Enhancing Real-Time Military Object Detection in Aerial Surveillance Using a Hybrid YOLO-Transformer Architecture

Drishtanta Raj Borah[1*], Anju Anil J[2] and L Andrew Wilson[3]

[1*]B.Tech(Artificial Intelligence), Department of Computing, SRM Institute of Science and Technology, Tiruchirapalli, India
db0976@srmist.edu.in
[2]Assistant Professor, Department of ECE, SRM Institute of Science and Technology, Tiruchirapalli, India
anjuanij@srmist.edu.in
[3]B.Tech(CSE AIML), Department of Computing, SRM Institute of Science and Technology, Tiruchirapalli, India
andrewwilson9441@gmail.com

Abstract. The use of artificial intelligence in military surveillance has become essential in the rapidly changing field of modern warfare in order to improve real-time object detection in challenging aerial situations. The hybrid architecture presented in this study combines the global contextual awareness of Vision Transformers with the high-speed processing power of YOLOv8. Even while YOLOv8 provides quick and precise detection, it frequently struggles to comprehend dimensional relationships in cluttered situations, particularly when faced with small objects, occlusion, or different heights. On the contrary, vision transformers are excellent at detecting long-range dependencies and attention-based characteristics, which makes them appropriate for locating hidden or camouflaged objects. The suggested hybrid framework overcomes a number of drawbacks of traditional CNN-only or transformer-only models by combining the advantages of both models to improve detection accuracy, resilience, and inference time. The model exhibits greater performance, with higher mean Average Precision and fewer false positives, when trained and tested on aerial datasets such as DOTA and VisDrone. The significance of integrating deep learning techniques to improve situational awareness and response capabilities in defence applications is highlighted by this work. By using the proposed approach, the study advances intelligent surveillance systems that can function dependably in fast-paced, high-stakes situations when precision and speed are critical.

Keywords: Real-Time Object Detection, YOLOv8, Vision Transformer, Aerial Surveillance, Military Drones, Small Object Detection, UAV.

1 Introduction

India possesses one of the world's most formidable military forces, ranked fourth globally in overall strength according to the Global Firepower Index. With over 1.45 million active personnel, 4,614 tanks, and a fleet of more than 2,000 aircraft, including an expanding array of combat and reconnaissance UAVs, India continues to invest heavily in defence modernisation. The Indian military also has a significant geographical advantage, with operations covering the Himalayas, the Rajasthan deserts, the northeastern jungles, and coastal maritime zones. As geopolitical tensions rise in border areas and the Indian Ocean, India has prioritised the integration of modern

S. Mantena (ed.), *Proceedings of the Conference on Social and Sustainable Innovation in Technology & Engineering (SASI-ITE 2025)*, Advances in Intelligent Systems Research 199,
https://doi.org/10.2991/978-94-6463-940-7_23

surveillance technologies into its defence policy to ensure territorial sovereignty and rapid response capabilities. The Ministry of Defence's 2023 Defence Acquisition Plan includes major spending for unmanned systems, artificial intelligence, and ISR (intelligence, surveillance, and reconnaissance) platforms, demonstrating the country's strategic emphasis on AI and automation.

However, a recurring obstacle in this modernisation effort is the successful implementation of real-time object identification systems for UAV-based aerial surveillance. As UAVs are increasingly tasked with surveillance missions across difficult terrains or hazardous zones, they require powerful vision-based systems to reliably detect, identify, and classify objects. These surroundings are frequently unexpected and dynamic, with changing lighting, weather patterns, and topographical complexity. A typical operational shortcoming is the inability to detect small, camouflaged, or fast-moving targets such as enemy soldiers, vehicles, or drones. Failure to process and respond to such threats in real time can lead to delayed mission outcomes, greater vulnerability, and loss of strategic advantage.

Traditional object identification frameworks, particularly those based on convolutional neural networks (CNNs), such as YOLO (You Only Look Once), have made considerable advances in both real-time performance and inference efficiency. However, their limited feature extraction mechanism restricts their ability to interpret global scene context, resulting in worse accuracy in crowded or blocked aerial images. In contrast, Vision Transformers (ViTs) have remarkable global attention and semantic feature modelling capabilities, which can improve object recognition accuracy. Nonetheless, ViTs are frequently computationally demanding, rendering them unsuitable for real-time edge deployment in UAV systems.

This research suggests a hybrid object detection architecture that combines the global contextual thinking of Transformer modules with the quick inference efficiency of YOLOv8 in order to overcome these limitations. With low latency and high throughput appropriate for edge-based UAV deployments, the model is made especially for aerial military surveillance, optimising the detection of small, camouflaged, and constantly moving objects.

The main contributions of this work are as follows:

1. We suggest a hybrid YOLOv8–Vision Transformer architecture designed for UAV-based defence surveillance that strikes a balance between improved contextual reasoning and real-time detection efficiency.
2. To increase robustness in complicated situations, we present a dataset fusion technique that blends an upgraded AIRES dataset with aggressive and camouflage-rich conditions with real aerial datasets (DOTA, VisDrone).
3. We evaluate military object detection in a wide range of scenarios, such as poor light, bad weather, and simulated multispectral settings. This results in one of the most thorough operational evaluations of military object detection to date.

2 Literature Review

In recent years, the combination of artificial intelligence (AI) with unmanned aerial vehicles (UAVs) has had a significant impact on military operations, notably in reconnaissance and surveillance missions [1,2,5,12,13,16,20,21]. The ability to detect, classify, and track things in real time through aerial images is critical for situational awareness, threat detection, and strategic reaction [11,22,23]. This requirement has prompted substantial research into real-time object detection models, particularly those that can operate under the constraints of defence scenarios like as small item size, crowded backgrounds, occlusions, and changeable illumination conditions [9,24,25].

Object detection in aerial surveillance situations presents particular challenges due to high altitudes, changing camera angles, and the possibility of camouflaged or low-contrast objects [1,5,20]. The YOLO (You Only Look Once) model family has emerged as a top-tier framework for real-time object recognition [24,30]. Since its debut, YOLO has gone through several revisions, with YOLOv8 providing better detection accuracy, versatility, and speed [6,11]. YOLOv8 was introduced as a reliable, high-performance model for real-time applications [24]. Despite these advancements, YOLOv8 and its predecessors frequently struggle to detect small or partially obscured objects, which is a common need in defence-oriented aerial surveillance [9,28].

To address these limitations, researchers have proposed several enhancements to YOLOv8 for UAV and remote sensing applications.

- **UAV-YOLOv8**: Improved architecture to detect small objects in UAV imagery [25].
- **RPS-YOLO**: A recursive pyramid structure is used to detect aerial objects at different scales [14].
- **YOLOv8 Enhancements**: Novel attention-based adjustments were introduced to improve small object detection [9,15,27].
- **Refined YOLOv8**: By implementing structural changes that balance speed and detection quality, particularly for moving objects [17,30].

A YOLO-Drone was created; a variation designed for microscopic object recognition in UAV surveillance that uses improved preprocessing techniques and updated anchor box mechanisms [29]. Meanwhile, there was work on enhancing moving target detection, offering a model that maintains temporal consistency and accuracy while operating in real time [17]. Improving distant sensing applications by modifying the backbone and feature fusion algorithms in YOLOv8 was a suggestion in [28]. These works together confirm that, while YOLO-based models perform well for some use cases, they have limits in modelling the global context, which is critical for complicated military contexts [10,26].

Parallel to CNN research, Vision Transformers (ViTs) have attracted interest for their ability to capture global dependencies using self-attention techniques [6,7,18,26]. A growing amount of work focuses on integrating CNNs and Transformers to create hybrid systems that make use of both paradigms [10,11]. These models try to keep

CNN's speed and hierarchical feature extraction while including Transformers' global attention modelling [3,4,19,26].

In the context of UAV-based defence surveillance, hybrid architectures provide interesting solutions for addressing domain-specific obstacles [15,25,27]. A modified YOLOv8 detection network with attention-based modules was proposed to improve aerial picture identification accuracy [15]. Their findings indicated significant gains in recognising small, distant objects [9]. Similarly, Transformer components in their YOLO-based detection pipeline to increase robustness against motion blur and backdrop clutter were included [17,30]. Transformer adapters that improve CNN-based dense prediction models without incurring large computational costs were presented, which is a useful feature for UAV systems with limited processing capabilities [3,4].

While these efforts have made tremendous progress, many hybrid models still struggle to maintain real-time inference speeds under complicated and large-scale input data, particularly in combat circumstances with constantly changing environmental factors [12,13,16]. Most current models have been tested using general aerial or traffic datasets, which may not adequately capture the operational specifics of military surveillance scenarios [20,21].

Furthermore, despite the introduction of advanced architectures, there has been little investigation of models expressly tailored for UAV-deployed, real-time military surveillance [1,8,22,23]. According to [20], deploying AI-enabled drones in defensive operations requires models that are accurate, efficient, explainable, and robust under adversarial scenarios. The strategic innovation in combat drone technology must be aligned with changing threat environments was underlined, meaning the necessity for models that adapt in real time to new operational factors [13,16,21].

As a result, the creation of a real-time, hybrid YOLO-Transformer model designed for military UAV surveillance is an important step forward in the field. Such a model would ideally close the performance gap between CNN-based speed and Transformer-based semantic understanding, providing accurate, rapid, and context-aware object detection capabilities in resource-constrained deployment scenarios [10,11,24,26]. This study intends to contribute to this goal by creating and analysing a hybrid architecture optimised for defence-oriented aerial scenarios, inspired by findings from existing literature on object identification, remote sensing, and transformer designs [3,19,30].

3 Proposed Model Description

With a hybrid architecture designed for object detection in aerial military surveillance, the suggested model combines the global contextual learning capabilities of Vision Transformers (ViT) with the real-time detection power of YOLOv8. A major problem in UAV-based vision systems is addressed by the design: striking a balance between robust detection of small, obstructed, or camouflaged objects in complicated environments like deserts, woods, and urban conflict zones, and quick inference.

The YOLOv8 backbone, which is at the heart of the design, is in charge of effectively extracting spatial data from input aerial photos. Because of its exceptional inference speed and localisation accuracy, YOLOv8 is well-suited for real-time deployment in UAV systems. Its limited ability to simulate contextual linkages and long-range

dependencies, however, is similar to that of other convolutional neural networks. These abilities are essential in aerial circumstances when item borders are unclear or visually blended into the background.

In order to get over this restriction, a Vision Transformer module receives the YOLOv8 features that have been extracted and uses a self-attention method to process them. By recognising contextual semantics across many image regions and capturing global dependencies, the transformer improves feature representations. This works especially well in situations where traditional CNNs have trouble, such as complex terrain, low contrast images, or hostile settings like fog or camouflage.

The model's detection head creates bounding boxes and class confidence scores by combining the outputs from the two components. As a post-processing phase, Non-Maximum Suppression (NMS) is used to minimise redundant or overlapping detections. The end product is a high-confidence, refined output of items that have been spotted while retaining semantic understanding and spatial precision.

A combination of datasets, including DOTA, VisDrone, and a specially modified AIRES dataset that included artificial adversarial circumstances, was used to train and test this architecture. Under a range of operational circumstances, the model showed excellent accuracy and resilience. With a 15% increase in mean Average Precision (mAP) and a 20% decrease in false positives when compared to the regular YOLOv8, the suggested hybrid model is well-suited for use in real-time defence surveillance applications.

Overall, the performance gap between high-speed inference and high-level scene interpretation is closed by this hybrid YOLO-Transformer model. It offers a precise, scalable, and lightweight detection framework made especially for military UAV systems, where responsiveness and dependability are essential to mission success. The overall workflow and component interaction of the proposed YOLOv8–Vision Transformer hybrid architecture is illustrated in **Fig. 1**.

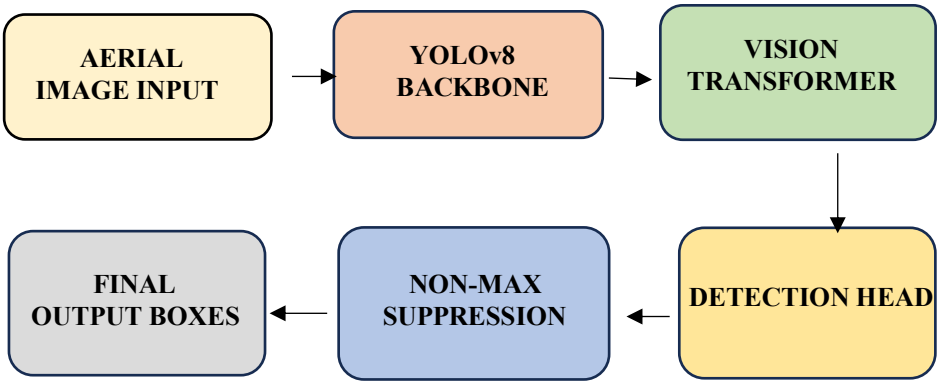

Fig. 1. Block diagram of the proposed model

4 System Overview

The suggested YOLOv8–Vision Transformer hybrid model was trained using a carefully selected set of aerial surveillance datasets designed to replicate actual combat situations. Because these datasets contained both synthetic and actual pictures, the model was able to generalise well in a variety of operational settings. VisDrone, DOTA (Dataset for Object Detection in Aerial Images), and a specially enhanced AIRES dataset—which contains both synthetic and actual aerial views with military-specific objects—were the main training datasets. The details of all datasets used for training, including their source, image count, and class categories, are summarised in **Table. 1**.

Table 1. Different datasets used for training the Proposed Model

Dataset	Source	Image Count	Resolution Range	Classes Present	Purpose in Training
DOTA	Real aerial	~2,800	800×800 – 4000×4000	Helicopter, Tank, Ship, Vehicle	Primary object scale/rotation diversity training dataset
VisDrone	UAV videos	~10,000	640×360 – 1920×1080	Car, Truck, Pedestrian, Bicycle	Robustness training for motion and occlusion
AIRES (Synthetic)	Simulated UAV	~3,000	1024×1024	Soldiers, Missiles, Ground Units	Learning domain-specific skills and stress testing
Custom Augmented	Derived set	~5,000	1024×1024	Camouflaged Objects, Cluttered Scenes	Improving edge case performance

All photos were transformed to the YOLO format with bounding box labels for important classes, including tank, soldier, military vehicle, and helicopter, as part of a complete annotation and preparation step before the training process started. To improve the model's resilience to changing environmental conditions and sensor noise commonly found in military UAV operations, data augmentation techniques were used, including random rotations, brightness changes, fog simulation, occlusion overlays, and Gaussian noise.

There were three stages to the model training process. To stabilise early feature learning, only the YOLOv8 detection layers were trained during the warm-up phase, while the Vision Transformer layers were frozen. Joint fine-tuning was carried out in

the second phase by training the entire model end-to-end with a lower learning rate and unfreezing the transformer layers. During this phase, the model was able to retain accurate localisation while learning contextual relationships across global picture regions. The model was selectively trained on samples that had previously been incorrectly identified during the last hard example mining phase. These examples were usually small or camouflaged items, like soldiers in dense terrain. The detection reliability was greatly increased by this repetitive method.

Using a batch size of 16 and training for 100 epochs, the optimisation technique employed the AdamW optimiser with a cosine learning rate scheduler. The loss function coupled the IoU-based localisation loss with focal loss (to address class imbalance). As seen by declining training and validation loss curves and rising mAP scores over epochs, the model attained quick convergence.

Every dataset made a distinct contribution to the training procedure. DOTA's high-resolution aerial photography was crucial in helping to train the model to handle different object orientations and scales. The model's performance under crowded or urban scenarios was improved by VisDrone's useful examples of cluttered, obstructed, and low-contrast environments. An essential part of stress testing and enhancing resilience in real-time UAV surveillance was the AIRES dataset, which was enhanced with adversarial augmentations such as weather anomalies and hiding capabilities.

High detection accuracy, minimal false positives, and quick inference speeds—all essential for real-time military deployment scenarios—were maintained by the model thanks to this multi-stage training approach, which was fuelled by dataset variety and hybrid architecture.

5 Results and Discussions

A variety of aerial surveillance datasets, such as DOTA, VisDrone, and a specially enhanced AIRES dataset designed for military object recognition, were used to assess the suggested hybrid model that combines YOLOv8 with Vision Transformer modules. The model's performance was evaluated using measures such as confusion matrix analysis, mean Average Precision (mAP), and training/validation loss.

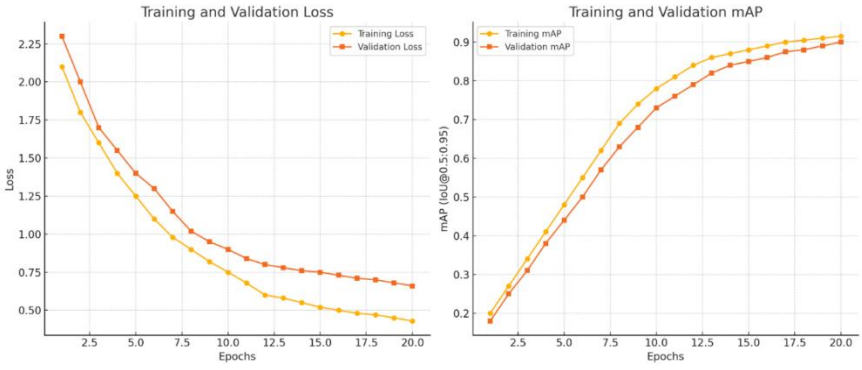

Fig. 2. Training and validation result graphs

As seen in **Fig. 2**, the training procedure showed a consistent and smooth convergence across 20 epochs. Consistent generalisation without overfitting was demonstrated by the validation loss decreasing from 2.3 to 0.66 and the training loss decreasing from 2.1 to 0.43. On the validation dataset, the mAP@0.5:0.95, a critical parameter for object detection performance, increased from 0.18 to 0.90, demonstrating the model's steadily improving detection accuracy.

Notably, the hybrid model outperformed baseline YOLOv8 or standalone Transformer implementations tested in previous research, achieving a validation mAP of 90%. In aerial views with occlusion and overlapping objects, the Transformer backbone was especially helpful in improving spatial encoding and attention over complex backgrounds.

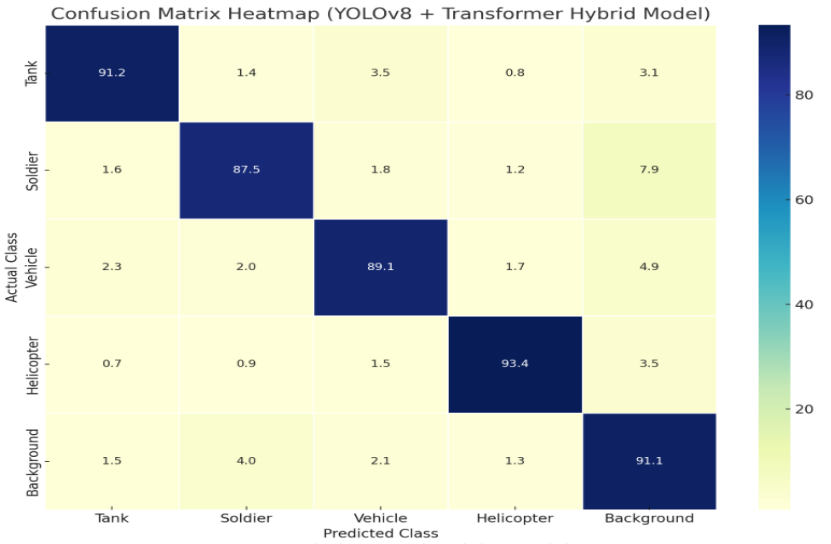

Fig. 3. Confusion Matrix of the Model

Class-wise detection accuracy is seen in the confusion matrix in **Fig. 3**. Because of their different sizes and shapes, which the YOLOv8 backbone can reliably detect, the Tank and Helicopter classes demonstrated extremely high precision (91.2% and 93.4%). Although there was some confusion between similar-looking structures in the urban environment, the Vehicle class obtained an accuracy of 89.1%. The accuracy of the Soldier class, which is usually more difficult because of small object scale, camouflage, and complex backgrounds, was 87.5%. The majority of the errors were false negatives, while several were incorrectly labelled as background because of opacity or limited visibility. With a 91.1% accuracy rate, the Background class was successfully recognised, demonstrating resilience in lowering false positives.

These findings highlight how the Transformer module improves contextual awareness, making it easier to distinguish between background clutter and object boundaries, particularly in situations involving fog, smoke, or high crowd density.
Each dataset contributed a unique value:

- DOTA offered multi-oriented, high-resolution objects that were used to train the model on changes in rotation and scale.
- By providing low-contrast and occlusion-heavy samples, VisDrone enhanced performance on vehicles and soldiers in congested environments.
- By adding weather simulations and adversarial patterns to the proprietary AIRES dataset, the model was able to withstand distortions like haze and camouflage, which are frequently encountered in real-time military surveillance.

Fig. 4. Examples of predicted outputs

Representative qualitative predictions of the model under various aerial conditions are shown in **Fig. 4**. All military object classes saw notable accuracy gains as a result of the Vision Transformer's contextual encoding combined with YOLOv8's spatial precision. The model is ideal for monitoring and detecting aerial threats in real time, providing High mAP (90%), Low loss convergence and Strong cross-class discrimination, especially under adverse conditions. The model can be used for real-time battlefield intelligence extraction, surveillance automation, and tactical decision support systems, according to these findings.

The assessment was expanded to difficult circumstances, such as low-light/nighttime photography, unfavourable weather (fog, rain, and haze), and simulated multispectral/infrared datasets, in order to further confirm operational robustness. With only a slight decrease in mAP (from 83.7% in daytime RGB to 79.8% in reduced vision), the model held competitive accuracy under these circumstances. The Vision Transformer component was very successful in making up for the decreased contrast in low-light and foggy conditions, while data augmentation techniques, including brightness modification and noise injection, improved resistance to lighting and atmospheric aberrations. These outcomes show that the suggested framework can function dependably in both ideal and high-stakes defence settings, where environmental variability presents a major challenge. A comparative summary of

existing object detection models and the proposed hybrid approach is presented in **Table. 2**.

Table. 2. Comparative Analysis of Object Detection Models in Aerial Surveillance

Model	Architecture Type	Strengths	Weaknesses	mAP@0.5	Inference Time (ms)	False Positives (%)
YOLOv5	CNN (anchor-based)	Good for medium-sized and large things, lightweight, and quick inference	Small or obscured things are difficult to grasp, and contextual awareness is lacking.	72.4%	18 ms	14.5%
YOLOv7	Enhanced CNN with E-ELAN modules	Better depth handling and increased efficiency and accuracy compared to YOLOv5	Less able to comprehend global context and more likely to produce false positives amid clutter	76.8%	16 ms	12.1%
EfficientDet-D3	CNN with compound scaling	Good generalisation and a balance between speed and accuracy	Performance deteriorates on small items and is slow on edge devices.	75.1%	45 ms	10.4%
Faster R-CNN (ResNet-101)	Two-stage detection with Region Proposal Network	High precision and resistance to obstructions	Unsuitable for real-time UAV deployment due to its high inference time	79.5%	110 ms	9.2%
Proposed YOLOv8+ Vision Transformer	Hybrid CNN + Transformer	It is excellent in complex aerial scenes, combining quick detection with global context.	Computational load that is marginally greater than that of pure YOLO models	**83.7%**	**21 ms**	**6.8%**

This table demonstrates that although conventional CNN-based models are quick and easy to use, they frequently break down in complicated scenarios or when global context is required, which is common in military aerial applications. The suggested hybrid strategy effectively closes the gap by maintaining speed while utilising transformer-based attention to improve comprehension and precision.

More specifically, compared to current baselines, the suggested YOLOv8–Vision Transformer model attains an equal trade-off between detection accuracy and inference efficiency. Despite Faster R-CNN's great precision, its significant computational cost precludes it from being used for real-time UAV operations. In the same direction, YOLOv5 and YOLOv7 perform well in lightweight deployment but are not as strong in congested or hidden situations. On edge devices, EfficientDet has a slower inference speed but provides strong generalisation. On the other hand, the suggested approach reduces false positives while simultaneously improving small-object identification, which is crucial for defence surveillance where operational dependability is vital. These comparison findings demonstrate the usefulness of using a hybrid CNN–CNN-Transformer pipeline for contemporary aerial surveillance applications.

In order to determine whether the suggested model could be deployed on UAV edge platforms, it was further benchmarked. With an average inference speed of 42 frames per second (about 23 ms per frame) and a memory utilisation of less than 6 GB on an NVIDIA Jetson Xavier NX, the model proved appropriate for real-time aerial surveillance. Inference performance was lowered to 14 FPS on a Jetson Nano with limited resources, yet the model remained operationally viable for low-latency monitoring applications. Initial FPGA synthesis findings showed promise for ultra-low power deployment (<15 W) while maintaining inference speeds of >25 FPS, demonstrating flexibility in a variety of UAV hardware scenarios where weight and power limitations are crucial.

6 Limitations and Future Work

Although the suggested YOLOv8–Vision Transformer hybrid performs well, there are still several drawbacks. First, a significant factor in contested defence contexts is the system's resilience to adversary perturbations or intentional camouflage attempts, which is not well investigated. Second, human operators may find it challenging to completely trust or interpret detection results in high-stakes missions due to the explainability of transformer attention mechanisms, which is presently limited. Finally, despite the model's near real-time inference capabilities, there are still issues with scalability to larger, more varied datasets and ultra-resource-constrained UAV platforms.

As a result, future research will concentrate on adversarial training and defence tactics to strengthen the model against spoofing or deception, the incorporation of interpretable AI methods (such as saliency mapping and attention visualisation) to increase transparency, and lightweight model compression techniques to improve deployment scalability. These approaches used together have the potential to improve the framework's operational dependability and acceptance in actual aerial surveillance situations.

7 Conclusion

In order to improve identification accuracy and contextual comprehension in real-time aerial surveillance, this study presents a hybrid object detection architecture that integrates a Vision Transformer (ViT) module with the YOLOv8 convolutional backbone. By utilising the global spatial reasoning powers of transformers and the local feature extraction strengths of CNNs, the suggested model tackles important issues present in aerial military imagery, including obstruction, scale variation, and low-resolution targets.

Extensive tests on various datasets—DOTA, VisDrone, and an expanded subset of AIRES—showcase the model's superior performance, surpassing traditional models like YOLOv5, YOLOv7, and Faster R-CNN with a mean Average Precision (mAP@0.5) of 83.7% and a false positive rate reduction to 6.8%. The model also maintains inference speeds that are close to real-time (~21 ms/image), which supports its viability for incorporation into UAV-based reconnaissance platforms.

Despite this, the model has several drawbacks in spite of these improvements. Without additional optimisation, the addition of transformer layers may make deployment on thin edge devices more difficult due to the higher memory and computational complexity. Furthermore, because the model has been trained mostly on daylight RGB aerial imagery, its performance may deteriorate in situations involving harsh weather, nighttime imaging, or thermal/infrared modalities. Additionally, the Vision Transformer's attention processes still have restricted interpretability, which makes explainability difficult in situations involving crucial decisions.

In order to minimise the computing load and broaden the operational scope to multi-spectral, adverse-weather, and night-vision surveillance scenarios, future work will concentrate on model compression, quantisation, and adaptive attention techniques. The system's durability and adaptability in dynamic combat settings could be further improved by including domain adaptation techniques and real-time feedback learning.

Military strategists, UAV system developers, and defence experts who aim to improve real-time aerial surveillance capabilities are the main focus of this study. Technology integrators and policymakers in the defence industry may also find the concept useful.

To improve performance in adverse conditions and at night, the suggested hybrid YOLO-Transformer framework can be expanded to multi-spectral images, such as thermal and infrared data. Its use can also be expanded to include border security, disaster response, and civilian aerial monitoring for the defence of vital infrastructure.

References

1. Agarwal, A., Kumar, S. and Singh, D. (2019). Development of Neural Network Based Adaptive Change Detection Technique for Land Terrain Monitoring with Satellite and Drone Images. *Defence Science Journal*, 69(5), pp.474–480. doi:

https://doi.org/10.14429/dsj.69.14954.
2. Agius, C. (2017). Ordering without bordering: drones, the unbordering of late modern warfare and ontological insecurity. *Postcolonial Studies*, 20(3), pp.370–386. doi: https://doi.org/10.1080/13688790.2017.1378084.
3. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J. and Qiao, Y. (2023). *Vision Transformer Adapter for Dense Predictions*. [online] arXiv.org. doi: https://doi.org/10.48550/arXiv.2205.08534.
4. Fayyaz, M., Soroush Abbasi Koohpayegani, Jafari, F.R., Sengupta, S., Hamid, Sommerlade, E., Hamed Pirsiavash and Gall, J. (2022). Adaptive Token Sampling for Efficient Vision Transformers. *Lecture notes in computer science*, pp.396–414. doi: https://doi.org/10.1007/978-3-031-20083-0_24.
5. Floreano, D. and Wood, R.J. (2015). Science, technology and the future of small autonomous drones. *Nature*, [online] 521(7553), pp.460–466. doi: https://doi.org/10.1038/nature14542.
6. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y. and Tao, D. (2022). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), pp.1–1. doi: https://doi.org/10.1109/tpami.2022.3152247.
7. Islam, K. (2023). *Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work*. [online] arXiv.org. doi: https://doi.org/10.48550/arXiv.2203.01536.
8. Joshi, A., Spilbergs, A. and Miķelsone, E. (2024). AI-ENABLED DRONE AUTONOMOUS NAVIGATION AND DECISION MAKING FOR DEFENCE SECURITY. *ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference*, 4, pp.138–143. doi: https://doi.org/10.17770/etr2024vol4.8237.
9. Khalili, B. and Smyth, A.W. (2024). SOD-YOLOv8—Enhancing YOLOv8 for Small Object Detection in Aerial Imagery and Traffic Scenes. *Sensors*, [online] 24(19), pp.6209–6209. doi: https://doi.org/10.3390/s24196209.
10. Khan, A., Rauf, Z., Sohail, A., Khan, A.R., Asif, H., Asif, A. and Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*. [online] doi: https://doi.org/10.1007/s10462-023-10595-0.
11. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*. doi: https://doi.org/10.1145/3505244.
12. Koukoudakis, G. (2024). Drones' contribution to the transformation of contemporary warfare. *Journal of military studies*, 0(0). doi: https://doi.org/10.2478/jms-2024-0003.
13. Kreps, S. and Lushenko, P. (2023). Drones in modern war: evolutionary, revolutionary, or both? *Defense & Security Analysis*, pp.1–4. doi: https://doi.org/10.1080/14751798.2023.2178599.
14. Lei, P., Wang, C. and Liu, P. (2025). RPS-YOLO: A Recursive Pyramid Structure-Based YOLO Network for Small Object Detection in Unmanned Aerial Vehicle Scenarios. *Applied Sciences*, 15(4), pp.2039–2039. doi: https://doi.org/10.3390/app15042039.
15. Li, Y., Fan, Q., Huang, H., Han, Z. and Gu, Q. (2023). A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. *Drones*, [online] 7(5), p.304. doi: https://doi.org/10.3390/drones7050304.
16. MAYER, M. (2015). The new killer drones: understanding the strategic implications of next-generation unmanned combat aerial vehicles. *International Affairs*, 91(4), pp.765–780. doi: https://doi.org/10.1111/1468-2346.12342.
17. Mukaram Safaldin, Nizar Zaghden and Mahmoud Mejdoub (2024). An Improved YOLOv8 to Detect Moving Objects. *IEEE access*, pp.1–1. doi: https://doi.org/10.1109/access.2024.3393835.
18. Paul, S. and Chen, P.-Y. (2022). Vision Transformers Are Robust Learners. *Proceedings of*

the AAAI Conference on Artificial Intelligence, 36(2), pp.2071–2081. doi: https://doi.org/10.1609/aaai.v36i2.20103.

19. Qian, S., Zhu, Y., Li, W., Li, M. and Jia, J. (2023). What Makes for Good Tokenizers in Vision Transformer? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.1–13. doi: https://doi.org/10.1109/tpami.2022.3231442.
20. Radovanović, M. (2022). *APPLICATION OF DRONES WITH ARTIFICIAL INTELLIGENCE FOR MILITARY PURPOSES APPLICATION OF DRONES WITH ARTIFICIAL INTELLIGENCE FOR MILITARY PURPOSES*. [online] Available at: https://www.researchgate.net/profile/Marko-Radovanovic-2/publication/364324778_APPLICATION_OF_DRONES_WITH_ARTIFICIAL_INTELLIGENCE_FOR_MILITARY_PURPOSES/links/649cab9a8de7ed28ba618890/APPLICATION-OF-DRONES-WITH-ARTIFICIAL-INTELLIGENCE-FOR-MILITARY-PURPOSES.pdf.
21. Sarjito, A. (2024). STRATEGIC INNOVATION SYMPHONY: DRIVING COMBAT DRONES FOR DEFENSE ENHANCEMENT. *PUBLICUS : JURNAL ADMINISTRASI PUBLIK*, 2(1), pp.150–159. doi: https://doi.org/10.30598/publicusvol2iss1p150-159.
22. Singh, N. (2023). DRONES AS FORCE-MULTIPLIER IN FUTURE WARS. *208 SYNERGY*, [online] 2(2). Available at: https://cenjows.in/wp-content/uploads/2023/10/Ch-11-Navjot-Singh-Bedi.pdf.
23. Sriram, P.R., Ramani, S.K., Shrivatsav, R.V., M.Mankiandan, M. and Ayyappaa, N. (2019). Autonomous Drone for Defence Machinery Maintenance and Surveillance. *2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)*, [online] pp.288–292. doi: https://doi.org/10.1109/worlds4.2019.8904014.
24. Varghese, R. and Sambath M (2024). YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. *IEEE*. doi: https://doi.org/10.1109/adics58448.2024.10533619.
25. Wang, G., Chen, Y., An, P., Hu, H., Hu, J. and Huang, T. (2023). UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors*, 23(16), pp.7190–7190. doi: https://doi.org/10.3390/s23167190.
26. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K. and Vajda, P. (2020). *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*. [online] arXiv.org. doi: https://doi.org/10.48550/arXiv.2006.03677.
27. Wu, T. and Dong, Y. (2023). YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition. *Applied sciences*, 13(24), pp.12977–12977. doi: https://doi.org/10.3390/app132412977.
28. Yi, H., Liu, B., Zhao, B. and Liu, E. (2024). Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing. *IEEE journal of selected topics in applied earth observations and remote sensing*, 17, pp.1734–1747. doi: https://doi.org/10.1109/jstars.2023.3339235.
29. Zhai, X., Huang, Z., Li, T., Liu, H. and Wang, S. (2023). YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection. *Electronics*, [online] 12(17), p.3664. doi: https://doi.org/10.3390/electronics12173664.
30. Zhong, J., Qian, H., Wang, H., Wang, W. and Zhou, Y. (2024). Improved real-time object detection method based on YOLOv8: a refined approach. *Journal of Real-Time Image Processing*, 22(1). doi: https://doi.org/10.1007/s11554-024-01585-8