Available online at www.jpit.az

16 (2)
2025

Thermal-RGB fusion with lightweight CNNs for night-time drone surveillance and real-time adaptive sensor selection

Vugar Gasimov

Azerbaijan Technical University, H.Javid avenue, 25, AZ1073 Baku, Azerbaijan

vugarbstateu@gmail.com

https://orcid.org/0009-0008-5245-7498

ARTICLE INFO

Keywords:

Thermal-RGB fusion
Lightweight CNN
Adaptive sensor selection
Drone surveillance
Object detection

ABSTRACT

This paper presents a lightweight and adaptive object detection framework tailored for night-time drone surveillance using thermal-RGB fusion. The proposed system integrates a dual-branch CNN backbone to extract modality-specific features, followed by a mid-level fusion strategy that combines thermal and RGB representations into a unified feature map. A real-time adaptive sensor selection mechanism dynamically adjusts the contribution of each modality based on scene context, enhancing robustness under varying illumination conditions. The architecture employs YOLOv5-Nano as a compact feature extraction backbone, supported by quantization and pruning for real-time deployment. Extensive experiments conducted on publicly available datasets, demonstrate superior accuracy and efficiency compared to single-modality and non-adaptive baselines. Ablation studies validate the contribution of each component. The results highlight the system's capability to achieve high detection performance with low computational overhead, offering a practical and scalable solution for intelligent aerial surveillance in complex night-time environments.

1. Introduction

In recent years, the integration of unmanned aerial vehicles (UAVs), commonly known as drones, into surveillance systems has significantly transformed the landscape of real-time monitoring and security. Due to their mobility, aerial perspective, and increasing computational capabilities, drones have been widely adopted in diverse fields such as disaster management, military reconnaissance, traffic monitoring, and border security. However, object detection, the core component of intelligent surveillance, remains a critical challenge, especially under adverse conditions such as night-time or low-light environments where traditional RGB (Red-Green-Blue) imaging often fails to capture sufficient visual cues.

Thermal infrared imaging offers a complementary modality that can detect heat-emitting objects regardless of lighting conditions, making it especially valuable for nocturnal surveillance. Unlike RGB cameras that depend on ambient illumination, thermal sensors visualize the infrared radiation emitted by objects, enabling reliable detection of humans, vehicles, and other heat sources even in complete darkness. Nevertheless, thermal imaging alone is often limited in spatial resolution and object texture, which are critical for accurate detection and classification. Therefore, fusing RGB and thermal modalities can potentially combine the strengths of both sensors, offering more robust and context-aware object detection across varying environmental conditions (Ozcan & Cetin, 2022; Zhu & et al 2023).

Despite the advantages of multimodal sensing, achieving real-time, energy-efficient object detection on drone platforms remains a substantial technical challenge. Most state-of-the-art deep

Received 7 March 2025, Received in revised form 8 May 2025, Accepted 17 June 2025
http://doi.org/10.25045/jpit.v16.i2.04

learning-based object detectors such as YOLOv5, Faster R-CNN, and DETR require extensive computational resources, which are often incompatible with lightweight systems (Li, Wang, & Zhang, 2023). In addition, existing multimodal fusion methods tend to focus on accuracy while neglecting real-time performance and energy constraints, which are critical in battery-powered aerial platforms. The challenge, therefore, lies in designing lightweight yet effective architectures that can perform reliable object detection while respecting the stringent resource limitations (Lim, Jin, & Bang, 2023).

A major challenge in aerial object detection is the assumption that all sensor inputs are equally reliable. However, the usefulness of RGB and thermal data varies with factors like lighting, weather, and scene complexity. Static fusion may introduce noise or redundancy. Therefore, a dynamic sensor selection mechanism that adapts in real time to environmental context is essential for improving robustness and efficiency.

In this study, we propose a novel approach for thermal-RGB fusion using lightweight convolutional neural networks (CNNs) tailored for night-time drone surveillance, augmented by a real-time adaptive sensor selection mechanism. Our method leverages a dual-branch CNN architecture to independently process RGB and thermal inputs, followed by a mid-level fusion module that combines informative features from both modalities (He, Liu, & Chen, 2023). Furthermore, an attention-based sensor selection block is integrated to dynamically adjust the contribution of each modality based on scene characteristics, enabling context-aware perception. To ensure suitability for real-time inference, we incorporate lightweight design principles, including the use of a compact backbone such as YOLOv5-Nano, as well as model pruning and quantization techniques (Kim et al., & 2024).

The contributions of this paper are as follows. First, we propose a novel multimodal object detection architecture that fuses thermal and RGB information using a lightweight CNN, optimized specifically for real-time drone surveillance in low-light conditions. Second, we introduce a real-time adaptive sensor selection mechanism that dynamically adjusts the weighting of RGB and thermal inputs based on the environmental context, thereby enhancing detection performance and robustness. Third, we conduct extensive experiments using publicly available RGB-T datasets, such as KAIST (https://soonminhwang.github.io/rgbt-ped-detection/) and FLIR (https://www.flir.com/oem/adas/adas-dataset-form/), as well as our own custom aerial footage, benchmarking our method against state-of-the-art lightweight object detectors (Liu, Wang, & Zhang, 2024). Finally, we present a comprehensive ablation study and real-time performance analysis including metrics such as mAP, FPS, and energy usage to validate the efficiency and deployability of the proposed system.

2. Proposed approach

The proposed system is designed to perform robust and real-time object detection in night-time aerial surveillance scenarios by integrating thermal and RGB imagery through a lightweight, adaptive architecture. The core idea is to exploit the complementary strengths of thermal and RGB modalities while maintaining computational efficiency suitable for deployment in low-resource environments (Zhang, Ji, & Zeng, 2024). The system is composed of four key modules: a dual-branch feature extraction backbone for processing RGB and thermal inputs independently, a mid-level fusion module for combining the learned features, an adaptive sensor selection mechanism that dynamically weights modality contributions based on scene conditions, and a lightweight object detection head that produces final predictions.

The pipeline begins with two parallel input streams: one for RGB images and the other for thermal infrared images. These images are first passed through two separate convolutional branches, each equipped with a lightweight CNN backbone such as YOLOv5-Nano. These branches are responsible for extracting low-to-mid-level features specific to their respective modalities (Sun et al., 2024). The rationale for using separate pathways at this stage is to allow each branch to specialize in modality-specific characteristics. RGB imagery typically contains high spatial detail and color gradients, whereas thermal images encode temperature distributions that are invariant to lighting (Wang et al., 2023).

The outputs from both branches, namely the feature maps, are forwarded to the fusion module. Unlike early fusion strategies that merge inputs at the raw image level, our approach employs mid-level fusion, which allows for more semantically meaningful integration of information. This

module concatenates or merges the feature maps along the channel dimension and then processes them through shared convolutional layers to produce a unified representation. The fusion operation can be as simple as channel-wise concatenation followed by 1×1 convolutions, or it can incorporate attention-based mechanisms to learn adaptive fusion weights.

A key component of our system is the real-time adaptive sensor selection block, which runs alongside the fusion process. It assigns dynamic weights to RGB and thermal feature maps based on their informativeness in each frame. Implemented via global context pooling and a lightweight MLP, this module learns to prioritize modalities using scene-level cues like brightness or contrast, enhancing adaptability under varying night-time conditions.

The fused and adaptively weighted features are then passed to the detection head, which is based on a lightweight, anchor-free object detection framework. This head consists of several convolutional layers followed by classification and regression branches. The classification branch outputs the probability scores for predefined object classes, while the regression branch predicts the bounding box coordinates. To maintain real-time inference speeds, we minimize the number of convolutional layers in the detection head and apply depthwise separable convolutions where possible.

The modular design of the system enables efficient deployment on low-power edge devices. Supporting both batch and frame-by-frame inference, it suits continuous aerial surveillance. The adaptive sensor selection ensures robustness under rapidly changing lighting and thermal conditions (Sun et al., 2024).

In summary, the proposed architecture effectively tackles challenges of multimodal fusion, efficiency, and adaptability for night-time drone object detection. Its modular design ensures both performance and practicality for real-world surveillance applications.

In the context of UAV-based surveillance systems, the computational limitations and energy constraints of onboard hardware necessitate the use of compact and efficient deep learning architectures. Unlike high-performance server environments where large-scale models such as ResNet-101 or EfficientDet-D7 can be deployed without major concerns, low-power platforms typically operate under strict power, memory, and latency budgets (Lim et al., 2023). To address these constraints, the proposed system integrates a lightweight backbone design for feature extraction in both RGB and thermal input streams, enabling real-time performance without compromising detection accuracy.

The backbone network is the foundational component responsible for extracting meaningful features from raw input images. In our framework, we adopt architectures such as YOLOv5-Nano due to their proven efficiency in edge computing environments. These models are specifically engineered to reduce parameter count and floating-point operations per second (FLOPs), making them suitable for deployment in resource-constrained settings (Sun et al., 2024).

YOLOv5-Nano is an ultra-compact object detection model optimized for real-time inference. It retains the essential structure of the YOLO family while drastically reducing the number of parameters and computations. The model includes a backbone, neck, and head, where the backbone is typically constructed with CSP (Cross Stage Partial) connections to improve gradient flow and reduce redundancy. In our final implementation, YOLOv5-Nano was used solely as a feature extractor by omitting its detection head. It was selected over MobileNetV3 due to better feature richness and compatibility with our fusion module (Kim et al., 2024).

For both RGB and thermal inputs, independent backbone instances are used to extract low-level and mid-level feature maps. This dual-backbone approach ensures that modality-specific spatial and contextual information is preserved prior to fusion. Given the distinct nature of RGB and thermal imagery, sharing a single backbone could lead to a loss of modality-specific characteristics, especially in early layers where fine-grained features dominate.

To further optimize the architecture, we apply post-training techniques such as quantization and model pruning. Quantization reduces the precision of model weights and activations from 32-bit floating point to 8-bit integers, resulting in significant reductions in model size and inference time. Meanwhile, pruning eliminates less significant connections or filters, effectively compressing the model without degrading performance beyond acceptable thresholds (He et al., 2023). These techniques are especially beneficial when deploying the system in energy-limited environments.

The design of the backbone also takes input resolution into account. A balance must be struck between spatial resolution, which affects detection accuracy, and computational load, which affects inference speed. During experimentation, input resolution of 416×416 pixels was found to offer a good trade-off between precision and real-time responsiveness, making it suitable for both RGB and thermal branches under varying flight conditions.

Multimodal fusion plays a pivotal role in enhancing the performance of object detection systems operating under challenging conditions such as night-time aerial surveillance. Among various fusion paradigms, mid-level fusion offers a compelling trade-off between early fusion, which merges raw data at the input level, and late fusion, which combines decision outputs (Zhang et al., 2023). The proposed system adopts a mid-level fusion strategy for integrating RGB and thermal modalities, motivated by its ability to preserve modality-specific features while leveraging their complementary information in a unified representation.

In early fusion approaches, the raw RGB and thermal images are concatenated into a multi-channel input before being fed into a shared backbone. While computationally efficient, this method often suffers from modality misalignment and information loss, particularly when the sensors differ in spatial resolution or field of view. Moreover, early fusion is limited in its capacity to capture high-level semantic representations specific to each modality. In contrast, late fusion architectures process each modality separately through independent detection branches and then combine the outputs through voting, averaging, or non-maximum suppression (Zhu et al., 2023). Although this approach can retain modality-specific predictions, it introduces latency and redundancy and fails to model cross-modal interactions during feature learning.

Mid-level fusion extracts features from RGB and thermal inputs using separate lightweight CNN backbones (e.g., YOLOv5-Nano), then aligns and merges them into a joint representation for further processing by the detection head (Sun et al., 2024).

The fusion process uses channel-wise concatenation of RGB and thermal features, followed by shared convolutional layers to refine the joint representation. Features are spatially aligned via interpolation if needed, then processed through 1×1 and 3×3 convolutions with batch normalization and ReLU, enabling the network to learn more discriminative joint features.

To enhance fusion effectiveness, we integrate a lightweight attention module using a Squeeze-and-Excitation (SE) block. It compresses the fused feature map to a global descriptor and computes channel-wise weights via fully connected layers. These weights emphasize informative channels, helping the system adapt to scene conditions—favoring thermal features in darkness and RGB in partially lit areas (Zhu et al., 2023).

An important consideration in mid-level fusion is modality synchronization. While RGB and thermal cameras can operate at different frame rates or with slight spatial misalignment, our system assumes a pre-processing step where temporal and spatial alignment is achieved. This may involve hardware-level synchronization during data acquisition or software-level alignment using homography-based transformation. Accurate alignment ensures that corresponding features represent the same spatial regions, which is critical for effective fusion.

Additionally, the fusion module is designed to maintain computational efficiency. All convolutional layers use depthwise separable convolutions, and the attention block is kept lightweight to avoid increasing the inference time significantly (Zhang et al., 2023). During training, the fusion module is jointly optimized with the rest of the network using multi-task loss functions that include classification and bounding box regression. This end-to-end training strategy enables the model to learn not only modality-specific features but also optimal fusion patterns directly from the data.

Mid-level fusion thus emerges as a practical and effective approach for thermal-RGB integration in drone surveillance systems. It provides a rich joint representation that captures both structural details from RGB inputs and thermal signatures from infrared imagery, while maintaining compatibility with lightweight CNN backbones such as YOLOv5-Nano and enabling real-time deployment. By enabling cross-modal interaction at a semantically meaningful level, the fusion strategy enhances object localization and classification accuracy under diverse and dynamic environmental conditions.

While the fusion of RGB and thermal modalities improves object detection in low-light environments, not all frames or environmental conditions equally benefit from both modalities. In certain situations, such as scenes with partial illumination or thermal noise, one modality may

provide more informative and reliable features than the other. Consequently, treating both modalities with equal weight during fusion can introduce redundancy, noise, or even mislead the detection system. To address this challenge, our proposed architecture incorporates a real-time adaptive sensor selection mechanism that dynamically modulates the influence of each modality based on scene context and feature quality.

The core principle of this mechanism is to allow the model to learn which modality to prioritize at a given moment, rather than statically combining all information regardless of its relevance. This is particularly critical in drone-based surveillance systems where environmental conditions such as lighting, altitude, occlusion, or thermal diffusion may rapidly change within a single flight session. The proposed selection mechanism evaluates modality-specific features on a per-frame basis and outputs adaptive weighting coefficients, which are applied to the feature maps before fusion.

The adaptive selection module is implemented as a lightweight attention block that operates independently on the feature maps from the RGB and thermal branches. Specifically, given a feature map $F \in R^{C\times H\times W}$, we first apply global average pooling across the spatial dimensions to obtain a compact channel descriptor $z \in R^{C}$. This descriptor captures the global context of the feature map, serving as an abstract summary of its semantic richness. A shallow multi-layer perceptron (MLP) then processes this descriptor to produce a modality-specific relevance score. This operation is repeated independently for both RGB and thermal feature maps.

The two resulting scores are passed through a softmax function to produce normalized w_{RGB} and $w_{thermal}$, which sum to one. These weights are used to scale the corresponding feature maps prior to fusion, such that:

$$\mathrm{F_{fused}} = \mathrm{w_{RGB}} \cdot \mathrm{F_{RGB}} + \mathrm{w_{thermal}} \cdot \mathrm{F_{thermal}}\,, \quad (1)$$

This formulation ensures that the network focuses more on the modality providing higher-quality features in the current context, while still retaining information from the other modality when useful. Unlike hard selection or gating mechanisms that completely discard one modality, soft attention-based weighting maintains robustness and continuity across frames.

In practice, adaptive weighting improves detection in scenes where one modality is unreliable. For instance, thermal data is more useful under glare, while RGB dominates when thermal contrast is low. The system learns to adjust automatically from training data, without manual rules or supervision.

To ensure that the adaptive selection block remains lightweight and suitable for edge deployment, we constrain the MLP to a small number of parameters, typically two fully connected layers with ReLU activation and optional dropout for regularization. The entire module adds negligible overhead to the overall model size and inference time but contributes significantly to detection robustness in variable conditions.

During training, the adaptive selection module is learned end-to-end alongside the main network using a multi-task loss that combines classification and regression. It requires no extra supervision, as loss gradients naturally guide it toward better modality weighting for improved detection.

In addition to attention-based selection, we also experimented with simpler alternatives such as fixed heuristics based on scene brightness or entropy. While such rule-based methods offered marginal improvements in specific cases, they lacked the generalization and flexibility of the attention-driven approach. Furthermore, fixed rules do not adapt to subtle patterns in the data that neural networks can capture through learning.

The adaptive sensor selection mechanism also contributes to energy efficiency. By identifying frames where one modality is clearly dominant, the system can potentially downsample or skip the less informative sensor's processing, reducing computational load. This opens up avenues for future extensions involving energy-aware dynamic inference, where the system selectively activates parts of the network based on context.

Overall, real-time adaptive sensor selection boosts both the intelligence and robustness of the system, enabling context-aware perception that balances accuracy and efficiency, which is crucial for reliable night-time drone surveillance in dynamic environments.

The effectiveness of any deep learning-based object detection system significantly depends on the quality of its training procedure and the optimization strategies used to adapt it to the target platform. In the proposed thermal-RGB fusion framework with adaptive sensor selection, the training pipeline is carefully designed to ensure accuracy, generalization, and computational

efficiency. This section outlines the training methodology, loss function design, optimization setup, and post-training compression techniques that enable the deployment of the model on real-time, resource-constrained drone platforms.

The model is trained on a combination of publicly available multimodal datasets, including the KAIST Multispectral Pedestrian Dataset and the FLIR ADAS Dataset, both of which provide aligned thermal and RGB image pairs. In addition to these, custom drone-captured RGB-thermal video footage was used to enrich the dataset with aerial perspectives and night-time urban scenarios. To ensure modality consistency, all RGB and thermal images are resized to a fixed resolution (e.g., 416×416) and normalized independently. Data alignment is precomputed during preprocessing, either using intrinsic sensor calibration parameters or feature-based homography transformation when camera synchronization is imperfect.

Data augmentation techniques are applied independently to both modalities to increase generalization capability. These include horizontal flipping, scaling, rotation, and slight brightness or contrast adjustments (limited to the RGB branch). Care is taken not to apply augmentations that could disrupt inter-modality correspondence. During training, mini-batches are constructed to contain balanced samples from both day and night scenes to ensure robustness across varying lighting conditions.

The proposed architecture uses a multi-task loss function that jointly optimizes object classification and bounding box regression. The classification loss is computed using the binary cross-entropy (BCE) function for each anchor and class, while the regression loss is based on the Complete IoU (CIoU) metric, which accounts for overlap area, center distance, and aspect ratio differences between predicted and ground truth boxes. The total loss L is defined as:

$$L = \lambda_{cls} L_{\mathbf{cls}} + \lambda_{reg} L_{reg}, \quad (2)$$

Where L_{cls} is the classification loss, L_{reg} is the CIoU-based localization loss, and λ_{cls}, λ_{reg} are weighting coefficients empirically tuned to balance the two terms. In our experiments, values of $\lambda_{cls} = 1.0$ and $\lambda_{reg} = 5.0$ produced stable convergence.

The adaptive sensor selection mechanism is trained end-to-end without any explicit supervision. Its weighting decisions are guided by the gradients of the main detection loss. Since modality weights directly influence feature fusion, the network learns to assign higher importance to the more informative modality by minimizing the total detection error during backpropagation.

The network is trained using the Adam optimizer with decoupled weight decay (AdamW), which provides stability and fast convergence for complex multimodal architectures. The initial learning rate is set to 0.001 with a cosine annealing schedule that gradually decays the learning rate to a minimum value of 1e-6. Batch size is set based on the available GPU memory, typically ranging from 16 to 32, and the number of training epochs is between 100 and 150 depending on convergence behavior.

To prevent overfitting, label smoothing (with a factor of 0.1) is used in the classification loss, and early stopping is applied if validation accuracy does not improve for 10 consecutive epochs. Dropout layers (with dropout rate 0.3) are also used in the sensor selection block to regularize the modality weighting mechanism.

Following the training phase, two model compression techniques are applied to enhance deployment efficiency: quantization and structured pruning.

Quantization is performed using post-training static quantization with 8-bit integer precision for both weights and activations. This significantly reduces the model size and memory footprint while maintaining inference accuracy within 1–2% of the full-precision model. Quantization-aware training (QAT) is also optionally used during fine-tuning to improve robustness to reduced numerical precision, especially for deployment in resource-constrained environments.

Structured pruning is applied to remove redundant filters and channels from convolutional layers. A sparsity constraint is gradually introduced during retraining using L1-norm-based importance scoring. Only filters with low activation responses across the training data are pruned. After pruning, the model is fine-tuned for a few additional epochs to restore accuracy.

The final compressed model is exported in ONNX or TensorRT format depending on the deployment environment. Inference benchmarking shows that the optimized model maintains real-time performance with significant speed improvements. Moreover, energy profiling reveals a 30–40% reduction in power consumption after applying quantization and pruning, demonstrating the practicality of the model for deployment in resource-constrained environments.

3. Experimental results

All experiments were conducted on a standard PC without physical drone hardware or external sensors. The evaluation was based entirely on pre-recorded drone footage containing RGB and thermal video data, sourced from publicly available datasets and simulated aerial scenarios. The primary objective of this setup was to test the proposed thermal-RGB fusion and adaptive sensor selection framework under conditions that approximate real-world drone surveillance, while operating within the constraints of a typical desktop computing environment.

All training and evaluation procedures were conducted using the PyTorch framework (version 1.13), supported by auxiliary libraries such as NumPy, OpenCV, and Albumentations, on a desktop computer equipped with an Intel Core i5 processor, 16 GB of RAM, and an NVIDIA GTX 1650 GPU with 4 GB of VRAM; while this hardware setup does not fully replicate the resource constraints of embedded UAV platforms, it provides a practical and accessible environment for initial experimentation, model development, and performance validation.

Input video frames were extracted from thermal and RGB video files and resized to a fixed resolution of 416×416 pixels. Thermal images were normalized using min-max scaling, while RGB images were normalized based on standard ImageNet statistics. Frame-by-frame processing was employed to simulate real-time inference conditions, though no live camera feed or real-time sensor communication was used. Since all footage was pre-recorded, the RGB and thermal frames were either natively aligned or manually pre-aligned during preprocessing. Minor misalignments were corrected using affine transformations to ensure spatial consistency across modalities.

Training was conducted using the AdamW optimizer with an initial learning rate of 0.001 and a cosine annealing schedule. The model was trained for 120 epochs with a batch size of 16. Early stopping based on validation loss was used to prevent overfitting. Model checkpoints were saved at regular intervals, and the final model was selected based on the highest validation mean average precision (mAP). Label smoothing and data augmentation techniques, such as random flipping, rotation, and brightness variation, were applied during training to enhance generalization, particularly under variable lighting conditions.

Although the system was not deployed on embedded hardware such as NVIDIA Jetson Nano or Xavier NX, the model's inference time, memory usage, and parameter count were monitored to assess its suitability for lightweight deployment. Post-training quantization and pruning experiments were also conducted on the same PC to emulate performance under reduced computational budgets.

Performance evaluation was carried out using isolated frame processing to mimic drone-based streaming. Frame processing times and memory usage were logged during inference to estimate the feasibility of real-time operation. All results reported in the following sections were obtained using this offline, PC-based experimental setup, with the goal of ensuring reproducibility in non-specialized computing environments.

To evaluate the proposed thermal-RGB fusion and adaptive sensor selection method, we used publicly available datasets that contain paired RGB and thermal imagery suitable for night-time surveillance research. Since no physical drone or onboard sensor system was employed, all experimental data was sourced from pre-recorded aerial or surveillance-style footage that emulates drone views under varying lighting conditions.

The primary dataset used in this study is the KAIST Multispectral Pedestrian Dataset, which provides aligned RGB and thermal image pairs captured in urban environments. This dataset includes over 95,000 annotated frames, covering both day and night scenarios. It is particularly well-suited for evaluating thermal-RGB fusion methods due to its synchronized dual-modal data and pixel-level annotations for pedestrian detection. Only night-time and low-light sequences were used in our experiments to align with the goals of this research.

In addition to KAIST, we employed selected samples from the FLIR ADAS Dataset, which offers high-resolution thermal and RGB images taken from a moving vehicle in urban and semi-urban areas. Although not strictly drone footage, the perspective and scene dynamics are sufficiently similar to emulate low-altitude UAV surveillance. The dataset includes over 10,000 labeled frames with thermal and RGB modalities, and was used primarily for validation and generalization testing.

We also used open-access night-time drone footage with RGB videos from low altitudes. Selected frames were manually annotated, and thermal counterparts were synthetically aligned using frame-to-frame interpolation to simulate fusion scenarios.

All datasets were preprocessed to ensure modality alignment, resized to 416×416 resolution, and split into training (70%), validation (15%), and testing (15%) subsets. Augmentation was applied during training, and care was taken to maintain consistency between RGB and thermal frames.

Together, these datasets provide a diverse and realistic basis for evaluating the model's performance in night-time aerial object detection using multimodal input.

The performance of the proposed object detection system was evaluated using a combination of accuracy, efficiency, and robustness metrics. For accuracy assessment, we used the mean Average Precision (mAP) at Intersection over Union (IoU) thresholds of 0.5 (mAP@0.5) and the COCO-style averaged range (mAP@0.5:0.95), which provide a comprehensive measure of detection precision. To evaluate real-time capability, we measured inference speed in frames per second (FPS) during testing on a standard PC. Model efficiency was quantified through the total number of parameters, model size in megabytes (MB), and estimated floating-point operations per second (FLOPs). Additionally, energy usage was indirectly assessed by monitoring GPU memory and power consumption during inference. These metrics collectively offer a balanced evaluation of the model's effectiveness, lightweight design, and potential suitability for real-time deployment on aerial platforms.

To evaluate the effectiveness of the proposed thermal-RGB fusion model with adaptive sensor selection, we conducted a comprehensive comparison against several baseline models. These models represent a variety of design choices in object detection, including single-modality inputs, fixed fusion strategies, and lightweight architectures without adaptation mechanisms. The primary goal of these comparisons is to assess the added value of both multimodal fusion and real-time adaptive sensor selection under night-time aerial surveillance scenarios.

The first baseline model is the YOLOv5-Nano (RGB-only) configuration. YOLOv5-Nano is a highly compact version of the popular YOLO (You Only Look Once) object detection framework, designed specifically for real-time applications on edge devices. In this baseline, only the RGB input stream is used, and no fusion or thermal information is provided. This model serves as a reference for evaluating the performance of RGB-based detection in low-light conditions, where RGB imagery may lack sufficient contrast or clarity.

The second baseline is the YOLOv5-Nano (Thermal-only) model. It uses the same architecture as the RGB-only variant but is trained solely on thermal images. This allows us to understand the benefits and limitations of thermal data when used independently for object detection. While thermal imaging performs better than RGB under complete darkness, it may miss important structural or contextual information that RGB provides under partial lighting.

The third baseline is a Fixed Fusion Model without adaptive sensor selection. This configuration uses the same mid-level fusion strategy as the proposed method but applies a static weighting scheme where RGB and thermal features are always combined with equal importance. This model demonstrates how fusion alone improves performance over single-modal approaches, while also revealing the limitations of static fusion in dynamic lighting conditions. The fixed fusion model uses channel-wise concatenation followed by shared convolution layers, identical to the proposed method, but without the dynamic weighting block.

The fourth baseline is a Lightweight RGB-T Fusion Model that uses a naive early fusion strategy where RGB and thermal images are concatenated at the input level. Although this design is computationally efficient, it lacks the semantic awareness of mid-level fusion and fails to fully capture modality-specific features. This baseline helps demonstrate the impact of fusion level and architectural design on final detection accuracy.

The fifth and final comparison is the Proposed Model without Sensor Selection, in which the attention-based adaptive weighting block is removed. This version uses mid-level fusion as described but applies fixed equal weights to both modalities. By comparing this model with the full proposed system, we isolate and evaluate the contribution of real-time sensor selection to detection robustness and environmental adaptability.

Each of the baseline models was trained and evaluated under the same conditions as the proposed method, using identical datasets, input resolutions, and training parameters. The evaluation metrics applied include mAP@0.5, mAP@0.5:0.95, FPS, and model size, allowing for a balanced comparison across accuracy, efficiency, and real-time performance dimensions.

As shown in Table 1, the proposed thermal-RGB fusion model with adaptive sensor selection outperforms all baseline models across both accuracy and real-time performance metrics.

In Table 1 and Table 2, the column "Parameter Count (M)" refers to the total number of trainable parameters in the model, measured in millions (M). For example, a model with 1.9M parameters has 1.9 million trainable parameters. This value indicates the model's complexity, memory requirements, and computational demand.

The terms mAP@0.5 (%) and mAP@0.5:0.95 (%) represent different evaluation metrics used to assess object detection performance. mAP@0.5 (%) is Mean Average Precision at an IoU (Intersection over Union) threshold of 0.5. This measures the model's ability to correctly detect objects with at least 50% overlap with the ground truth.

mAP@0.5:0.95 (%) is Mean Average Precision calculated by averaging the precision scores over multiple IoU thresholds, ranging from 0.5 to 0.95 with a step size of 0.05. This is a stricter and more comprehensive evaluation that measures both detection accuracy and localization precision across varying levels of overlap.

We conducted ablation studies on mid-level fusion, adaptive sensor selection, attention mechanisms, and post-training optimizations to validate design choices and identify which components most influence performance in real-time drone surveillance.

The results of these ablation studies are summarized in Table 2, highlighting the contribution of each component to the overall system performance.

The first experiments compared the full thermal-RGB fusion model with RGB-only and thermal-only versions. While RGB performed poorly in dark scenes and thermal lacked structural detail, mid-level fusion significantly outperformed both by achieving higher mAP and better localization, confirming the benefits of combining both modalities.

To assess the adaptive sensor selection module, we compared the full model with a variant using fixed equal weights. While fixed fusion outperformed single modalities, it failed to adapt to changes like dynamic lighting or occlusion. In contrast, the adaptive module consistently improved accuracy and stability, especially under transitional lighting such as twilight or artificial illumination.

We also analyzed the role of attention in fusion, focusing on the Squeeze-and-Excitation (SE) block. Removing it reduced the model's focus on relevant features, while its inclusion improved performance, especially in cluttered scenes or with small targets, demonstrating that even lightweight attention enhances feature integration without notable overhead.

To assess the trade-off between performance and efficiency, we tested the full model before and after applying quantization and structured pruning. Post-training static quantization to 8-bit integer precision reduced model size by approximately 60% and improved inference speed by up to 35%, with a negligible drop in accuracy (less than 1.5% mAP). Structured pruning further reduced FLOPs and model size by removing redundant channels in convolutional layers, although a slight accuracy degradation (approximately 2%) was observed in heavily pruned versions. These results confirm that the proposed model remains suitable for edge deployment even after significant compression.

The simplified model without attention and sensor selection performed better than single-modality baselines but worse than the full model, confirming their importance. The ablation studies show that fusion, selection, attention, and optimization components together enable accurate and efficient night-time drone detection.

Table 1. Performance comparison of baseline models and the proposed method (Evaluated on input resolution of 416×416 pixels)

Model	Modality	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS	Model size (MB)	Parameter Count (M)
YOLOv5-Nano (RGB-only)	RGB	64.7	41.2	48	7.4	1.9
YOLOv5-Nano (Thermal-only)	Thermal	68.5	44.6	47	7.4	1.9
MobileNetV2 Early Fusion	RGB + Thermal	71.3	46.8	44	8.2	2.3
Fixed Mid-Level Fusion (no adapt)	RGB + Thermal	74.6	49.0	43	8.7	2.5
Proposed Model (w/o sensor adapt)	RGB + Thermal	**76.1**	**50.5**	**42**	**9.0**	**2.6**
Proposed Model (full)	RGB + Thermal	**79.4**	**53.7**	**41**	**9.3**	**2.8**

Table 2. Ablation study results: impact of individual components on performance

Model	Fusion type	Sensor selection	Attention module	Quantized / Pruned	mAP@0.5 (%)	mAP@0.5:0.95 (%)	FPS	Model Size (MB)	Parameter Count (M)
RGB-only Baseline	None	No	No	No	64.7	41.2	48	7.4	1.9
Thermal-only Baseline	None	No	No	No	68.5	44.6	47	7.4	1.9
Mid-Level Fusion (no adapt, no attention)	Mid-Level	No	No	No	74.6	49.0	43	8.7	2.5
Mid-Level Fusion + Attention (no adapt)	Mid-Level	No	Yes	No	75.5	50.1	42	8.9	2.6
Mid-Level Fusion + Sensor Selection (no attention)	Mid-Level	Yes	No	No	76.1	50.5	42	9.0	2.6
Full Proposed Model	**Mid-Level**	**Yes**	**Yes**	**No**	**79.4**	**53.7**	**41**	**9.3**	**2.8**
Full Model (Quantized + Pruned)	**Mid-Level**	**Yes**	**Yes**	**Yes**	**78.0**	**52.4**	**55**	**3.8**	**2.8**

4. Conclusion

This paper presents a lightweight and adaptive object detection framework for night-time drone surveillance based on thermal-RGB fusion. The system integrates mid-level feature fusion, adaptive sensor weighting, and model compression to address the challenges of low-light aerial perception, all validated using publicly available datasets and standard PC hardware.

First, the mid-level fusion strategy effectively combines RGB and thermal modalities, preserving critical spatial and contextual information while improving detection performance under complex lighting.

Second, the adaptive sensor selection module enhances robustness by dynamically adjusting the importance of each modality based on scene conditions. This allows the model to prioritize thermal features in darkness and RGB cues in partially lit environments without manual intervention.

Third, ablation studies confirmed that removing any core module—fusion, sensor selection, or attention—leads to a drop in accuracy, validating the necessity of each design component.

Fourth, quantization and pruning significantly reduced the model size and improved inference speed, demonstrating the feasibility of deployment on resource-limited platforms.

The proposed system achieves a strong balance between accuracy and efficiency, making it well-suited for real-time aerial surveillance. Its modular design and reliance on publicly accessible resources further support adaptability and reproducibility in real-world applications.

Acknowledgement

This work was supported by Baku State University and Azerbaijan Technical University.

References

He, X., Tang, C., Zou, X., & Zhang, W. (2023). Multispectral object detection via cross-modal conflict-aware learning. In 31st ACM International Conference on Multimedia, Ottawa, Canada, October 29 - November 3 (pp. 1465 - 1474). https://doi.org/10.1145/3581783.3612651

Kim, T., Shin S., Yu, Y., Kim, H.G., & Ro., Y.M. (2024). Causal mode multiplexer: A novel framework for unbiased multispectral pedestrian detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 16-22 June 2024 (pp. 26774-26783). https://doi.org/10.1109/CVPR52733.2024.02529

Li, Q., Zhang, C., Hu, Q., Fu, H., & Zhu, P. (2023). Confidence-aware fusion using Dempster-Shafer theory for multispectral pedestrian detection. IEEE Transactions on Multimedia, 25, 3420-3431. http://doi.org/10.1109/TMM.2022.3160589

Li, R., Xiang, J., Sun, F., Yuan, Y., Yuan, L., & Gou, S. (2024). Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection. IEEE Transactions on Multimedia, 26, 852-863. http://doi.org/10.1109/TMM.2023.3272471

Lim, D.Y., Jin, I.J., & Bang, I.C. (2023). Heat-vision based drone surveillance augmented by deep learning for critical industrial monitoring. Scientific Reports, 13(1), 1-12. https://doi.org/10.1038/s41598-023-49589-x

Liu, Y., & Jiang, W. (2024). Frequency mining and complementary fusion network for RGB-infrared object detection. IEEE Geoscience and Remote Sensing Letters, 21, 1–5. http://doi.org/10.1109/LGRS.2024.3448493

Ozcan, A., & Cetin, O. (2022). A novel fusion method with thermal and RGB-D sensor data for human detection. IEEE Access, 10, 66831–66841. https://doi.org/10.1109/ACCESS.2022.3185402

Sun, J., Yin, M., Wang, Z., Xie, T., & Bei, S. (2024). Multispectral object detection based on multilevel feature fusion and dual feature modulation. Electronics, 13(2), 443. https://doi.org/10.3390/electronics13020443

Sun, X., Zhu, Y., & Huang, H. (2025). Specificity-guided cross-modal feature reconstruction for RGB-Infrared object detection. IEEE Transactions on Intelligent Transportation Systems, 26(1), 950-961. http://doi.org/10.1109/TITS.2024.3495028

Zhang, X., Cao, S.-Y., Wang, F., Zhang, R., Wu, Z. & Zhang, X. (2024). Rethinking early-fusion strategies for improved multispectral object detection. IEEE Transactions on Intelligent Vehicles, 1–15. (in press) https://doi.org/10.1109/TIV.2024.3462488

Zhang, Y., Zeng, W., Jin, S., Qian, C., Luo, P., & Liu, W. (2025). When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset. In 18th European Conference on Computer Vision, Milan, Italy, September 29–October 4 (pp. 430–448). https://doi.org/10.1007/978-3-031-73195-2_25

Wang, Q., Chi, Y., Shen, T., Song, J., Zhang, Z., & Zhu, Y. (2022). Improving RGB-infrared pedestrian detection by reducing cross-modality redundancy," In IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16-19 October 2022 (pp. 526-530) http://doi.org/10.1109/ICIP46576.2022.9897682

Zhu, Y., Sun, X., Wang, M., & Huang, H. (2023). Multi-modal feature pyramid transformer for RGB-infrared object detection. IEEE Transactions on Intelligent Transportation Systems, 24(9), 9984-9995. http://doi.org/10.1109/TITS.2023.3266487