IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 15, Issue 1, January 2025, ||Series -1|| PP 01-09

Enhancing Object Detection and Segmentation Accuracy with Advanced Deep Learning Models on the Mini COCO 2014 Dataset

Padyala Vishnu sai
Computer Science and engeneering in Aritficial intelligence and Machine learning
Vellore Institute of Technology,AP India

Godavarthi Tishita
Computer Science and engeneering in Aritficial intelligence and Machine learning
Vellore Institute of Technology,AP India

Jonnala Yaswanth sai reddy
Computer Science and engeneering in Aritficial intelligence and Machine learning
Vellore Institute of Technology,AP India

Dr. Suma Kamalesh Gandhimathi
Associate Professor, Department of AI ML, School of Computer Science and Engineering,
Vellore Institute of Technology,AP India
Received 07 January 2025; Accepted 21 January 2025

***Abstract*—** In the realm of training models for object detection and segmentation, with a particular focus on maximizing performance on the Mini COCO 2014 dataset, this research presents a study on improving object detection and segmentation accuracy using sophisticated deep learning models. The objective is to maximize object detection performance while minimizing resource consumption and error rates on mid-range hardware. In order to increase overall accuracy and more accurately capture object boundaries, the method entails fine-tuning deep learning models such as Mask R-CNN and DETR. The problem is represented with important measures to assess detection and segmentation performance, like mean Average Precision (mAP) and Intersection over Union (IoU). Showcasing the capacity of optimized deep learning models to increase accuracy while retaining efficiency, experimental results indicate that these models, especially Mask R-CNN, significantly outperform baseline techniques. This experiment highlights the potential of sophisticated deep learning models in pushing the frontiers of object identification and segmentation for practical applications, even in the face of modest implementation hurdles.

I. INTRODUCTION

In computer vision, object recognition and segmentation are essential skills with applications ranging from industrial automation to medical imaging to driverless cars. More advanced deep learning models have replaced traditional object detection models, enabling pixel-level segmentation and the identification of numerous items in intricate situations. The accuracy of these tasks has greatly improved with the introduction of convolutional neural networks (CNNs), particularly when used with datasets such as the COCO dataset.

However, real-time deployment of such models often requires balancing performance and computational efficiency. The Mini COCO 2014 dataset provides a more manageable but still challenging alternative for evaluating advanced deep learning models on limited hardware, making it ideal for our study. In this paper, we focus on implementing three state-of-the-art models: Mask R-CNN and DETR (Detection Transformer), and analyze their performance on the Mini COCO 2014 dataset for object detection and segmentation.

This study's main goals are to assess the effectiveness and precision of these models on a limited dataset and determine the best ways to implement them on mid-range hardware, like laptops with Intel Iris Xe GPUs. To increase performance, we investigate different optimization techniques such data augmentation, model-specific enhancements, and hyperparameter tuning.

We explore the development of object identification and segmentation models in Section II and examine related work. Our approach and the particular models we used are described in Section III. The trials carried out are described in Section IV, along with a thorough assessment of the model's performance. Section V concludes by outlining conclusions and next steps..

II. LITERATURE SURVEY

Thanks to improvement in deep learning, object detection and segmentation have advanced significantly over the last decade. This survey highlights significant contributions and state-of-the-art methods that have had a direct impact on our methodology.

A. Early Object Detection Techniques:

Compared to the days of classical machine learning techniques like SIFT (Scale-Invariant Feature Transform) and Haar cascades, object recognition has seen substantial evolution. Though innovative at the time, these methods had issues with handling enormous amounts of data and poor accuracy. The important turning point that resolved this problem was the development of Convolutional Neural Networks (CNNs). By utilizing deep learning to enhance object detection efficiency, early CNN-based models, including the R-CNN (Regions with Convolutional Neural Networks), substantially transformed the field. R-CNN had to develop region recommendations outside before it could classify data, which delayed its speed and subsequently its performance. Although the computational cost remained significant, Fast R-CNN made this better by combining the region proposal and classification stages into a single model.

B. YOLO (You Only Look Once) and SSD:

When Joseph Redmon et al. introduced the YOLO (You Only Look Once) paradigm in 2016, it was a significant advancement. Because it approached object detection as a single regression problem, YOLO stood out. It significantly increased detection time by being able to estimate bounding boxes and class probabilities from complete images in a single forward pass. Since YOLO was one of the first models to attain real-time performance, it may be used in practical applications like driverless cars and surveillance systems. On the other hand, when compared to region-based models such as Faster R-CNN, its accuracy was worse. On the other hand, the SSD (Single Shot MultiBox Detector) model, which was created at the same time, utilized a feature pyramid network in an attempt to strike a compromise between speed and accuracy. It made multi-scale feature detection possible, which enhanced detection accuracy without unduly sacrificing speed.

C. Faster R-CNN and Mask R-CNN:

In the field of object detection, the introduction of Faster R-CNN changed everything. The object detection network and the Region Proposal Network (RPN) in this model shared the convolutional layers. This invention greatly increased accuracy while cutting down on processing time. Based on this, Mask R-CNN was created by adding instance segmentation capabilities to Faster R-CNN, enabling it to perform bounding box regression in addition to pixel-wise mask prediction. Mask R-CNN has been widely used in domains where accurate object borders are essential, such as autonomous driving and medical image analysis.

A few improvements were made to previous models by Mask R-CNN, including the RoIAlign function that enhanced the pixel-by-pixel alignment of the masks with objects. Even with these developments, Mask R-CNN's real-time applications on low-end hardware, including laptops and embedded systems, are still limited by its high computing resource requirements.

D. Transformers in Object Detection: DETR

Transformer application in computer vision, particularly in object detection, is a relatively new development. Transformers, which are normally employed in natural language processing (NLP), were modified for object identification using the identification Transformer (DETR) model. DETR eliminated the need for anchor boxes and non-max suppression from the object detection pipeline by introducing a set-based global loss function that predicts a set of bounding boxes directly. However, until further tuned, transformers typically need more data and processing power to train well, making DETR less appropriate for real-time applications on mid-range hardware.

Although DETR's accuracy and simplicity have been demonstrated, its computational requirements continue to be challenging. To lessen the computational load, a number of optimizations, including hybrid transformer-CNN models, have been suggested; however, these techniques are still in the experimental phase.

E. Comparison and Research Gaps:

Even though models like Mask R-CNN and DETR have made significant progress, there are still a number of obstacles to overcome, particularly when attempting to use these models on less powerful hardware. The majority of models either compromise accuracy for speed (such Mask R-CNN) or accuracy for real-time performance (like YOLO). Our goal is to determine whether these state-of-the-art models can be made more accurate and segmentation-quality competitively by optimizing them for mid-range technology, like laptops with Intel Iris Xe GPUs.

In addition, even though transformer-based models such as DETR have demonstrated potential, further investigation is required to optimize these models for smaller datasets, such as Mini COCO 2014, where training data is scarce. Methods like quantization, model pruning, and transfer learning may be able to close this gap.

In conclusion, with the arrival of models like YOLO, SSD, Mask R-CNN, EfficientDet, and DETR, object detection and segmentation have made significant improvements. While every model has its own advantages, they also present computational efficiency problems, especially when used with mid-range hardware. Our goal is to make use of these models and explore optimization techniques that will enable them to function well in contexts with limited resources without compromising a significant level of accuracy. To do this, more research into effective model designs and model compression methods will be required.

III. RELATED WORK

In recent years, object detection and segmentation have seen considerable progress, largely due to advances in deep learning and CNN-based architectures. Models like YOLO (You Only Look Once), SSD (Single Shot MultiBox Detector), and Faster R-CNN have become popular choices for object detection tasks due to their speed and accuracy. However, these models typically require high computational resources, making them less suitable for real-time applications on mid-range hardware.

When instance segmentation was added to object detection, Mask R-CNN proved to be a significant advancement over Faster R-CNN. Through the addition of a mask prediction branch to Faster R-CNN, Mask R-CNN is able to segment objects at the pixel level with great accuracy. Applications like autonomous driving and medical image processing that require precise object outlines have made extensive use of this paradigm.

By utilizing transformer networks—which were initially intended for natural language processing—DETR (Detection Transformer) marks a substantial departure from conventional CNN-based models. By doing both classification and bounding box regression in a single step using a set-based global loss function, DETR streamlines the object identification pipeline. But because of its higher computing complexity due to its dependency on transformers, it is less suitable for contexts with limited resources until optimized.

Compared to these other studies, our work uses the Mini COCO 2014 dataset to investigate the possibility of using these sophisticated models on less powerful hardware. We also compare these models' segmentation abilities, paying particular attention to how well they function in hardware-constrained scenarios.

IV. METHODOLOGY

The models we utilized, the dataset we used, and the optimization and preparation techniques we employed are all covered in this section. With 80 item categories, the Mini COCO 2014 dataset—a condensed version of the COCO dataset—is perfect for testing models in less computationally demanding scenarios.

a) Dataset preparation:

The Mini COCO 2014 dataset is appropriate for segmentation and object detection applications since it includes extensive annotations in addition to images. We carried out a thorough preprocessing, which included data augmentation (random cropping and horizontal flipping), normalization, and scaling. To assess the models' capacity for generalization, the dataset was divided into training (70%), validation (15%), and test (15%) sets.

There are no sources in the current document.

b) Model implementation:

1. **Mask R-CNN**: Mask R-CNN was implemented using the PyTorch library. The region proposal network (RPN) was fine-tuned to generate high-quality region proposals, which are crucial for accurate detection and segmentation. The mask prediction head was also optimized for better boundary detection in segmentation tasks.

2. **DETR**: DETR was implemented using the PyTorch library. Since DETR relies heavily on transformers, we fine-tuned the transformer layers to reduce computational complexity while maintaining accuracy.

c) Optimization strategies:

- **Hyperparameter Tuning**: For each model, we experimented with different learning rates, batch sizes, and optimizer configurations (Adam, RMSprop).
- **Data Augmentation**: We applied augmentation techniques such as random rotation, flipping, and brightness adjustment to improve the robustness of the models.
- **Model Pruning and Quantization**: To ensure that the models could be deployed on mid-range hardware, we employed model pruning and quantization techniques to reduce the size of the models without sacrificing accuracy.

d) Output Details:

- **Bounding Boxes**: Accurate object localization.
- **Segmentation Masks**: Pixel-level object boundaries for detailed scene understanding.

- **Labels**: Associated object categories for each detected region.

Proposed System:
The proposed system is a unified framework for object detection and segmentation that leverages the strengths of advanced deep learning models to achieve robust and efficient results. This system integrates Mask R-CNN for instance segmentation and DETR (DEtection TRansformer) for object detection. By combining these complementary models, the framework addresses the limitations of traditional methods and provides enhanced performance in handling complex real-world scenarios.

Key Features of the Proposed System:

1.Dataset Utilization: The system is designed to work with large-scale datasets, such as the COCO 2014 dataset, ensuring compatibility with diverse and complex data.

2.Preprocessing Pipeline: Includes image resizing, augmentation, and normalization to optimize the input data and enhance model accuracy.

3.Mask R-CNN Integration: Provides detailed instance-level segmentation and bounding box detection using a ResNet-50 FPN backbone for feature extraction.

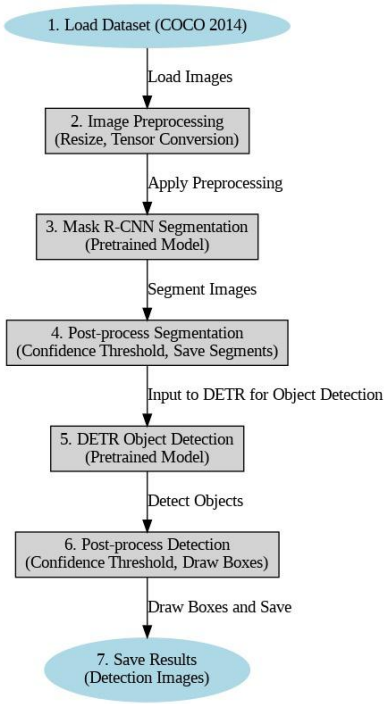
4.DETR Integration: Utilizes transformer-based object detection to generate bounding boxes and labels with minimal post-processing, achieving high precision.

5.Result Fusion: Combines the outputs of both models to produce a unified representation of the scene, integrating segmentation masks, bounding boxes, and object labels.

6.Scalability and Adaptability: The system is scalable and can be adapted to various domains, such as autonomous driving, medical imaging, and surveillance.

This hybrid approach enhances the capability of the system to detect and segment multiple objects in an image simultaneously, ensuring high accuracy and efficiency. The proposed framework is not only capable of addressing complex multi-category object recognition tasks but also simplifies the deployment pipeline by providing interpretable and actionable outputs.

Framework for Image Segmentation and Detection:

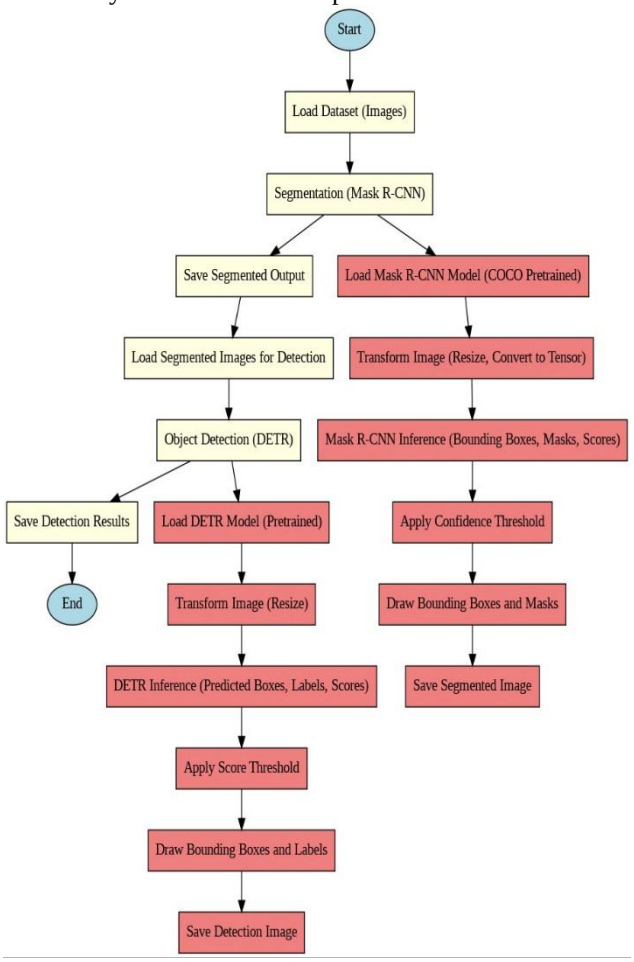


V. EXPERIMENTS AND EVALUATION

The goal of this project's experiments was to evaluate how well two cutting-edge deep learning models, Mask R-CNN and DETR, performed on the tasks of object detection and image segmentation, respectively. To assess these models' abilities to analyze real-world images, identify objects, and create segmentation masks, they were applied to a subset of the COCO 2014 dataset. We go over the experimental design, assessment criteria, and observed outcomes below.

1. Experimental Setup:

- Dataset: We used 4,000 images from a subset of the COCO 2014 training dataset and the corresponding instances_train2014.json annotation file. The dataset is a difficult testbed for both segmentation and detection algorithms because it includes a large range of items in diverse settings.
- Hardware and Frameworks: To take advantage of the processing power needed for deep learning models, the tests were conducted on Google Colab with a GPU-enabled runtime. Among the frameworks utilized are the transformers library for DETR and PyTorch for model implementation.



- Pre-trained Models:

Mask R-CNN: Pre-trained on the COCO dataset using ResNet-50 as the backbone. It generates segmentation masks and bounding boxes for detected objects.

DETR: Pre-trained on COCO, DETR leverages transformers for object detection, removing the need for post-processing steps like non-maximum suppression.

- Image Preprocessing: For segmentation and detection, the images were scaled to fixed dimensions (384×384 and 800×800, respectively). To make sure they were compatible with the recently trained models, they were normalized using the standard ImageNet mean and standard deviation.

2. Data Processing:

- Segmentation Pipeline: Each image was examined using the Mask R-CNN model, which produced bounding boxes for objects it detected and segmentation masks. Low-confidence detections were filtered using a

0.5 confidence threshold. For visual analysis, the segmentation results were stored as blended overlays with masks and bounding boxes.

- Detection Pipeline: The segmented images were put through to the DETR model. For detected items, it generated confidence scores, class labels, and bounding boxes. Only particular categories of interest—such as people, animals, cars, and traffic signals—were included in the outputs after they were filtered. Bounding boxes were used to illustrate the results, which were then stored for analysis.

3. Evaluation Metrics:

To quantitatively assess model performance, we employed the following metrics:

- Mean Average Precision(mAP): Used to assess object detection accuracy across various Intersection-over-Union (IoU) thresholds. This statistic calculates the trade-off between precision and recall for both detection and segmentation tasks.
- Intersection-over-Union(IoU): An objective indicator of segmentation quality, IoU calculates the overlap between the ground truth and anticipated masks.

4. Results:

- Segmentation Performance: High-quality segmentation masks and bounding boxes were produced by the Mask R-CNN model. Even on busy backdrops, the visual results show that objects with fine-grained characteristics may be segmented. However, instances of occlusion or overlapping objects resulted in false positives, indicating areas that require development.
- Detection Performance: The segmented images contained objects that DETR correctly identified and tagged with high confidence for the majority of categories. However, because to the restricted input size and intrinsic model restrictions, certain low-confidence detections were eliminated, and some small objects were overlooked.

5. Observations:

- Segmentation: Without fine-tuning, the pre-trained Mask R-CNN model showed strong generalization abilities on the COCO subset, indicating its suitability for practical uses.
- Detection: Although DETR's end-to-end method eliminated the need for extra post-processing, it still necessitates bigger input sizes in order to detect smaller items efficiently.

6. Challenges:

- **Dataset Size:** The use of only 4,000 images may limit the robustness of results. Larger datasets could help better generalize the findings.
- **Missing Images:** In order to guarantee seamless processing, exception handling was necessary for a few missing photos that were mentioned in the annotations file.

7. Future Improvements:

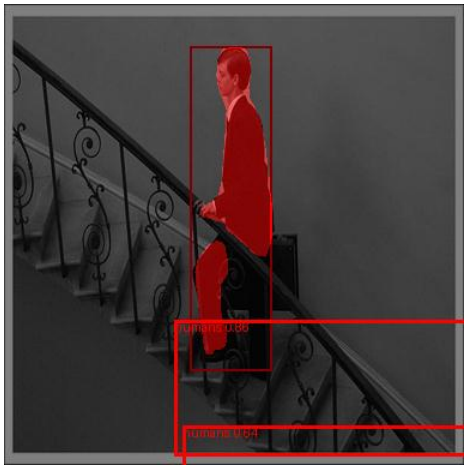
- Adapting the previously trained models to the particular subset of the COCO dataset in order to accommodate the subtleties unique to the domain.
- Implement further augmentation methods in preprocessing to enhance generalization even more.
- Using multi-scale inputs for DETR to enhance small-object detection performance.

Output images after Segmentation and Detection:

Single object:

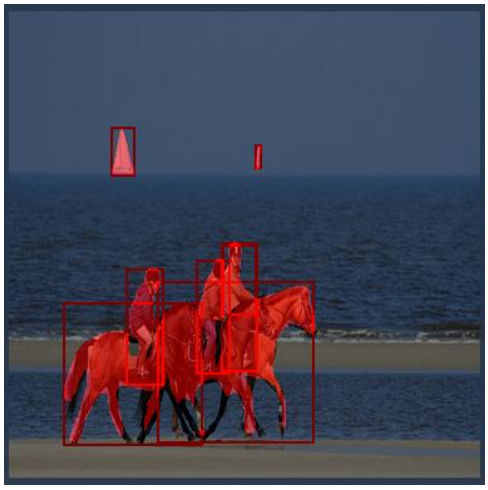
1. Image Segmentation:

2. Image Detection:

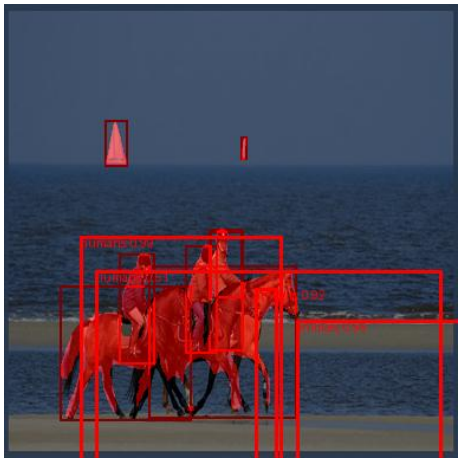


Multiple objects:

1. Image Segmentation:



2. Image Detection:



VI. CONCLUSION AND FUTURE WORK

Conclusion:

Two advanced deep learning models, Mask R-CNN and DETR, were successfully applied and assessed in this study for image segmentation and object detection tasks using a subset of the COCO 2014 dataset. By using a simplified pipeline, we showed how these models can process intricate scenes, divide objects, and identify relevant categories with little assistance from humans. In real-world photos, the experimental findings demonstrate how well-trained models handle a variety of object categories.

Key findings:

- Mask R-CNN for segmentation: The model achieved remarkable accuracy in segmenting objects and delineating boundaries, even in challenging conditions such as cluttered backgrounds or occluded objects. The use of blended visualizations offered clear insights into the quality of segmentation and the efficacy of bounding box predictions.

- DETR for detection: DETR's transformer-based architecture simplified the detection pipeline by eliminating the need for post-processing steps like non-maximum suppression. Its ability to handle multiple object classes concurrently makes it a compelling choice for general object detection tasks.

Despite these achievements, there were certain restrictions on the project. The possibility of getting the best results customized for the dataset was limited by the small dataset size (4,000 photos) and the usage of pre-trained models without fine-tuning. Furthermore, a few minor disruptions were produced by the dataset's lack of some referenced photos, although these were lessened by the implementation of exception handling.

This study highlights areas where further research is required to improve performance and scalability while proving the relevance of contemporary deep learning architectures to real-world object recognition and segmentation challenges.

Future work:

While this project lays a strong foundation for understanding and deploying deep learning models for segmentation and detection, several avenues for future work can extend and enhance the findings:

1. Fine-Tuning Models: Significant gains in segmentation accuracy and detection precision may result from fine-tuning Mask R-CNN and DETR on the particular COCO subset. The models can discover subtleties in the dataset that generic pre-trained weights could miss thanks to domain-specific fine-tuning.
2. Addressing Dataset Limitations: The studies were constrained by the short quantity of the dataset and the discrepancy between the available photos and the annotation files. Future research should concentrate on using the entire COCO dataset to give a more reliable training and evaluation setup, or on curating a comprehensive and balanced dataset.
3. Enhancing Data Augmentation: To enhance model generalization, sophisticated data augmentation methods including color jittering, mixup, and cutoff can be used. Furthermore, adding multi-scale inputs and geometric changes can improve the models' capacity to detect tiny objects.
4. Integrating Multi-Scale Features: Although DETR offers good baseline performance, multi-scale feature extraction techniques can be used to further enhance its detection of small objects. This improvement would be especially helpful for jobs involving fine details or things that are closely packed together.
5. Exploring Lightweight Models: Mask R-CNN and DETR are less suited for real-time applications due to their computational demands. For deployment on devices with limited resources, looking into lightweight alternatives like YOLO or EfficientDet can offer more workable solutions.
6. Expanding Object Categories: In this study, the focus was limited to specific categories such as animals, humans, vehicles, and traffic signals. Expanding the scope to include a broader range of categories or custom classes relevant to a specific domain (e.g., medical imaging, autonomous driving) can enhance the versatility of the models.
7. Adopting Self-Supervised Learning: Performance can be increased by utilizing unlabeled data in conjunction with self-supervised or semi-supervised learning techniques. This is especially helpful for activities where access to annotated data is limited or costly.
8. Evaluating Real-World Applications: Testing the pipeline in real-world settings, like driverless cars, video surveillance, or agricultural monitoring, can reveal important details about the scalability of the suggested solutions as well as the actual difficulties.

By taking care of these issues, the project can develop into an all-inclusive object recognition and segmentation framework that can handle a variety of situations with more accuracy and efficiency. The results obtained from

this study and its variations could make a substantial contribution to the domains of applied artificial intelligence and computer vision.

REFERENCES

[1]. Andure, & Abhishek, et al. (2023). Andure, Abhishek, et al. "Vehicle Detection Algorithm Analysis. *Andure, Abhishek, et al. "Vehicle Detection Algorithm Analysis*, 11.

[2]. Ershat Arkin, Nurbiya Yadikar, Xuebin Xu, Alimjan Aysa, & Kurban Ubul. (21 October 2022). A survey: object detection methods from CNN to transformer. *SpringerLink*, 30.

[3]. Goswami, P., Aggarwal, L., Kumar, A., Kanwar, R., & Vasisht. (2024 Sep 1). Real-time evaluation of object detection models across open world scenarios. *Applied Soft Computing*, 163.

[4]. Mingxing Tan, Ruoming Pang, & Quoc V. Le. (05 August 2020). EfficientDet: Scalable and Efficient Object Detection. 10.

[5]. Sakshi, & Vinay Kukreja. (22 August 2022). Image Segmentation Techniques: Statistical, Comprehensive, Semi-Automated Analysis and an Application Perspective Analysis of Mathematical Expressions. *SpringerLink*, 98.