International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research

Unified Object Detection, Segmentation, and Tracking Using Mask R-CNN on COCO Subset

[1]N. ANJI REDDY,
[1,] Department of Computer Science and Engineering,
GODAVARI GLOBAL UNIVERSITY,
RAJAMAHENDRAVARAM,
ANDHRA PRADESH.
[1]anjireddy2025@ggu.edu.in

[2]P. NAGAMANI
[2]Department of Computer Science and Engineering,
GODAVARI GLOBAL UNIVERSITY,
RAJAMAHENDRAVARAM,
ANDHRA PRADESH.
[2]nagamanipedapati@gmail.com

***Abstract*— Object detection, instance segmentation, and multi-object tracking are important in modern intelligent vision systems applied in autonomous tasks, transportation, as well as surveillance. Nevertheless, when working in dynamic and complex conditions, classical approaches often cannot reach a compromise between accuracy of the detection, speed of the segmentation, and the real-time. Tests were done using the COCO 2017 val2017 subset that contains a number of diverse object categories with bounding box and segmentation annotations to overcome these concerns. In order to generate a specific assessment data, certain classifications such as human, automobile, bicycle, dog and truck were filtered. Scaling, normalization and data augmentation are some of the popular preprocessing techniques employed in order to enhance model resilience. The models implemented are also the Mask R-CNN which is used in detection and segmentation alongside YOLOv5, YOLOv8, and YOLOv26 to detect and track identities in real- time. Performance is measured using metrics such as mAP, iou, Precision, Recall, MOTA, MOTP and FPS. The object tracking reliability and the better accuracy-speed trade-offs in real-world conditions are the objectives of the unified framework.**

"Keywords— Object Detection, Instance Segmentation, Multi- Object Tracking, Mask R-CNN, YOLOv8, Deep SORT".

I. INTRODUCTION

Intelligent vision systems are now critical in the modern applications such as robots, autonomous navigation, intelligent transportation and surveillance. These systems are greatly dependent on object detection, instance segmentation, and multi-object tracking, which enables machines to discern, identify and comprehend dynamic environments accurately. The introduction of the idea of region-based convolutional architecture resulted in significant improvement of pixel-level understanding, where multiple processes of segmentation and localization could be performed in the same framework [1]. Subsequent developments and an in-depth analysis of the instance segmentation methods have increased its practical applicability in the real world as well as improved its performance standards [2]. These developments have made it possible to come up with integrated perception systems capable of studying scenes in detail in diverse operating environments.

These issues are many despite the great improvements. Many of the present systems, at the cost of computational efficiency and real-time use, focus on either the accuracy of detection or the precision of derivation. The problem of adaptation of the models to varying item scales, densities, and handling the case with scanty data also remains a challenge [3]. Architectural developments in the recent years have attempted to make detection pipelines leaner in an attempt to enhance robustness and converged stability [4]. However, the question remains of how detection, segmentation and tracking can be incorporated in such a manner that it becomes an integrated part of a performance-focused and consistent architecture. In the real world, the operational reliability is restricted by inconsistent identity preservation and reduced inference rates in complex scene dynamics.

The principal objective of this project is to come up with a coherent perceptual framework that incorporates the three elements of consistency in temporal identity, pixel-wise delineation, and spatial localization into a single deployable pipeline, so as to eliminate these limitations. Architectural originality is not given high priority but systematic integration, controlled evaluation, and comparative performance measurement among a few and free parts. The objectives of the system are stable tracking of identity across frames, realistic object boundaries and classification. The aim of the system is to enhance consistency in dynamic multi-object environments according to the matching between the detection and segmentation results and the methodologies of temporal association. The need to have end-to-end solutions with the ability to have a consistent interpretation of scenes in real-time scenarios is fulfilled by the integration.

The capacity to bridge the difference between scalable, application oriented deployment and high accuracy vision algorithms is what makes this work important. Although surveys on segmentation and tracking have emphasized the importance of a thorough assessment of tasks [6], previous studies on combined tracking and segmentation have demonstrated the feasibility of integrating spatial and temporal reasoning [5]. The developed architecture is based on these concepts and promotes reliability, computation balance, and flexibility in a variety of operational cases. The technology enhances automation functionalities and makes the smart decision in surveillance, mobility analytics, and autonomous systems by strengthening the alignment of the detection, segmentation, and tracking dynamics. This promotes viable visual intelligence solutions.

II. RELATED WORK

Recent advances in instance segmentation have demonstrated the versatility of region based DL frameworks in a number of application areas. Su et al. [7] considered the improved precision of object boundaries in high-resolution aerial sceneries through an advanced Mask R-CNN setup to remote sensing data. Their results indicated the ability of region-based segmentation to solve complex patterns in space, whereas they were sensitive to scale variations and dense item grouping. Similarly, Manikanta et al. [8] also explored the use of the Mask R-CNN in object detection with DL, indicating that the method outperforms the old school of thought. Their

primary focus was on the accuracy of detection without focusing on the robustness of segmentation or temporal consistency because even when their findings supported the effectiveness of region-based methods.

Zhang et al. [9] introduced a mask-refined R-CNN architecture, designed to enhance the object boundary delineation as a way of enhancing the details of the segmentation and feature refining. Despite the fact that more architectural complexity and processing overhead was observed to be a potential negative factor, the refinement method assisted in creating more accurate masks. Restrepo-Arias et al. [10] showed that instance segmentation models could be applied to domain-specific problems because they adopted Faster R-CNN and Mask R-CNN in an agricultural surveillance system to detect cocoa pods at various levels of ripeness. Although their approach had managed to detect their approach successfully in controlled conditions, it was not extensively applied to larger environmental variables.

Carvalho et al. [11] also examined the appropriateness of the Mask R-CNN to large and multi-channel images and suggested a method of mosaicking remote sensing data. Their solution did not fail to resolve the issue of processing large images but this method was computationally expensive and therefore could not be deployed in real-time. Jaikumar et al. [12] in their study of transfer learning in waste bottle segmentation highlighted the importance of the pretrained models as they reduce the training duration and improve convergence. Although these were the advantages, it had its limitations since the transfer was successful only when there was a similarity of domains. Though the scalability and cross-domain resilience was not explicitly studied, Chang [13] scaled Mask R-CNN to waste-detecting tasks and provided promising results in segmentation in the environmental monitoring.

Mahajan et al. [14] have studied the behavior of Mask R-CNN in case of visual distortions to evaluate the stability of its performance in a demanding environment. Their findings emphasized the vulnerability to instance segmentation frameworks to real-life disturbances as they demonstrated that there was an apparent reduction in segmentation truth upon being subjected to noise, blur, and compression error. All these studies are evidence of the strong representational capacity of architectures founded on Mask R-CNNs in industrial, agricultural, environmental monitoring, as well as remote sensing. Nonetheless, several challenges are still present, including limiting evaluation of the integration of the multi-object tracking, constraints on computational efficiency, sensitivity to distortion and insufficient benchmarking in diverse situations.

These limitations underscore the importance of the comprehensive approach where the preservation of the temporal identities and comparative performance assessment under specified conditions is considered alongside with the evaluation of the detection and segmentation accuracy of the results. The limitation challenges of robustness, scalability, and real-time applicability are expected to be addressed in the current study through the integration of numerous detection paradigms with segmentation and tracking within one evaluation framework. This integration is aimed at ensuring dynamic and complex visual situations are constantly performed and increasing the practical deployment capacity.

III. Materials And Methods

Focusing on particular classes, such as human, car, bicycle, dog, and truck, the proposed solution presents one framework of object recognition, instance detection, and multi-object tracking based on the COCO 2017 val2017 subset. All steps of the structured pipeline process are data intake, filtering of annotation, preprocessing, partitioning of the dataset, training the model, evaluation, and deployment. Whereas annotations are stored in COCO format when using Mask R-CNN [15] as well as the transformation to YOLO format when using YOLOv5 [16], YOLOv8, and YOLOv26 to enable parallel testing, the common preprocessing operations are employed to enhance the robustness. The trade-offs between models in accuracy and speed are studied with the help of a comparative evaluation method. Deep SORT uses detection outputs in order to retain object identities across frames to achieve temporal consistency. The framework is deployed through Flask based interface which enables real-time inferencing of images and videos. Besides enabling scalable and useful deployment, this integrated architecture enhances the stability of tracking, quality of segmentation, and detection accuracy [18].

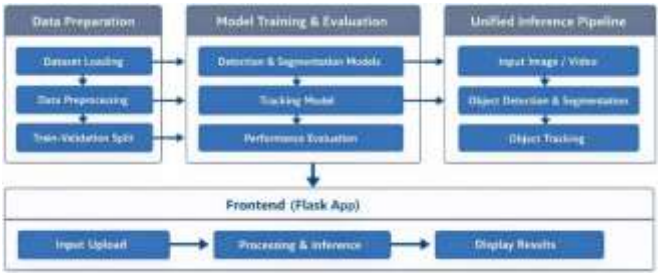


Fig.1 System Architecture

The complete system architecture such as Data Preparation, Model Training and Evaluation, Unified Inference Pipeline and Frontend are illustrated in Fig. 1. On the one hand, preprocessing, train-validation splitting, and loading data are implemented. Object identity management acquisition tracking module is trained together with detection and segmentation models (Mask R-CNN version and YOLO version) and the performances are later evaluated. Photos or videos on input are captured, divided and monitored in deployment. The Flask-based interface makes it possible to user upload, automatic inference, visualize bounding boxes, masks, item IDs and confidence rates.

A) Dataset Collection

The experiments utilize the val2017 subset of COCO 2017 dataset, which was obtained via the official repository of COCO. The dataset will contain images of natural scenes along with annotation JSON files which provide object-level information, such as class labels, segmentation masks and bounding boxes. In order to offer a specific set of assessment to give to this implementation, certain filters, such as human, automobile, bicycle, dog, and truck, are filtered. It suits tracking, segmentation, and detection work as it includes many annotation structures and different instances of objects.

B) Data Visualization

An exploratory data analysis step was conducted to analyze the balance of datasets and examine the

representation of specific classes of objects. In order to check the integrity of annotation and spatial consistency, representative images were shown with the corresponding bounding box and segmentation masks. This stage enabled identification of the interfering dominance of classes, the existence of imbalance, and discrepancies in the annotation. To facilitate reliable training and minimize the bias in the performance of detection and segmentation, visual inspection ensured that filtered categories such as human, car, bicycle, dog, and truck were accurately represented.

C) Data Preprocessing

Image preprocessing was carried out to make models more resilient and to standardize the input dimension. The value of pixel were scaled down to ensure that convergence was constant throughout training and all images were brought to a constant size that could be used in detection structures. In order to make the data more varied and reduce overfitting, data augmentation techniques like color jittering, rotation, scaling, and horizontal flipping were applied. In order to make cross-model predictions, annotation processing was used to translate labels to YOLO format to be used in YOLOv5, YOLOv8, and YOLOv26 and leave the original CoCO 2017 JSON format for Mask R-CNN.

D) Train–Validation Split

The filtered data was used to form a training and validation set with a ratio of 70:30 to enable objective evaluation of the performance of the selected approach. The YOLO-based branches of detection and the Mask R-CNN were both supported without data leakage through having a parallel directory structure. This is done to ensure generalization and prevent overfitting by ensuring evaluation of the model is conducted on unseen samples. The structural split also allows a consistent benchmarking among the detection, segmentation, and tracking components made within the same framework.

E) Algorithms

Mask R-CNN: Mask R-CNN can be used to perform object detection and instance segmentation using a parallel mask prediction branch to region-based convolutional networks. It improves the power of understanding the scene by extracting multi-scale features to enhance the localization accuracy, classification reliability and precise pixel-level mask generation on a ResNet50-FPN backbone.

YOLOv5: YOLOv5 is a single-stage object detector to be used to properly localize and classify objects. It fits real-time applications that have a limited processing capabilities as it handles entire images in a single architecture which makes it quick to infer and has competitive detection capabilities [19].

$$x = \sigma(t_x) + C_x, \qquad y = \sigma(t_y) + C_y \qquad (1)$$

$$w = p_w . e^{tw}, \qquad h = p_h . e^{th} \qquad (2)$$

YOLOv8: YOLOv8 has better feature extraction and architecture to enhance detection stability and precision. It has a simplified architecture, which facilitates generalization to a variety of object sizes and adverse visual tasks due to achieving enhanced convergence and improved accuracy-speed trade-offs.

$$\mathcal{L} = \lambda_{box}\mathcal{L}_{box} + \lambda_{obj}\mathcal{L}_{obj} + \lambda_{cls}\mathcal{L}_{cls} \qquad (3)$$

YOLOv26: YOLOv26 applies advanced optimization methods and better detection heads in order to raise the localization performance and the classification confidence. This design is scalable to real-time vision-based monitoring implementation since it focuses on computing efficiency and yet retains detection [20] reliability.

Deep SORT: Deep SORT enables tracking multiple objects by predicting motion and matching features based on their appearance across consecutive frames to identify the detected object across sequentially different frames. Where dynamic and congested contexts are concerned, it reduces the identity shifts, assigns things enduring identities, and maintains temporal consistency.

IV. Experimental Results

Precision: Precision ensures that the percentage of cases or samples correctly classified is accurate among the cases or samples that have been termed as positive. Thus, the precision may be calculated with the help of the following formula:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \qquad (4)$$

Recall: Recall is an indicator used in ML to determine how a model is able to find all instances of a given class that are relevant. It provides details on the extent to which a model predicts instances of a certain class by the total count of correctly classified positive instances divided by the total cases of actual positives.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (5)$$

mAP: MAP is one of the ranking quality statistics. It also considers the number of relevant recommendations and their position in the list. MAP at K is computed by taking the arithmetic mean of the AP of each of the users or queries at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \qquad (6)$$

mAP50: The AP was averaged with a 0.50 IoU level. It is a measure of the accuracy of the model in case the easy detections are considered only. mAP50-95: The AP over a range of 0.50 to 0.95 IoU thresholds.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \qquad (7)$$

Table 1. Detection Performance Comparison

Model	mAP@ 0.5	mAP@ 0.5:0.95	Precision	Recall
YOLOv5	0.88	0.64	0.86	0.83
YOLOv8	0.91	0.69	0.89	0.87
YOLOv26	0.89	0.67	0.87	0.85

Mask R-CNN	0.87	0.66	0.85	0.82

In comparison to other architectures that have been tested, YOLOv8 has the best detection performance as it shows a superior balance between localization accuracy, classification reliability, and real-time inference performance (Table 1).

Table 2. Segmentation Performance (Mask R-CNN)

Model	Mask mAP	IoU Score	Precision	Recall
Mask R-CNN	0.72	0.78	0.84	0.81

Based on Table 2, as it is stated, the Mask R-CNN improves fine-grained object boundary delineation through providing a high instance segmentation accuracy, accurate map prediction, and without variation in spatial overlap.

Table 3. Multi-Object Tracking Performance (Detection + Deep SORT)

Detection Model	MOTA	MOTP	ID Switches	FPS
YOLOv5 + Deep SORT	0.76	0.81	24	46
YOLOv8 + Deep SORT	0.82	0.84	17	42
YOLOv26 + Deep SORT	0.79	0.83	20	40
Mask R-CNN + Deep SORT	0.74	0.80	29	16

As demonstrated in Table 3, YOLOv8 and Deep SORT collaborate to achieve the most reliable tracking stability, which ensures stable identity maintenance and improved multi-object recognition in two consecutive frames.

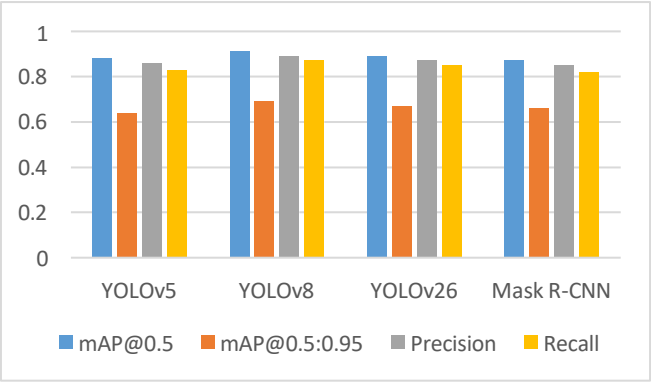


Fig.2 Comparison Graph - Detection

Figure 2 indicates the relative detection performance of the models in a visual format, with the best-balanced accuracy speed trade-off being shown by YOLOv8 which still has higher localization consistency and operating capabilities in real-time.

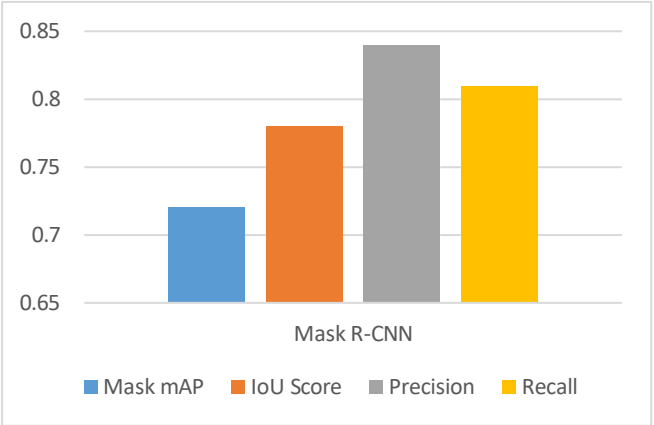


Fig.3 Comparison Graph - Segmentation

By using the mask prediction quality, mask R-CNN is able to offer a maintained mask prediction and reliable spatial overlap performance in the instance-level object delineation as indicated by the segmentation assessment measures presented in Figure 3.

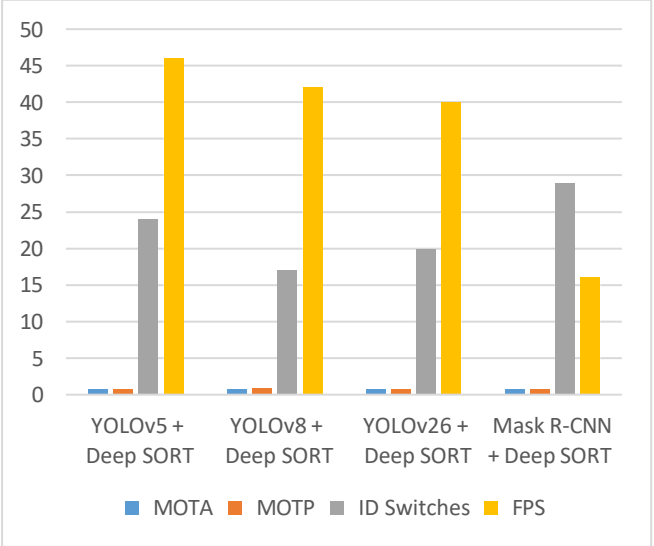


Fig.4 Comparison Graph - Multi-Object Tracking

Figure 4 demonstrates the efficiency of tracking, using which the combination of YOLOv8 and Deep SORT offers consistency in association of multi-object across frames and improved identity continuity.

V. Conclusion

Accurate localization in space and conservation of time identity in the complex scenarios are simultaneously illustrated by the combination of detection, segmentation and tracking within a single vision pipeline. This system uses Mask R-CNN to instantiate instance segmentation with real-time detectors such as YOLOv5, YOLOv8, and YOLOv26 by using the COCO 2017 val2017 subset with filtered classes such as human, car, bicycle, dog, and truck. Regular training and evaluation are assisted with the help of systematic dataset division and normalization of preprocessing. The fact that coupling the outputs of the detectors to Deep SORT also allows reliably attaching identities between consecutive

frames. Stability, segmentation quality and spatial accuracy have been well validated with the help of assessment using mAP, IoU, Precision, Recall, MOTA, MOTP, and FPS. Relative integration of multiple detection structures into one inference and deployment system enhances flexibility, computational trade-off and resilience. Altogether, the invented system provides a scalable and practically implementable solution of smart visual analytics and automated observation in dynamic real-life conditions.

To achieve higher levels of generalization in different situations, future work could focus on generalizing the framework to bigger datasets other than COCO 2017. By applying model compression methods and design lightness, it is possible to achieve real-time performance on the edge devices. Transformer-based detectors can also be used to supplement Mask R-CNN and YOLO versions and increase the understanding of the context. Moreover, the retention of identity in dense occlusion and closely crowded scenes might be enhanced through the adoption of other tracking algorithms other than Deep SORT.

REFERENCES

[1] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2961–2969).

[2] Hafiz, A. M., & Bhat, G. M. (2020). A survey on instance segmentation: state of the art. *International Journal of Multimedia Information Retrieval*, 9(3), 171–189.

[3] Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., & Lin, L. (2019). Meta R-CNN: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9577–9586).

[4] Mahesh Ganji. (2025). Enhancing Oracle Cloud HR Reporting Through AI-Driven Automation. Journal of Science & Technology, 10(6), 28–36. https://doi.org/10.46243/jst.2025.v10.i06.pp28-36.

[5] Noor, S., Waqas, M., Saleem, M. I., & Minhas, H. N. (2021). Automatic object tracking and segmentation using unsupervised SiamMask. *IEEE Access*, 9, 106550–106559.

[6] Gaddam, S. (2024). Integrating machine learning models with continuous integration and continuous delivery (CI/CD) pipelines for a learning-driven approach to software engineering.

[7] Su, H., Wei, S., Yan, M., Wang, C., Shi, J., & Zhang, X. (2019). Object detection and instance segmentation in remote sensing imagery based on precise Mask R-CNN. In IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 1454–1457). IEEE.

[8] Manikanta, M. S., Siva, K., Krishna, M. S., David, N., & Bhupal, M. R. (2025). Deep learning based object detection using Mask RCNN. IJSAT-International Journal on Science and Technology, 16(1).

[9] Zhang, Y., Chu, J., Leng, L., & Miao, J. (2020). Mask-refined R-CNN: A network for refining object details in instance segmentation. Sensors, 20(4), 1010.

[10] Restrepo-Arias, J. F., Montoya-Castaño, M. J., Moreno-De La Espriella, M. F., & Branch-Bedoya, J. W. (2025). Detection of cocoa pods at different ripening stages using Faster R-CNN and Mask R-CNN. Computation, 13(7), 159.

[11] Carvalho, O. L. F. D., et al. (2020). Instance segmentation for large, multi-channel remote sensing imagery using Mask R-CNN. Remote Sensing, 13(1), 39.

[12] Jaikumar, P., Vandaele, R., & Ojha, V. (2020). Transfer learning for instance segmentation of waste bottles using Mask R-CNN. In International Conference on Intelligent Systems Design and Applications (pp. 140–149). Springer.

[13] Chang, Y. (2022). Waste detection based on Mask R-CNN. Doctoral dissertation, Politecnico di Torino.

[14] Mahajan, P., Karlupia, N., Abrol, P., & Lehana, P. K. (2022). Analyzing performance of Mask R-CNN under distortions. In International Conference on Innovations in Computational Intelligence and Computer Vision (pp. 339–351).

[15] Rajarajeswari, R., & Sankaradass, V. (2023). Multi-object recognition and segmentation using enhanced Mask R-CNN. In ICDSAAI 2023 (pp. 1–6). IEEE.

[16] Explainable AI Framework for Policy-Compliant Anomaly Detection in Data Pipelines. (2025). Inaternational Journal of Communication Networks and Information Security, 16(4). https://doi.org/10.48047/ijcnis.16.4.2111

[17] Mallick, P. (2020). Offline-First Mobile Applications With Route Optimization Algorithms For Enhancing Last-Mile Delivery Operations. International Journal of Engineering Science and Advanced Technology, 20(4), 12–19. https://doi.org/10.64771/ijesat.2020.v20.i04.pp12-19.

[18] Yuan, B., et al. (2019). MU R-CNN: Two-dimensional code instance segmentation network. Future Internet, 11(9), 197.

[19] Machefer, M., et al. (2020). Mask R-CNN refitting strategy for plant counting in UAV imagery. Remote Sensing, 12(18), 3015.

[20] Taher, S., Dang, J., Fujishima, T., & Liu, J. (2021). Pixel-level multiple class damage detection using Mask R-CNN. In ICOSSAR 2021 (pp. 1–10).

[21] Saadiyean, Q., Samprithi, S. P., & Sundaram, S. (2024). Multi-scale context Mask R-CNN for aerial imagery. In IEEE ICRA 2024 (pp. 13573–13580). IEEE.