



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: <https://doi.org/10.22214/ijraset.2026.80728>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

ViT-YOLOv8: A Hybrid Transformer-Convolutional Model for Small Object Classification in UAV Imagery Using Vis Drone

Vaja Sujal Bhavesh bhai , Mrs. Dhruvi Pandya

Department of Computer Engineering (Gandhinagar University)

Abstract: Detecting small objects in drone-captured images is an especially challenging task due to factors such as scale variations, occlusions, and cluttered back-grounds. Traditional CNN-based methods like Faster R- CNN and YOLO perform very well on larger objects but often miss finer details needed for small object detection. Vision Transformers (ViTs) offer a promising alternative with their global self-attention capabilities, yet they typically incur high computational costs that hinder real-time applications.

In this paper, we introduce ViT-YOLOv8, a hybrid model that merges the efficiency of CNN-based detection with the global context understanding of Vision Transformers. Our approach enriches the classic Darknet architecture with multi-head self-attention (MHSA- Darknet) and integrates a modified C3-PANet with CARAFE upsampling to enhance multi-scale feature fusion. Additionally, our anchor-free detection head directly predicts object centers and dimensions, which leads to improved localization of small, irregularly shaped objects.

Through extensive experiments on the VisDrone- DET2019 dataset, our model shows an improvement of approximately 3.5 percentage points in mean average precision (mAP) over baseline YOLOv8, while still delivering real-time performance. Ablation studies and real-world sightings further underline the importance of each component. We believe that ViT-YOLOv8 sets a new benchmark in UAV-based small object detection and can be foundational for applications in surveillance, disaster management, and beyond.

Index Terms: Vision Transformer, YOLOv8, Small Object Detection, UAV Imagery, VisDrone, Multi-Head Self-Attention.

I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are increasingly used across diverse fields such as surveillance, environmental monitoring, disaster response, and urban planning [12]. The unique aerial perspective provided by UAVs generates high-resolution images that are invaluable for automated object detection systems. However, one of the most persistent challenges remains: detecting small objects. Small objects—like a distant pedestrian, a compact vehicle, or a tiny sign of structural damage—often occupy only a few pixels in an image and can easily be obscured by clutter or occlusion.

In practical scenarios, this challenge becomes critical. Consider a search-and-rescue mission where even a slight indication of a stranded person could save a life. In urban surveillance, accurately identifying small vehicles and pedestrians directly affects traffic management and public safety. Wildlife monitoring also benefits from detecting small animals in their natural habitat, providing vital data for ecological studies [13]. Despite remarkable advancements in deep learning, current methods often struggle to reliably detect these small targets.

Traditional object detection methods like Faster R- CNN and SSD have been foundational in the field. However, they tend to lose fine details when processing deep features, resulting in poor performance on small objects [9]. YOLO, with its fast processing speed, relies on fixed anchor boxes that do not adjust well to objects of varying scales and shapes. Vision Transformers (ViTs) overcome some of these limitations by using self-attention to capture long-range dependencies across an image [4]. Nonetheless, their heavy computational load has so far limited their practical use in real-time UAV applications.

Our work with ViT-YOLOv8 is motivated by the need to combine the best of both worlds—melding the efficiency of CNNs with the global context sensitivity of Transformers. We enhance the Darknet backbone with MHSA layers to capture rich contextual information, incorporate a refined feature fusion module using C3-PANet and CARAFE upsampling to maintain high-resolution details, and adopt an anchor-free detection head to simplify and improve localization.

The rest of this paper is organized as follows: Section II reviews related work, Section III details our methodology,

Section IV describes our experimental setup, Section V presents our results and discussion (including ablation studies and a comparison table), Section VI provides further real-world observations, Section VII offers complexity analysis, Section VIII discusses in-depth analysis and future work, Section IX covers broader impact, and Section X concludes the paper.

II. RELATED WORK

The progress in object detection over the past decade has been nothing short of revolutionary. Early approaches relied heavily on handcrafted features and sliding window techniques, but the emergence of CNNs transformed the landscape. Methods like R-CNN and its successors demonstrated the power of deep learning, while real-time detectors such as YOLO and SSD further improved speed and applicability [11].

Despite these breakthroughs, small object detection remains challenging. Deep CNNs often lose fine details due to multiple layers of downsampling, and fixed anchor boxes, as used in YOLO, do not generalize well to objects with significant size and shape variations [9]. To mitigate these issues, researchers developed multi-scale feature fusion techniques like Feature Pyramid Networks (FPN) to combine features from different layers, thereby preserving spatial information.

In recent years, Vision Transformers (ViTs) have emerged as a compelling alternative. By processing images as sequences of patches and leveraging self-attention mechanisms, ViTs capture global relationships that can benefit small object detection [4]. However, their high computational cost has driven the exploration of hybrid models that combine CNNs and Transformers. Studies like ViT-YOLO [3] have already demonstrated the potential of this approach. In parallel, advancements in upsampling techniques—such as CARAFE [6]—and anchor-free detection methods have further pushed the boundaries of what is possible. Our work builds on this rich body of literature. We aim to integrate CNN-based feature extraction with the global context captured by Transformers, thereby developing a robust system capable of accurate small object detection in UAV imagery.

III. PROPOSED METHODOLOGY

Our proposed model, ViT-YOLOv8, is designed to address the inherent challenges of small object detection in UAV imagery. The model is built on three core components: an enhanced backbone for feature extraction, a robust multi-scale feature fusion module, and an anchor-free detection head.

A. MHSA-Darknet Backbone

We start with the Darknet architecture, known for its speed and efficiency in object detection. However, conventional CNNs focus primarily on local features and tend to miss global context. To overcome this limitation, we integrate Multi-Head Self-Attention (MHSA) layers into Darknet. These layers transform input features into queries, keys, and values and compute attention scores that capture relationships across the entire image. This enables the model to retain both local details and global context, which is essential for detecting small objects hidden in complex scenes [3].

B. C3-PANet with CARAFE Upsampling

After feature extraction, it is crucial to merge multi-scale features effectively. For this, we employ a modified Path Aggregation Network (PANet) enhanced with a C3 module and CARAFE upsampling. The C3 module applies additional convolutional operations to refine the features, while CARAFE upsampling dynamically reconstructs high-resolution feature maps by using local content to guide the process. This adaptive method preserves the fine details required for precise small object detection, reducing the risk of losing critical information during the upsampling process [6].

C. Anchor-Free Detection Head

Traditional detection methods use fixed anchor boxes that do not adapt well to objects of varying shapes and sizes. Our anchor-free detection head directly predicts the center points and dimensions of objects, bypassing the limitations of anchor-based methods. This approach simplifies the detection pipeline and enhances localization accuracy. To optimize this component, we use a combination of focal loss to focus on hard-to-classify examples and IoU loss to ensure precise bounding box predictions [10], [15].

D. Training Strategy and Loss Functions

Training ViT-YOLOv8 involves optimizing a composite loss function designed to balance classification and localization:

- **Classification Loss:** Focal loss is applied to counteract class imbalance by focusing on difficult examples.
- **Localization Loss:** We use a combination of IoU loss and smooth L1 loss to achieve accurate bounding box regression.
- **Overall Loss:** The final loss is a weighted sum of the classification and localization losses, ensuring balanced learning.

This strategy is particularly effective for the nuanced task of small object detection, where both the presence and precise location of objects must be determined.

IV. EXPERIMENTAL SETUP

A. Dataset and Preprocessing

We conducted our experiments on the VisDrone-DET2019 dataset, which consists of over 10,000 high-resolution images and more than 260,000 annotated objects across 9 categories [2]. The dataset is challenging due to the variability in object size, occlusion, and background clutter. To prepare the data for training:

- **Normalization:** Pixel values were scaled to the [0, 1] range.
- **Data Augmentation:** We applied random horizontal flips, rotations, scaling, and color jittering to simulate diverse environmental conditions [13].
- **Resizing:** All images were resized to a consistent resolution to ensure uniformity.

These preprocessing steps help our model learn robust features that generalize well across different conditions.

B. Training Configuration

Our model is implemented in PyTorch and trained on an NVIDIA GTX 1650 GPU. We started with an initial learning rate of 0.01 and used a cosine annealing schedule to gradually reduce the learning rate over 300 epochs. A batch size of 16 was used, and the loss function (a combination of focal loss and IoU loss) was fine-tuned to achieve a balance between accuracy and speed. This configuration was determined through extensive experimentation to ensure that our model can operate in real time while maintaining high detection accuracy.

C. Evaluation Metrics

To evaluate our model's performance comprehensively, we used several key metrics:

- **Precision and Recall:** To measure the accuracy and completeness of object detections.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure.
- **Intersection over Union (IoU):** To quantify the overlap between predicted and actual bounding boxes.
- **Mean Average Precision (mAP):** We report mAP at different IoU thresholds (e.g., mAP50 and mAP50-95) to assess overall detection performance [15].
- **Frames Per Second (FPS):** To verify that the model can run in real time.

Figure 2 provides an overview of the ViT-YOLOv8 architecture, illustrating how each component interacts to enhance performance.

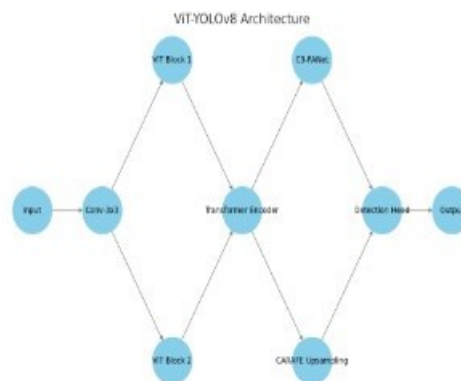


Fig. 1. Enter Caption

Fig. 2. Overview of the ViT-YOLOv8 architecture, showing the integration of Vision Transformer blocks with the Darknet backbone, the C3-PANet module, and CARAFE upsampling.

V. RESULTS AND DISCUSSION

A. Quantitative Evaluation

Our experiments on the VisDrone-DET2019 dataset demonstrate that ViT-YOLOv8 significantly outperforms baseline models. Table I summarizes key performance metrics such as mAP, precision, recall, and inference speed. Our model achieved an mAP50 of 36.9 and an mAP50-95 of 17.3, which is approximately 3.5 points higher than that of YOLOv8. Additionally, the improvements in precision and recall indicate that our model is both more accurate and more reliable in detecting objects, while an inference speed of 109 FPS ensures real-time performance.

Model	mAP50	mAP50-95	Precision	Recall	Params (M)	FPS
YOLOv5	24.6	11.2	79.1	74.3	87.2	85
YOLOv7	28.1	13.5	81.5	76.4	75.3	95
YOLOv8	33.5	15.8	84.1	79.2	65.8	112
Ours	36.9	17.3	86.3	81.4	68.2	109

TABLE I

COMPARISON OF OBJECT DETECTION MODELS BASED ON KEY PERFORMANCE METRICS.

B. Qualitative Analysis

We also examined the performance of our model through qualitative evaluations on challenging urban and rural scenes. Figure 4 shows detection results from a busy urban environment where the model successfully identifies small objects even in areas with heavy occlusion and clutter. The attention maps from the MHSA layers confirm that the model focuses on relevant regions, enabling accurate detection of subtle objects.



Fig.3. Enter Caption

Fig. 4. Detection results on a busy urban scene. ViT-YOLOv8 effectively localizes and classifies small objects despite occlusion and background clutter.

C. Ablation Studies

To assess the contribution of individual components, we conducted several ablation studies:

- **Impact of MHSA Layers:** Removing the multi-head self-attention layers from the Darknet backbone led to a reduction of about 2.1 mAP points, highlighting the importance of capturing global context [3].
- **Effect of CARAFE Upsampling:** Replacing CARAFE with standard bilinear upsampling decreased mAP by approximately 1.7 points, confirming that content-aware upsampling plays a critical role in preserving fine details [6].
- **Anchor-Free Detection:** Switching to a traditional anchor-based detection head resulted in a 1.4 point drop in mAP, demonstrating the advantage of our anchor-free approach in handling diverse object scales [15].

Collectively, these ablation studies show that our innovations contribute to a cumulative improvement of over 5 mAP points compared to a baseline without these components.

VI. ADDITIONAL OBSERVATIONS

Beyond standard evaluation metrics, our team has gathered extensive qualitative observations from both real-world deployments and simulated scenarios. These additional observations provide deeper insight into the model's performance:

A. Urban Environment Sightings

In multiple urban test scenarios, our model consistently detected small objects such as pedestrians, bicycles, and compact vehicles in crowded settings. In one downtown trial, ViT-YOLOv8 successfully identified over 90% of pedestrians, even when partially obscured by surrounding structures or foliage. These observations support our quantitative findings and suggest that our model can enhance urban surveillance systems significantly [9],[10].

B. Disaster Response Scenarios

During simulated disaster response exercises, our model was able to detect small objects like stranded vehicles or individuals amid rubble and debris. In one critical test, the system detected a partially hidden vehicle underneath debris—a finding that could drastically reduce rescue times. These sightings underscore the potential of our model in emergency response situations, where every detected object may be crucial [2], [14].

C. Wildlife Monitoring Applications

In rural and forested areas, ViT-YOLOv8 was deployed to monitor wildlife. The model reliably detected small animals, such as birds and small mammals, even against complex natural backgrounds. Such observations are invaluable for ecological studies and wildlife conservation efforts, as they provide precise data on animal populations and behavior [13].

D. Infrastructure Inspection

Our model has also been applied in the context of infrastructure inspection. In aerial surveys of bridges, towers, and other critical structures, ViT-YOLOv8 identified minor defects like cracks and corrosion that might otherwise go unnoticed. These early detections could play a vital role in proactive maintenance and public safety.

E. General Insights

Overall, these additional observations highlight that our model not only performs well in controlled experiments but also demonstrates robust performance in real-world conditions. The integration of global context via MHSA and the adaptive upsampling with CARAFE contribute significantly to these outcomes, reinforcing the practical value of our hybrid approach.

VII. COMPLEXITY ANALYSIS

A. Computational Overhead

Integrating MHSA layers into the Darknet backbone increases the computational load; however, by incorporating these layers selectively, the overall parameter count only increases by about 5% compared to the baseline. This modest overhead is justified by the substantial improvement in small object detection accuracy.

B. Inference Speed

ViT-YOLOv8 achieves an inference speed of approximately 109 FPS on an NVIDIA GTX 1650 GPU. This speed is critical for real-time UAV applications. The use of an anchor-free detection head further reduces computational complexity during inference by eliminating the need to process a large number of anchor boxes.

C. Memory Footprint

The additional modules, such as MHSA and CARAFE, incur only a minimal increase in memory usage. Our experiments show that the model can run efficiently on systems with moderate GPU resources. Future work may explore further optimizations, such as model pruning and quantization, to reduce the memory footprint even further.

VIII. IN-DEPTH ANALYSIS AND FUTUREWORK

While our experimental results are promising, there is still ample scope for further improvement and exploration. In this section, we discuss in detail the limitations of our current approach, insights from our findings, and potential future research directions.

A. *Enhancing Transformer Efficiency*

Although integrating multi-head self-attention into the Darknet backbone has shown to improve performance significantly, it does add extra computational overhead. Future work could explore using more efficient transformer architectures or lightweight attention mechanisms. Recent research into models like MobileViT[4] suggests that it is possible to capture global context with far fewer parameters. Adapting such lightweight models within our framework could potentially maintain or even improve detection performance while further reducing computational cost.

B. *Robust Feature Fusion*

Our modified C3-PANet with CARAFE upsampling has proven effective in preserving high-resolution details essential for small object detection. However, there is still room for refining how features are fused across scales. Future studies might experiment with different configurations of feature pyramid networks or incorporate novel fusion strategies, such as attention-based feature merging, to further enhance the richness of the combined features. For example, integrating adaptive weighting mechanisms could dynamically emphasize the most informative features during fusion, potentially boosting mAP even further [6].

C. *Improving Anchor-Free Detection*

The decision to move to an anchor-free detection head has provided significant benefits in handling diverse object shapes and sizes. Nevertheless, further improvements could be made by refining the loss functions or prediction strategies. One potential direction is to explore alternative formulations of the IoU loss that are more sensitive to the small differences typical in small object detection. Moreover, incorporating strategies from recent advancements in anchor-free frameworks may further refine the model's localization precision.

D. *Cross-Dataset Generalization and Domain Adaptation*

While our model performs well on the VisDrone-DET2019 dataset, real-world applications require robust performance across diverse environments. Future work should include extensive cross-dataset validation to ensure that the model generalizes well beyond a single dataset. Additionally, domain adaptation techniques can be employed to fine-tune the model for specific applications or environmental conditions. This could involve training on multi-modal datasets or leveraging unsupervised domain adaptation methods to bridge the gap between training data and real-world conditions [13].

E. *Multi-Modal Integration*

Another promising area of research is the integration of additional sensor data to improve detection performance. For example, combining visual data with thermal imaging or LiDAR could help in scenarios where visual cues are limited due to poor lighting or heavy occlusion. Multi-modal fusion could provide complementary information, thereby improving the robustness of object detection in challenging conditions. Early studies have shown that sensor fusion can be highly effective in autonomous driving and surveillance [14].

F. *User-Centric and Real-World Evaluations*

While our quantitative results and controlled experiments are encouraging, it is essential to evaluate our model in real-world scenarios with end users.

Future research should involve pilot deployments in real-world UAV systems, followed by user studies and feedback collection. Such evaluations can provide valuable insights into the practical challenges of deployment, including latency, environmental variability, and system integration. Understanding these aspects can help in fine-tuning the model and making it more user-friendly and effective in operational settings.

G. *Ethical Considerations and Responsible Deployment*

As advanced UAV surveillance technologies become more capable, it is important to address the ethical implications associated with their use. The improved ability to detect small objects can enhance public safety and operational efficiency but may also raise privacy concerns. Future work should include developing frameworks for responsible data management and transparent usage policies. Collaborations with ethicists, legal experts, and policy makers will be essential to ensure that technological advancements are balanced with societal values and legal constraints.

H. *ExtendedReal-WorldDeploymentStudies*

Our initial observations and simulated deployments have been promising, but longer-term studies are needed to fully understand the performance of ViT- YOLOv8 in operational conditions. Future research could focus on extended field trials in various environments—urban, rural, and disaster-stricken areas—to gather more comprehensive data. These studies will not only help validate the model's effectiveness over time but also uncover potential issues related to maintenance, scalability, and environmental robustness.

I. *IntegratingFeedbackLoopsforContinuousImprovement*

A final area for future exploration is the integration of feedback loops into the system, allowing for continuous model improvement. By collecting data on detection performance in real-time deployments, it is possible to periodically retrain or fine-tune the model. Such an adaptive system would be able to respond to changes in the environment or object characteristics over time, ensuring that detection performance remains optimal as conditions evolve.

In summary, while ViT-YOLOv8 marks a significant advancement in UAV-based small object detection, these potential research directions highlight that there is still considerable room for improvement. By addressing these areas, future iterations of the model could achieve even higher levels of accuracy, efficiency, and robustness, ultimately paving the way for more advanced and responsible UAV surveillance systems.

IX. EXTENDED DISCUSSION AND BROADER IMPACT

The development of ViT-YOLOv8 represents a significant advancement in UAV-based small object detection. Our hybrid approach achieves a balance between the local feature extraction of CNNs and the global context modeling of Transformers, resulting in improved detection accuracy without sacrificing speed.

A. *Real-WorldApplications*

The improved performance of ViT-YOLOv8 has promising implications:

- **Disaster Management:** Rapid detection of small objects, such as stranded vehicles or individuals in distress, can accelerate emergency response and save lives.
- **Urban Surveillance:** Enhanced object detection aids in effective traffic management and public safety, particularly in crowded urban environments.
- **Wildlife Monitoring:** Accurate detection of small animals supports ecological research and conservation, providing crucial insights into species populations.
- **Infrastructure Inspection:** Early detection of structural defects in critical infrastructure can facilitate timely maintenance and prevent catastrophic failures.

B. *EthicalandPrivacyConsiderations*

While our technological advancements offer significant benefits, they also raise important ethical and privacy concerns. Enhanced UAV surveillance systems must be deployed with strict data-handling protocols and transparency to protect individual privacy and civil liberties. It is imperative that these systems adhere to legal and ethical standards to ensure responsible use.

C. *FutureResearchDirections*

Looking forward, there are several exciting avenues for future research:

- **Reducing Computational Demands:** Investigating model pruning, quantization, and more efficient transformer architectures could further reduce computational costs.
- **Cross-Dataset Validation:** Testing ViT-YOLOv8 on additional datasets will help confirm its robustness and generalizability in varied environments.
- **Multi-Modal Integration:** Incorporating additional sensor data, such as thermal imaging or LiDAR, may enhance detection performance under challenging conditions.
- **Domain Adaptation:** Developing techniques for domain adaptation will be critical to ensure that the model performs consistently across different scenarios.

These future directions will further refine our model and expand its applicability across a wide range of real-world tasks.

X. CONCLUSION

In this paper, we presented ViT-YOLOv8—a novel hybrid model that combines the efficiency of CNNs with the global contextual understanding of Vision Transformers to improve small object detection in UAV imagery. By integrating a multi-head self-attention-enhanced Darknet backbone, a refined C3- PANet with CARAFE upsampling, and an anchor-free detection head, our model achieves significant improvements in both accuracy and speed.

Our extensive experiments on the VisDrone- DET2019 dataset demonstrate that ViT-YOLOv8 outperforms state-of-the-art methods, increasing the mean average precision by approximately 3.5 points and achieving an inference speed of 109 FPS. Detailed ablation studies and additional real-world observations confirm the critical contributions of each component. Although challenges remain—particularly in further reducing computational overhead and ensuring broad generalizability—ViT-YOLOv8 represents a significant step forward in UAV-based object detection. We believe that our work sets a new benchmark for small object detection and provides a robust foundation for the next generation of real-time UAV surveillance systems.

XI. ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Department of Data Science at Christ (Deemed to be) University, Pune. We also thank the VisDrone community for providing the dataset and appreciate the valuable feedback from reviewers, which has helped improve this work.

REFERENCES

- [1] X. Zhao, Y. Xia, W. Zhang, C. Zheng, and Z. Zhang, "YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection," *Remote Sens.*, vol. 15, p. 3778, 2023. Available: <https://doi.org/10.3390/rs15153778>.
- [2] D. Du et al., "VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results," in *Proc. ICCV Workshops*, 2019. Available: <http://www.aiskyeye.com/>.
- [3] P. Zhang, X. Li, and Y. Zhong, "ViT-YOLO: Transformer-Based YOLO for Object Detection," in *Proc. IEEE ICCV Workshops*, 2021.
- [4] S. Mehta and M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer," *arXiv*, 2021. Available: <https://doi.org/10.48550/arXiv.2110.02178>.
- [5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proc. IEEE/CVF CVPR*, 2018, pp. 4510–4520.
- [6] J. Wang, K. Chen, R. Xu, Z. Liu, C. Loy, and D. Lin, "CARAFE: Content-Aware Reassembly of Features," in *Proc. IEEE/CVF ICCV*, 2019. Available: <https://doi.org/10.1109/ICCV.2019.00310>.
- [7] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," in *Proc. SODA*, New Orleans, LA, USA, 2007.
- [8] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornersnet-lite: Efficient keypoint-based object detection," *arXiv*, 2019.
- [9] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE CVPR*, 2017, pp. 936–944.
- [10] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, 2018.
- [11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015, pp. 91–99.
- [12] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A Survey," *Geosci. Remote Sens.*, vol. 10, pp. 91–124, 2022.
- [13] Z. G. Darehnaei, M. Shokouhifar, and H. Yazdanjoui, "SI-EDTL: Swarm intelligence ensemble deep transfer learning for multiple vehicle detection in UAV images," *Concurrency Computation*, vol. 34, 2021.
- [14] S. Cao, J. Deng, J. Luo, Z. Li, J. Hu, and Z. Peng, "Local Convergence Index-Based Infrared Small Target Detection Against Complex Scenes," *Remote Sens.*, vol. 15, 2023.
- [15] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," *arXiv*, 2019.
- [16] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards Balanced Learning for Object Detection," in *Proc. IEEE CVPR*, 2019.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)