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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.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 objectsbut often miss finer details needed for small object detection. Vision Transformers (ViTs) offer a promising alternativewiththeirglobalself-attentioncapabilities,yet theytypicallyincurhighcomputationalcoststhathinder real-time applications.***

In this paper, we introduce ViT-YOLOv8, a hybrid model that merges the efficiency of CNN-based detec- tion with the global context understanding of Vision Transformers. Our approach enriches the classic Dark- net 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 di- rectlypredictsobjectcentersanddimensions,whichleads 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)overbaselineYOLOv8,whilestilldeliv- ering real-time performance. Ablation studies and real- worldsightingsfurtherunderlinetheimportanceofeach component. We believe that ViT-YOLOv8 sets a new benchmarkinUAV-basedsmallobjectdetectionandcan 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

UnmannedAerialVehicles(UAVs)areincreasingly used across diverse fields such as surveillance, en- vironmental 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 re- mains: detecting small objects. Small objects—like a distantpedestrian,acompactvehicle,oratinysign ofstructuraldamage—oftenoccupyonlyafewpixels in an image and can easily be obscured by clutter or occlusion.

In practical scenarios, this challenge becomes criti- cal. Consider a search-and-rescue mission where even aslightindicationofastrandedpersoncouldsave 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].Despiteremarkableadvancementsindeep 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,theytendtolosefinedetailswhenprocessing deep features, resulting in poor performance on small objects [9]. YOLO, with its fast processing speed, reliesonfixedanchorboxesthatdonotadjustwell to objects of varying scales and shapes. Vision Trans- formers (ViTs) overcome some of these limitations by usingself-attentiontocapturelong-rangedependencies across an image [4]. Nonetheless, their heavy compu- tational 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 informa- tion,incorporatearefinedfeaturefusionmoduleusing C3-PANetandCARAFEupsamplingtomaintainhigh- 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,

SectionIVdescribesourexperimentalsetup,SectionV presents our results and discussion (including ablation studies and a comparison table), Section VI provides furtherreal-worldobservations,SectionVIIoffers 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 approachesreliedheavilyonhandcraftedfeatures andslidingwindowtechniques,buttheemergence ofCNNstransformedthelandscape.Methodslike R-CNNanditssuccessorsdemonstratedthepower of deep learning, while real-time detectors such as YOLO and SSD further improved speed and appli- cability[11].

Despite these breakthroughs, small object detection remains challenging. Deep CNNs often lose fine de- tailsduetomultiplelayersofdownsampling,andfixed anchorboxes,asusedinYOLO,donotgeneralizewell toobjectswithsignificantsizeandshapevariations[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.

Inrecentyears,VisionTransformers(ViTs)have emergedasacompellingalternative.Byprocessing imagesassequencesof patchesandleveragingself- attentionmechanisms,ViTscaptureglobalrelation- ships that can benefit small object detection [4]. How- ever,theirhighcomputationalcosthasdriventhe exploration of hybrid models that combine CNNs and Transformers.StudieslikeViT-YOLO[3]havealready demonstratedthepotentialofthisapproach.Inpar- allel,advancementsinupsamplingtechniques—such asCARAFE[6]—andanchor-freedetectionmethods havefurtherpushedtheboundariesofwhatispossible. Our work builds on this rich body of literature. WeaimtointegrateCNN-basedfeatureextractionwith theglobalcontextcapturedbyTransformers,thereby developingarobustsystemcapableofaccuratesmallobjectdetectioninUAVimagery.

III. PROPOSED METHODOLOGY

Ourproposedmodel,ViT-YOLOv8,isdesigned to address the inherent challenges of small object detectionin UAVimagery.Themodelisbuiltonthree core components: an enhanced backbone for feature extraction, a robust multi-scale feature fusion module, and an anchor-free detection head.

A. MHSA-DarknetBackbone

We start with the Darknet architecture, known forits speed and efficiency in object detection. How-ever, conventional CNNs focus primarily on local features and tend to miss global context. To overcome thislimitation,weintegrateMulti-HeadSelf-Attention (MHSA) layers into Darknet. These layers transform input features into queries, keys, and values and com- pute attention scores that capture relationships across theentireimage.Thisenablesthemodeltoretainboth local details and global context, which is essential for detectingsmallobjectshiddenincomplexscenes[3].

B. C3-PANetwithCARAFEUpsampling

After feature extraction, it is crucial to merge multi- scale features effectively. For this, we employ a mod- ified Path Aggregation Network (PANet) enhanced withaC3moduleandCARAFEupsampling.The C3moduleappliesadditionalconvolutionaloperations 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 requiredforprecisesmallobjectdetection,reducingtherisk of losing critical information during the upsampling process [6].

C. Anchor-FreeDetectionHead

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

D. TrainingStrategyandLossFunctions

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

- Classification Loss: Focal loss is applied to counteract class imbalance by focusing on dif- ficult examples.
- LocalizationLoss: We use a combination ofIoU loss and smooth L1 loss to achieve accurate bounding box regression.
- OverallLoss:Thefinallossisaweightedsumof theclassificationandlocalizationlosses,ensuring balanced learning.

Thisstrategyisparticularlyeffectiveforthenuanced taskofsmallobjectdetection,whereboththeпresence and precise location of objects must be determined.

IV. EXPERIMENTAL SETUP

A. DatasetandPreprocessing

We conducted our experiments on the VisDrone- DET2019 dataset, which consists of over 10,000 high- resolution images and more than 260,000 annotated objectsacross9categories[2].Thedatasetischalleng- ing 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 hori- zontal flips, rotations, scaling, and color jittering tosimulatediverseenvironmentalconditions[13].
- Resizing:Allimageswereresizedtoaconsistent resolution to ensure uniformity.

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

B. TrainingConfiguration

Our model is implemented in PyTorch and trained on an NVIDIA GTX 1650 GPU. We started with an initiallearningrateof0.01andusedacosineannealing scheduletograduallyreducethelearningrateover300 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 andspeed.Thisconfigurationwasdeterminedthrough extensiveexperimentationtoensurethatourmodelcan operate in real time while maintaining high detection accuracy.

C. EvaluationMetrics

To evaluate our model's performance comprehen- sively, 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., mAP50and mAP50-95) to assess overall detection per- formance[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,illustratinghoweachcomponentinteracts to enhance performance.

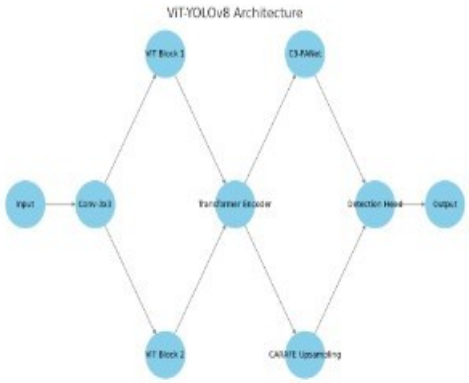


Fig.1.EnterCaption

Fig. 2.Overview of the ViT-YOLOv8 architecture, showing theintegrationofVisionTransformerblockswiththeDarknetbackbone,the C3-PANet module, and CARAFE upsampling.

V. RESULTS AND DISCUSSION

A. QuantitativeEvaluation

OurexperimentsontheVisDrone-DET2019dataset demonstrate that ViT-YOLOv8 significantly outper- formsbaselinemodels.TableIsummarizeskeyperfor- mance metrics such as mAP, precision, recall, and in- ference speed. Our model achieved an mAP50 of 36.9 andanmAP50-95of17.3,whichisapproximately3.5 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 detectingobjects,whileaninferencespeedof109FPS 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
ViT-YOLOv8 (Ours)	36.9	17.3	86.3	81.4	68.2	109

TABLEI
COMPARISONOFOBJECTDETECTIONMODELSBASEDONKEYPERFORMANCE METRICS.

B. QualitativeAnalysis

We also examined the performance of our model through qualitative evaluations on challenging urban and rural scenes. Figure 4 shows detection resultsfrom a busy urban environment where the model successfullyidentifiessmallobjectseveninareaswith 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.EnterCaption

Fig. 4.Detection results on a busy urban scene. ViT-YOLOv8effectively localizes and classifies small objects despite occlusionand background clutter.

C. AblationStudies

Toassessthecontributionofindividualcomponents, we conducted several ablation studies:

- Impact of MHSA Layers: Removing the multi- head self-attention layers from the Darknet back- bone 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 de- creased mAP by approximately 1.7 points, con- firming that content-aware upsampling plays a critical role in preserving fine details [6].
- Anchor-Free Detection**:** Switching to a tradi- tionalanchor-baseddetectionheadresultedina1.4 point drop in mAP, demonstrating the ad- vantage of our anchor-free approach in handling diverse object scales [15].

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

VI. ADDITIONAL OBSERVATIONS

Beyondstandardevaluationmetrics,ourteamhas gathered extensive qualitative observations from both real-world deployments and simulated scenarios. These additional observations provide deeper insight into the model's performance:

A. UrbanEnvironmentSightings

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

B. DisasterResponseScenarios

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 vehicleunderneathdebris—afindingthatcoulddrasti- cally reduce rescue times. These sightings underscore thepotentialofourmodelinemergencyresponsesitu- ations,whereeverydetected objectmaybecrucial[2], [14].

C. WildlifeMonitoringApplications

In rural and forested areas, ViT-YOLOv8 was de- ployed to monitor wildlife. The model reliably de- tected small animals, such as birds and small mam- mals,evenagainstcomplexnaturalbackgrounds.Such observations are invaluable for ecological studies and wildlife conservation efforts, as they provide precise data on animal populations and behavior [13].

D. InfrastructureInspection

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 mightotherwisegounnoticed.Theseearlydetections couldplayavitalroleinproactivemaintenanceand public safety.

E. GeneralInsights

Overall, these additional observations highlight that our model not only performs well in controlled ex- periments but also demonstrates robust performancein real-world conditions. The integration of global context via MHSA and the adaptive upsampling with CARAFE contribute significantly to these outcomes, reinforcingthepracticalvalueofourhybridapproach.

VII. COMPLEXITY ANALYSIS

A. ComputationalOverhead

Integrating MHSA layers into the Darknet back- bone increases the computational load; however, by incorporating these layers selectively, the overall pa- rameter count only increases by about 5% comparedto the baseline. This modest overhead is justified by the substantial improvement in small object detection accuracy.

B. InferenceSpeed

ViT-YOLOv8 achieves an inference speed of ap- proximately 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 re- duces computational complexity during inference by eliminating the need to process a large number of anchor boxes.

C. MemoryFootprint

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. EnhancingTransformerEfficiency

Although integrating multi-head self-attention into the Darknet backbone has shown to improve perfor- mance significantly, it does add extra computational overhead. Future work could explore using more ef- ficient transformer architectures or lightweight atten- tion 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 perfor- mance while further reducing computational cost.

B. RobustFeature Fusion

OurmodifiedC3-PANetwithCARAFEupsampling has proven effective in preserving high-resolution de- tails essential for small object detection. However, there is still room for refining how features are fused across scales. Future studies might experiment with differentconfigurationsoffeaturepyramidnetworksor 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 emphasizethemostinformativefeaturesduringfusion, potentially boosting mAP even further [6].

C. ImprovingAnchor-FreeDetection

The decision to move to an anchor-free detection head has provided significant benefits in handling diverse object shapes and sizes. Nevertheless, further improvementscouldbemadebyrefiningthelossfunc- tions or prediction strategies. One potential directionis to explore alternative formulations of the IoU loss that are more sensitive to the small differences typi- cal in small object detection. Moreover, incorporating strategies from recent advancements in anchor-free frameworksmayfurtherrefinethemodel'slocalization precision.

D. Cross-Dataset Generalization and Domain Adap- tation

While our model performs well on the VisDrone- DET2019 dataset, real-world applications require ro- bust performance across diverse environments. Future workshouldincludeextensivecross-datasetvalidation to ensure that the model generalizes well beyond a single dataset. Additionally, domain adaptation tech- niques can be employed to fine-tune the model for specificapplicationsorenvironmentalconditions.This could involve training on multi-modal datasets or leveragingunsupervisedomainadaptationmethodsto bridge the gap between training data and real-world conditions [13].

E. Multi-ModalIntegration

Anotherpromisingareaofresearchistheintegration of additional sensor data to improve detection per- formance. For example, combining visual data with thermal imaging or LiDAR could help in scenarios wherevisualcuesarelimitedduetopoorlighting or heavy occlusion. Multi-modal fusion could provide complementary information, thereby improving the robustness of object detection in challenging condi- tions.Earlystudieshaveshownthatsensorfusion can be highly effective in autonomous driving and surveillance [14].

F. User-CentricandReal-WorldEvaluations

While our quantitative results and controlled ex- periments are encouraging, it is essential to evaluate ourmodelinreal-worldscenarioswithendusers.

Future research should involve pilot deployments in real-world UAV systems, followed by user studiesand feedback collection. Such evaluations can provide valuable insights into the practical challenges of de- ployment,includinglatency,environmentalvariability, 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. EthicalConsiderationsandResponsibleDeploy- ment

AsadvancedUAVsurveillancetechnologiesbecome 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 devel- oping frameworks for responsible data management and transparent usage policies. Collaborations with ethicists, legal experts, and policy makers will be essentialtoensurethattechnologicaladvancementsare 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 couldfocusonextendedfieldtrialsinvariousenviron- ments—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 main- tenance, scalability, and environmental robustness.

I. IntegratingFeedbackLoopsforContinuousIm- provement

A final area for future exploration is the integra-tion 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. Suchanadaptivesystemwouldbeabletorespond to changes in the environment or object characteristics overtime,ensuringthatdetectionperformanceremains optimal as conditions evolve.

In summary, while ViT-YOLOv8 marks a signif- icant advancement in UAV-based small object de- tection, 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 BROADERIMPACT

The development of ViT-YOLOv8 represents a sig- nificant advancement in UAV-based small object de- tection. Our hybrid approach achieves a balance be- tween the local feature extraction of CNNs and the global context modeling of Transformers, resulting in improveddetectionaccuracywithoutsacrificingspeed.

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 environ- ments.
- 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 catas- trophic failures.
-

B. EthicalandPrivacyConsiderations

While our technological advancements offer signif- icant benefits, they also raise important ethical and privacyconcerns.EnhancedUAVsurveillancesystems must be deployed with strict data-handling protocols andtransparencytoprotectindividualprivacyandcivil liberties. It is imperative that these systems adhere to legal and ethical standards to ensure responsible use.

C. FutureResearchDirections

Lookingforward,thereareseveralexcitingavenues for future research:

- ReducingComputationalDemands:Investigat- ing model pruning, quantization, and more effi- cient transformer architectures could further re- duce computational costs.
- Cross-DatasetValidation:TestingViT-YOLOv8 onadditionaldatasetswillhelpconfirmitsrobust- ness and generalizability in varied environments.
- Multi-Modal Integration: Incorporating addi- tional sensor data, such as thermal imaging or LiDAR, may enhance detection performance un- der challenging conditions.
- Domain Adaptation: Developing techniques for domain adaptation will be critical to ensure that the model performs consistently across different scenarios.

Thesefuturedirectionswillfurtherrefineourmodel 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.

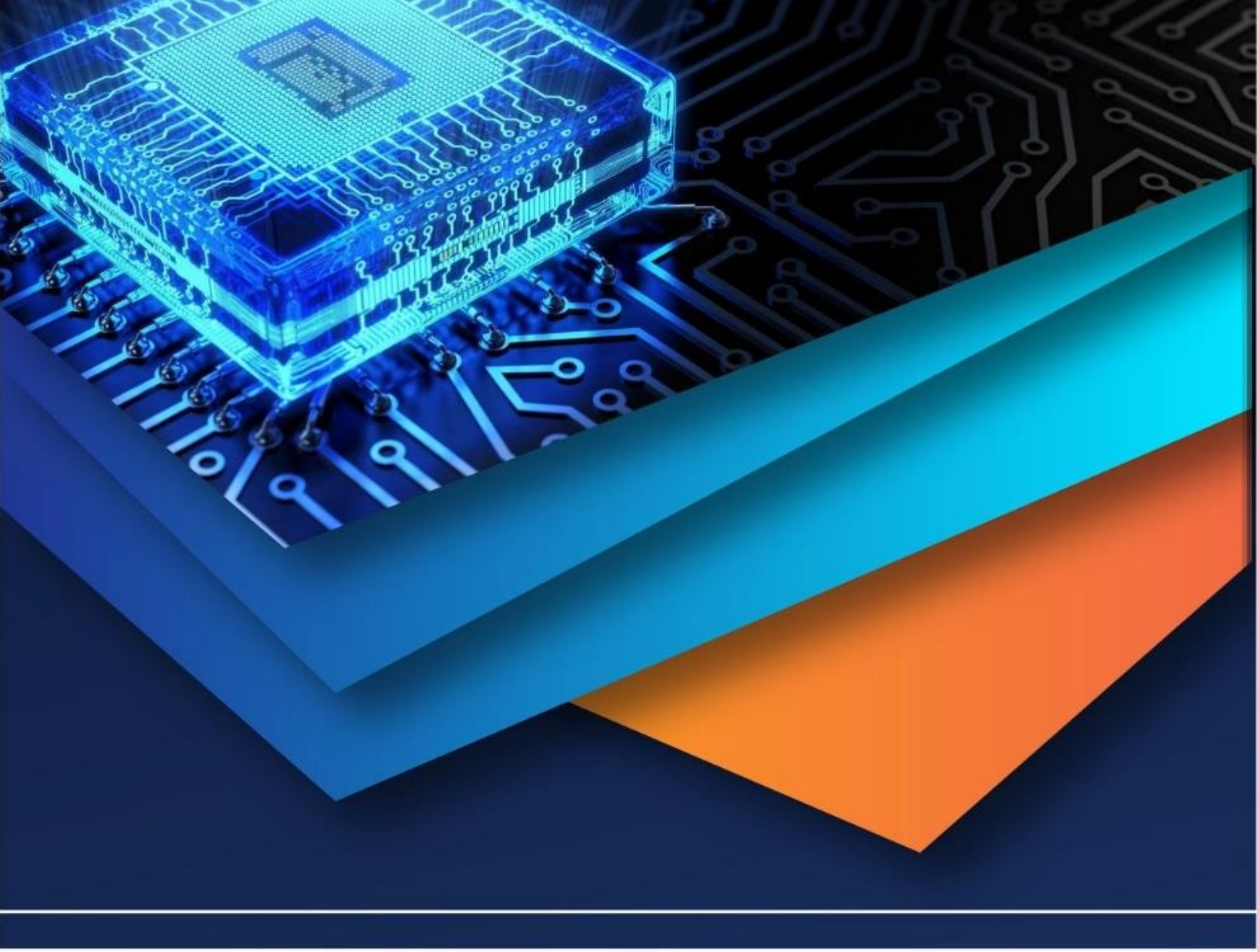
OurextensiveexperimentsontheVisDrone- DET2019 dataset demonstrate that ViT-YOLOv8 out- performsstate-of-the-artmethods,increasingthemean averageprecisionbyapproximately3.5pointsand achievinganinferencespeedof109FPS.Detailed ablationstudiesandadditionalreal-worldobservations confirm the critical contributions of each component. Althoughchallengesremain—particularlyinfur- therreducingcomputationaloverheadandensuring broad generalizability—ViT-YOLOv8 represents a sig- nificantstepforwardinUAV-basedobjectdetection. Webelievethatourworksetsanewbenchmarkfor smallobjectdetectionandprovidesarobustfoundation forthenextgenerationofreal-timeUAVsurveillance systems.

XI. ACKNOWLEDGMENTS

We gratefully acknowledge the support of the De- partment of Data Science at Christ (Deemed to be) University, Pune. We also thank the VisDrone com- munity 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,andZ.Zhang,"YOLO-ViT-Based Method for Unmanned Aerial Vehicle InfraredVehicle 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 DroneObject Detection in Image Challenge Results," in Proc. ICCVWorkshops, 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 ICCVWorkshops, 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:InvertedResidualsandLinearBottlenecks,"inProc. IEEE/CVF CVPR, 2018, pp. 4510–4520.
[6] J. Wang, K. Chen, R. Xu, Z. Liu, C. Loy, and D. Lin,"CARAFE:Content-AwareReassemblyofFEatures,"inProc.IEEE/CVF ICCV, 2019. Available: https://doi.org/10.1109/ICCV.2019.00310.
[7] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantagesof Careful Seeding," in Proc. SODA, New Orleans, LA, USA,2007.
[8] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "Cornernet-lite: Efficient keypoint-based object detection," arXiv, 2019.
[9] T.Lin,P.Dolla´r,R.B.Girshick,K.He,B.Hariharan,andS.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 Im-provement," arXiv, 2018.
[11] S.Ren,K.He,R.B.Girshick,andJ.Sun,"FasterR-CNN:Towardsreal-timeobjectdetectionwithregionproposalnetworks," in Proc. NeurIPS, 2015, pp. 91–99.
[12] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep Learningfor Unmanned Aerial Vehicle-Based Object Detection andTracking: A Survey," Geosci. Remote Sens., vol. 10, pp. 91–124, 2022.
[13] Z. G. Darehnaei, M. Shokouhifar, and H. Yazdanjouei, "SI-EDTL: Swarm intelligence ensemble deep transfer learningfor multiple vehicle detection in UAV images," ConcurrencyComputation, vol. 34, 2021.
[14] S. Cao, J. Deng, J. Luo, Z. Li, J. Hu, and Z. Peng, "LocalConvergence Index-Based Infrared Small Target DetectionAgainst Complex Scenes," Remote Sens., vol. 15, 2023.
[15] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convo-lutional One-Stage Object Detection," arXiv, 2019.
[16] J.Pang,K.Chen,J.Shi,H.Feng,W.Ouyang,andD.Lin, "Libra R-CNN: Towards Balanced Learning for ObjectDetection," 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