallow Feature Extractors:** Lightweight CNNs extract early-stage spatial and thermal features from each modality.
- **Dynamic Transformer Module (DTRM):** Captures long-range dependencies and semantic context from local and global receptive fields.

Features from both branches are fused in a central DTRM block, then passed to a shared decoder. The output of YDTR is fed into a YOLOv8/YOLOv10/YOLOv12 detector trained to distinguish drones from birds.

We experimented with:

- **RGB-only:** YOLOv8, YOLOv10, YOLOv12 (100 epochs)
- **IR-only:** YOLOv8, YOLOv10 (100 epochs)
- **Fusion:** YDTR + YOLOv10 and YOLOv12

Common preprocessing includes resizing to 320 × 256, normalization, and augmentation (flip, jitter, noise). Detection performance is evaluated using precision, recall, F1-score, and mAP.

B. UAV Tracking

For real-time tracking, we integrate BoT-SORT [16] with the YDTR-enhanced YOLO detections. The system assigns persistent IDs, analyzes trajectory, and classifies motion behavior (approaching or receding).

*1) **Detection + Tracking Pipeline**:* The detection input to BoT-SORT is generated by YOLOv8 + YDTR. A confidence threshold of 0.25 is used. BoT-SORT employs hybrid appearance-motion matching, Kalman filtering, and robust ID association.

*2) **Optimized Tracker Architecture**:* We designed a class called OptimizedDroneTracker with the following capabilities:

- Deque-based buffers: Efficient history tracking of position, area, and bounding boxes.
- Distance & area analysis: Behavior classification based on motion toward/away from the camera.
- Velocity smoothing: Adaptive windowing (5–20 frames) reduces jitter and noise.
- Lazy evaluation & caching: Minimizes computation during inference.

*3) **Multi-Class Identity Management**:* We maintain separate ID counters for drones and birds. Green boxes are used for drones and blue for birds. IDs persist through occlusion and are cleaned after 15 missed frames.

*4) **Real-Time Visualization**:* Our system supports:

- Class-labeled bounding boxes and IDs
- Velocity vectors and motion trails
- Frame-wise detection scores and FPS display

*5) **Performance and Scalability**:* The combined YDTR + YOLOv8 + BoT-SORT system achieves 25–30 FPS on GPU. Tracking persistence exceeds 90%, while approach/departure classification achieves around 85%. The pipeline is optimized for low-latency execution with minimal memory usage and can be extended to multi-agent UAV scenarios and additional modalities.

C. Payload Classification (RGB-IR Fusion)

Payload classification is conducted separately from drone tracking, using RGB-IR image pairs. We use a standard CNN-based classifier trained to distinguish between harmful and normal payloads.

*1) **Architecture and Training**:* The classifier uses early fusion (channel-stacking) of RGB and IR inputs, followed by:

- 3 convolutional layers (ReLU + BatchNorm)
- 2 fully connected layers
- Softmax output

The model is trained on labeled payload images (as per dataset spec) with cross-entropy loss and evaluated using accuracy, precision, recall, F1-score, and mAP.

Algorithm 1 Multimodal UAV Detection and Classification Pipeline

```
Require: RGB frame $I_{RGB}$, IR frame $I_{IR}$
Ensure: Drone class, tracking ID, and payload classification
Step 1: Feature Fusion with YDTR
Extract shallow features: $F_{RGB} \leftarrow \text{CNN}(I_{RGB})$, $F_{IR} \leftarrow \text{CNN}(I_{IR})$
Fuse features using Dynamic Transformer Module:
    $F_{fused} \leftarrow \text{DTRM}(F_{RGB}, F_{IR})$
Step 2: Object Detection
Pass fused features to YOLO: $\mathcal{D} \leftarrow \text{YOLO}(F_{fused})$
Filter detections above confidence threshold
Step 3: Object Tracking
for each detection $d \in \mathcal{D}$ do
    Compute motion/appearance embeddings
    Update object state with BoT-SORT:
        $ID_d \leftarrow \text{BoT-SORT}(d)$
    Update tracker memory
end for
Step 4: Motion Behavior Analysis
for each ID do
    Calculate velocity and area change
    Classify as approaching or receding
end for
Step 5: Payload Classification (Early Fusion)
Stack channels: $I_{stacked} \leftarrow [I_{RGB}, I_{IR}]$
Predict payload class:
    $y \leftarrow \text{CNN}_{\text{payload}}(I_{stacked})$
return Detections $\mathcal{D}$ with tracking IDs and payload labels $y$
```

*2) **Limitations and Future Work**:* While our current payload classification is effective in visible and thermal domains, future work may include extending YDTR for direct semantic fusion of payload features, or using transformer-based attention layers for better payload region localization.

D. Box-Level late fusion

In this approach, object detection is performed independently on both RGB and infrared (IR) images using two separate YOLOv8 models (YOLOv8-RGB and YOLOv8-IR). The detection results from each modality, including bounding boxes, confidence scores, and class labels, are first extracted separately. These results are then combined into a unified set of detections. To eliminate redundant or overlapping detections from both sources, Non-Maximum Suppression (NMS) is applied based on the Intersection over Union (IoU) and confidence scores. The final output consists of refined detections that integrate information from both RGB and IR inputs, enhancing detection robustness through complementary sensing.

IV. Experimental Results

We evaluate our system using the official IEEE VIP Cup 2025 datasets across three tasks: drone detection, tracking, and payload classification. The evaluation pipeline is explained as **Algorithm-2**

A. Evaluation Metrics

The following metrics are used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Mean Average Precision (mAP) is used for detection and classification performance where applicable.

B. Drone Detection Performance

TABLE I
Drone Detection Accuracy on Validation Set

Model	Modality	F1 Score
YOLOv8	RGB	0.9759
YOLOv10	RGB	0.9726
YOLOv12	RGB	0.9687
YOLOv8	IR	0.9827
YOLOv10	IR	0.9804
YOLOv8	Fusion	0.9846

Table I summarizes the F1 scores achieved by different YOLO variants across RGB, IR, and fused modalities. The highest performance is observed with the YOLOv10 model trained on IR imagery (F1 = 0.9804), closely followed by YOLOv8 on the fusion dataset (F1 = 0.9846). These results highlight the complementary nature of thermal and visible modalities.

The comparatively lower scores of YOLOv12 across all modalities suggest that deeper or more complex detection architectures may not necessarily yield superior results in this task, particularly when trained on limited or noise-affected UAV datasets. YOLOv8 and YOLOv10, with their balanced complexity and efficiency, show better generalization.

Notably, IR-only models (YOLOv8 and YOLOv10) outperform their RGB-only counterparts, which confirms the robustness of thermal imaging in conditions such as low light, fog, and occlusion. However, the best performance overall is achieved by the fusion model (YOLOv8 + RGB-IR), indicating that combining textural cues from RGB with thermal saliency from IR provides richer features for more reliable drone detection.

These findings validate our fusion-based approach and support the design decision to incorporate both RGB and

Algorithm 2 Performance Evaluation of UAV Detection and Classification

```
Input: Ground Truth Annotations $G = \{g_1, g_2, \dots, g_n\}$, Predictions $P = \{p_1, p_2, \dots, p_m\}$
Set IoU threshold $\theta_{IoU} = 0.5$
for each predicted box $p_i \in P$ do
    for each ground truth $g_j \in G$ do
        Compute Intersection-over-Union: $IoU(p_i, g_j)$
        if $IoU(p_i, g_j) \geq \theta_{IoU}$ and $g_j$ not matched then
            Count as True Positive (TP)
            Mark $g_j$ as matched
        end if
    end for
end for
Count remaining unmatched predictions as False Positives (FP)
Count remaining unmatched ground truths as False Negatives (FN)
Compute metrics:
    $\text{Precision} = \frac{TP}{TP + FP}$
    $\text{Recall} = \frac{TP}{TP + FN}$
    $\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
For detection tasks, compute:
    $\text{mAP@50} = \text{mean of AP at IoU = 0.5}$
    $\text{mAP@50-95} = \frac{1}{10} \sum_{\theta \in \{0.5, 0.55, \dots, 0.95\}} \text{AP}_\theta$
Output: Evaluation metrics: Precision, Recall, F1, mAP@50, mAP@50–95
```

IR information via the YDTR module for UAV detection in challenging visual environments.

C. Comparison of RGB and IR Modalities for Payload Classification

To assess the individual contributions of RGB and infrared (IR) modalities in payload classification, we trained two separate models using RGB-only and IR-only image inputs, respectively. The classification performance is summarized in Table II.

TABLE II
PAYLOAD CLASSIFICATION RESULTS BY MODALITY

Modality	Precision	Recall	F1 Score	mAP50	mAP50–95
IR-only	0.9926	0.9984	0.9955	0.9949	0.9947
RGB-only	0.9881	1.0000	0.9940	0.9949	0.9949

As shown in the results, both modalities deliver excellent classification performance, achieving F1 scores above 0.99. However, the IR-only model slightly outperforms the RGB-only model across most metrics, including precision (0.9926 vs. 0.9881), F1 score (0.9955 vs. 0.9940), and mAP@50–95 (0.9947 vs. 0.9949).

This indicates that thermal signatures from payloads provide highly discriminative features for classification, especially under challenging conditions like poor illumination or visual occlusion where RGB may lose detail. Nevertheless, the near-perfect recall of the RGB model (1.0000) suggests that it can still reliably detect all payload instances, though at a slightly higher false positive rate.

These findings validate the utility of the IR modality in enhancing classification precision and robustness. In future work, we aim to exploit RGB-IR fusion to combine the complementary strengths of both modalities and push performance even further.

V. CONCLUSION

In this paper, we presented a comprehensive deep learning framework for robust drone and payload recognition using RGB-IR multimodal fusion. Our system, developed in response to the IEEE VIP Cup 2025 challenge, integrates Y-shape Dynamic Transformers (YDTR), state-of-the-art YOLO object detectors, and BoT-SORT trackers to achieve high accuracy in drone detection, real-time tracking, and payload classification under diverse environmental conditions.

Experimental results demonstrate the effectiveness of RGB-IR fusion, with fused models outperforming single-modality counterparts, especially in low-visibility and occluded scenes. The YDTR fusion module, combined with optimized object detection and tracking pipelines, enables high inference speed and strong generalization across modalities.

By leveraging both spatial and thermal cues, our approach offers a scalable and practical solution for UAV surveillance systems in real-world security and defense contexts. Future work will explore extending the YDTR architecture for end-to-end multimodal semantic segmentation and incorporating self-supervised learning for better adaptation to unseen environments.

REFERENCES

[1] B. Nassi, A. Shabtai, R. Masuoka, and Y. Elovici, “Sok – security and privacy in the age of drones: Threats, challenges, solution mechanisms, and scientific gaps,” *arXiv preprint arXiv:1903.05155*, Mar 2019. [Online]. Available: https://arxiv.org/abs/1903.05155

[2] J. Kim, J. Lee, and B. Han, “Self-supervised learning for drone-based surveillance systems: A comprehensive review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, early Access.

[3] Anonymous, “A comprehensive survey on anti-uav methods, benchmarking, and future challenges,” *arXiv preprint arXiv:2504.11967*, 2025. [Online]. Available: https://arxiv.org/abs/2504.11967

[4] S. Samaras, E. Diamantidou, D. Ataloglou, and D. Tzovaras, “Deep learning on multi sensor data for counter uav applications—a systematic review,” *Sensors*, vol. 19, no. 22, p. 4973, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/22/4973

[5] L. Yang, R. Ma, and A. Zakhor, “Drone object detection using rgb/ir fusion,” in *arXiv preprint arXiv:2201.03786*, 2022. [Online]. Available: https://arxiv.org/abs/2201.03786

[6] C. Li, J. Liang, Y. Wang, S. Yan, and X. Xu, “Multispectral pedestrian detection: Benchmark dataset and baseline,” *ECCV*, 2018.

[7] Z. Liu, P. An, Y. Yang, S. Qiu, Q. Liu, and X. Xu, "Vision-based drone detection in complex environments: A survey," *Drones*, vol. 8, no. 11, p. 643, 2024.

[8] L. Yang, R. Ma, and A. Zakhor, "Drone object detection using rgb/ir fusion," in *arXiv preprint arXiv:2201.03786*, 2022.

[9] H. Wang, W. Liu, X. Zhang *et al.*, "Tsifnet: Transformer-based simultaneous infrared and visible image fusion network," *Information Fusion*, vol. 91, pp. 1–13, 2023.

[10] Q. Fang, D. Han, and Z. Wang, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021. [Online]. Available: https://arxiv.org/abs/2111.00273

[11] W. Tang, F. He, and Y. Liu, "Ydtr: Infrared and visible image fusion via y-shape dynamic transformer," *IEEE Transactions on Multimedia*, vol. 25, pp. 5413–5428, 2023.

[12] Y. Zhao, L. Zhang, and Y. Zheng, "Image fusion transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 3512–3527, 2022.

[13] G. Jocher, A. Chaurasia, T. Qiu *et al.*, "Yolov8: Next-generation object detection," *Ultralytics Technical Report*, 2023, https://github.com/ultralytics/ultralytics.

[14] J. Li, Y. Chen, G. Wang, J. Li, Z. Liu *et al.*, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2403.04344*, 2024.

[15] J. E. Gallagher and E. J. Oughton, "Surveying you only look once (yolo) multispectral object detection: Advancements, applications and challenges," *arXiv preprint arXiv:2409.12977*, 2024.

[16] Y. Xu, L. Zhang, and X. Zhang, "Bot-sort: Robust association tracking with appearance embedding," *arXiv preprint arXiv:2206.14651*, 2022.