Enhancing Small Object Detection in Aerial Imagery: A Comparative Study of YOLO and RT-DETR Models Using Slicing Aided Hyper Inference

MSc Research Project
MSc in Data Analytics

Ajaykkumar Rajendran
Student ID: 23197439

School of Computing
National College of Ireland

Supervisor: Prof. Jaswinder Singh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ajaykkumar Rajendran		
Student ID:	23197439		
Programme:	MSc in Data Analytics	**Year:**	2024-2025
Module:	MSc Research Project		
Supervisor:	Jaswinder Singh		
Submission Due Date:	12-12-2024		
Project Title:	Enhancing small object detection in aerial Imagery: A comparative study of YOLO and RT-DETR models using Slicing Aided Hyper Inference		

Word Count: 7003 words **Page Count:** 18 pages

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ajaykkumar Rajendran

Date: 12-12-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Small Object Detection in Aerial Imagery: A Comparative Study of YOLO and RT-DETR Models Using Slicing Aided Hyper Inference

Ajaykkumar Rajendran
Master's in Data Analytics
National College of Ireland, Ireland.
x23197439@student.ncirl.ie

Abstract. Detecting small objects in many aerial images caught by drones is now an important challenge that many experts face in the field of computer vision. Traditional object detection algorithms often struggle to identify small objects. Their low resolution and complicated aerial environments cause this issue. Traffic monitoring, urban planning, and disaster management will grow in relevance with this issue. The performance of deep learning models like YOLO and RT-DETR are investigated in this study. The Slicing Helped Hyper Inference technique is used to enhance these models for detection improvements. In this study, three models were trained on a VisDrone dataset for 200 epochs and their performance were compared with standard inference and SAHI techniques to determine their efficiency. The outcome of this study showed that with SAHI models were performing better with higher mAP metrics, especially in detecting small objects. By increasing accuracy and reliability, many practical applications can greatly benefit from improved object detection for drone surveillance systems, which are important in many fields such as urban planning and emergency responses. SAHI technique and pre-trained models are trained and combined in this study, drawing attention to opportunities to solve problems in computer vision.

Keywords - Deep Learning, Object Detection, Small Objects, YOLO (You Only Look Once), RT-DETR (Real-Time Detection Transformer), SAHI (Slicing Aided Hyper Inference

1. Introduction

Deep learning has fast improved, impacting computer vision and, in specific terms, pictures taken from the sky. The detection of small objects is one of the big challenging areas in this domain. This forms an integral part of modern society since drones are gradually becoming an integral part, used for city planning, monitoring of traffic, and rescue operations after disasters.

Traditional object detection algorithms have long met challenges in terms of small object identification with low resolution and complex backgrounds in aerial images. This paper is to test the comparative effectiveness of two state-of-the-art deep learning models, YOLO (You Only Look Once) and RT-DETR (Real-Time Detection Transformer) enhanced by a novel technique, Slicing Aided Hyper Inference (SAHI) on their effectiveness of small object detection from aerial drone-captured imagery

Before the advancement in data learning models, small objects detection was performed manually, which is a tedious process with full of errors. Traditional methods like Haar cascades and Edge detection algorithms, shown weaknesses in their ability to accurately detect small objects in complex aerial images. As put by Doloriel & Cajote (2023), the currently used aerial detectors often don't consider the challenges of detecting small objects, laying more focus on orientation modelling and doesn't take size limitations into account. This drawback encourages more promising approaches that are able to overcome these barriers.

Deep learning arose and revolutionized object detection through new architectures capable of learning hierarchical features and contextual relationships from large datasets. The R-CNN model was one of the major inventions by Girshick et al. (2014), in using a region proposal network to improve the understanding of object representations. Models like YOLO, SSD and RT-DETR have advanced the field with multiscale feature extraction which improves the robustness of detection of object with variable sizes. These models perform better by capturing context information across various images with different image sizes which is an important aspect of detecting small objects.

In recent years, data augmentation techniques had a significant growth in increasing the model performance. In this study, three models were trained on VisDrone dataset for a 200 epochs and they were compared based on the standard testing technique and SAHI testing technique to determine their effectiveness. Apart from that the transformer-based architectures proved to be a robust tool to aid the complex within the images.

Advancements in computational technology have increased GPU performance, resulting in deep learning frameworks contributing to the tremendous improvement of neural network functionality for the training of increasingly complex architectures with extraordinary precision for small object detection in a diversity of imaging contexts. In various applications such as urban planning, disaster management, deep learning have the potential to do image analysis in very high detail. Although there was advancements in computational technologies, detection of small objects still remain an issue for numerous reasons.

Traditional approaches usually cannot achieve satisfactory performance for objects with small pixel sizes, which might cause misinterpretations. Besides, traditional approaches with horizontal bounding boxes are also restricted to a certain extent in the precise localization of objects in drone-captured images. The aim of this research is to investigate advanced deep learning models namely, YOLO and RT-DETR improved by the SAHI technique to handle such issues. This study tests the performance of those models on different object categories using the VisDrone dataset, hence providing imperative insight into the development of object detection methods for difficult conditions. Improving object detection approach contain significant potential for practical uses in various fields. Finding solution for improving small objects detection, this research has significant impact for the wider technological advances in both computer vision and artificial intelligence.

1.1 Research Question

How do deep learning models, specifically YOLO and RT-DETR enhanced with the Slicing Aided Hyper Inference (SAHI) technique, perform in the detection of small objects in drone-captured aerial imagery, and how does their performance vary across different object classes within the VisDrone dataset?

In this work, deep learning models are introduced to investigate the performance in the domain of small object detection on aerial images captured by drones, enhanced by slicing aided hyper inference with YOLO and RT-DETR. It is intended to compare these models against the VisDrone dataset for the improvement of the detection accuracy of different classes of objects. The following section, Literature Review, situates our research within the extant work done about object detection in aerial images. This gives an idea of the state-of-the-art techniques and methodologies that guided our approach and leads up to the detailed discussion of findings and their implications for future research in this very important domain of computer vision.

2. Related Work

Object detection and image classification are the major field areas where deep learning models get used. In this part, the overview will consider the work of past researchers regarding techniques pursued to detect small objects. The main goal here will be finding the strong points and weaknesses of their approach to advance these detection methods in this very research project.

2.1 Review on Traditional methodologies

Object detection in aerial imagery has improved greatly with the development of new machine learning and computer vision approaches. Previous approaches have been at pains to tackle certain inherent challenges in aerial images, especially the detection of small objects that are hard to recognize against complex backgrounds and changes in scale. Yao & Zhang (2005) respond to this challenge by developing an Enhanced Semi-Supervised Learning (ESL) framework to increase the accuracy of detection through classifying two stages, candidate generation and classification. The study shows that the current semi-supervised learning algorithms have several limitations, for example, it is costly to set up the ground truth manually and the semi-supervised learning process needs to be carried out in iteration because the whole detection depends on implementing the supervised learning several times.

In the relevant studies, Ramli et al. (2017) construct an extensive obstacle detection system explicitly for small unmanned aerial vehicles. The authors present a hybrid approach that combines Lidar technology and a camera sensor to enhance the identification of obstacles. The Lidar sensor, in this setup, acts as the primary detector, with the camera providing pictures that the SURF algorithm analyses and estimates obstacle size based on feature point identification. Their findings confirm the applicability of such an approach in real time in indoor scenarios and show the ability of UAVs to spot obstacles and safely navigate the environment.

Narayanan et al. (2018) pointed to the use of real time perception for UAV applications with enhanced situational awareness for surveillance and reconnaissance applications. Authors propose a perception framework based on an embedded computer vision platform running on-board a hexacopter carrying a camera with its respective driver interacting with the deep neural network-based object detection module. Preliminary tests using the YOLO algorithm show great promise, with image sequences streamed for further analysis using the Multi-Expert Region-based CNN algorithm. The authors have also suggested exploiting a mix of aerial data and synthetic images using video gaming engines to help improve neural network training.

Thus, combination of various object detection methods becomes a very crucial area for research, especially in improvements over the Haar cascade classifier. As pointed out by North (2017), most detectors are oriented to specific features like a face or body parts. Therefore, running several detectors consecutively may yield a better result in target detection reliability for complex and deformable objects like horses. From this approach, the role of comprehensive datasets in the training of Haar cascade detectors is underlined.

Mentari et al. (2023) came up with a simulation tool that could help the visually impaired in the detection and counting of human objects using Haar cascades, realizing a 71 percent accuracy rate in the detection of human objects in video footage. This integration of hardware and software components shows the classifier to be effective despite its inherent limitations.

Further works by Ohn-Bar & Trivedi (2017) point out the limits of boosted decision trees in object detection architectures. The results showed that boosted decision trees were competitive with deep learning models and achieved a rate of 9.71 percent on the Caltech Pedestrian Detection benchmark, even though they still suffer from limited modeling power and required architectural improvements.

Xuet al.(2022) highlighted that modern object detection systems problem when applied to environments characterized by high density and low quality of proposal regions, which seriously degrades the accuracy of object detection. The experiment is evaluated using Cascade R-CNN reported a mean Average Precision (mAP) of 0.413 on the SKU-110K dataset. This shows that quality improvement of region proposals will result in improved detection performances in complex situations. Taken as a whole, these studies represent a heterogeneous approach toward object detection, fusing strengths from established methods with those of more recent developments. They emphasize the refinement of conventional algorithms while also encouraging exploration into new architectures able to handle the increased complexity brought about by detection problems in aerial imagery. These will be the main obstacles this research will have to handle, contributing to the development of methodologies aimed at improving small object detection in aerial imagery, which facilitates improvements in several applications, from urban planning to disaster management.

2.2 Review on Convolutional Neural Network Object detection methodologies

High-resolution object detection is nowadays becoming one of the most necessary research areas in computer vision, especially in situations when environmental complexity would tend to cause misidentifications. Wu et al. (2018) argued that it has been obvious how the superior capability in feature extraction provided by DCNN has propelled numerous state-of-the-art computer vision performances. The authors introduce the YOLO model, which regards object detection as a regression problem to speed up the prediction bound with one neural network. Their experiments using the NWPU VHR-10 dataset and a self-collected airport/airplane dataset from Google Earth demonstrate that a YOLO-based model is faster and more precise than traditional detection pipelines. Parallel limitations have been enumerated, poor performance in the case of compactly packed or small objects, positioning errors due to part of the loss function problems and difficult to accurate recognition of an object's position. Proposal of a new model and training algorithm, designed to detect small size objects and to be applied on a compact aerial drone, were made by Moskalenko et al. (2019).

Their approach uses a convolutional Squeezenet network with different layers such as convolutional sparse coding and PCA-transformer. Its training algorithm is multi-stage, including transfer learning and fine-tuning by a simulated annealing algorithm. With its major stride in unsupervised learning, their model achieves an impressive 96 percent detection accuracy on the Inria Aerial Image Labeling dataset, hence marking a key contribution in the field of aerial imagery object detection. Lin et al. (2020) discuss in detail problems related to the detection of small objects in an aerial view, especially in inspection tasks carried out by UAVs.

Their study exposes the limitations of the YoloV3 network at high altitude scenarios with down sampling in the feature maps, which leads to reduced informational integrity. Additionally, further fine-tuning of the YOLOv3 network is added by tuning the anchor values and adding a 4x down sampling prediction layer to improve the converging capabilities and detection precision using the LaSOT dataset.

Pham et al. (2020) provide an extensive review of progress made in object detection from aerial and satellite remote sensing images over the past decade. They underline the success made by deep learning-based approaches but also the persistent difficulties for small object detection in large scenes. The paper introduces YOLO fine, an improved one stage deep learning based detection model for both high accuracy and speed in small object detection, where robustness to new backgrounds is an important factor for domain adaptation in remote sensing applications. Experimental results showing the improvements of YOLO fine over other state of the art detectors.

The techniques of object detection have been significantly explored with the advance in deep learning methodologies. Wang et al. (2019) investigates the authors on small target detection using unmanned aerial vehicles mounted by pan-tilt-zoom cameras through the comparison of traditional support vector machines and YOLO-based neural networks. Their results indicate that while SVMs are constrained by the methods of target area selection, YOLO exhibits better performance in small object detection when modified accordingly.

The authors, Reddy et al. (2024), tested a few CNN based algorithms like Faster R-CNN, YOLO and SSD. SSD has higher processing speed than rest of the models, with higher mAP. This study shows that Faster R-CNN is better than YOLO in terms of accuracy but process at a slow speed. YOLOv3 does well with an mAP score of 0.81.Huang et al. (2024) emphasize optimizations made to the YOLOv8 model for small object detection, showcasing improvements over previous versions like YOLOv3 and YOLOv5n that reinforce its relevance for real-time applications.

Aboyomi & Daniel (2023) discuss the significance of object detection across various applications such as autonomous driving and surveillance, comparing YOLO, SSD and Faster R-CNN models based on their speed and accuracy. Xue (2024), the author showed the shift from the traditional machine learning approaches to the modern deep learning techniques. The result of this study showed that CNNs automate the extraction of features with much better accuracy than that of the previous approach, which was SVM.

Overall, the studies showed a deep understanding of techniques adopted in recent years to detect small objects in aerial images. It also highlighted that there is a strong improvement needed to ever be developing technologies, as there are more practical applications where the enhancement of the detection capabilities are crucial.

2.3 Review on Advanced detection techniques

The detection of small objects in high resolution images has become a common task in modern research. The work of Zhang et al. (2023) deals with this problem by proposing a new adaptive slicing method called ASAHI, which is short for Adaptive Slicing Aided Hyper Inference. It tries to improve the detection performance with limited redundant computation cost. The paper critiques traditional slicing methods and proposes the use of Cluster-DIoU-NMS for better accuracy and speed in postprocessing. Extensive experiments on VisDrone and xView verify that ASAHI achieves the competitive mAP50, while reducing the computation time compared to existing methods.

According to Akyon et al. (2022), recent learning based object detection techniques can be divided into two categories, single-stage and two-stage methods. Among them, it is claimed that single-stage detectors like SSD and YOLO have faster performance, but generally, two-stage methods like Faster R-CNN achieve higher accuracy. This review then brings out the emergence of anchor-free detectors, like FCOS, which do not need any predefined anchor boxes. Despite these advances, the general object detection algorithms still remain inefficient for high resolution images that may contain small and densely packed objects. This has provoked specific approaches aimed at small object detection. Akyon proposes a framework offering a generic solution applicable to existing detectors without major modifications.

Best et al. (2024) explain in detail an open access dataset for wind turbine surface damage, fine-tuning the YOLOv5 and YOLOv7 models to improve detection abilities. This study demonstrates that YOLOv7 outperforms YOLOv5 through superior recall scores and improved F1-confidence metrics. The research makes use of Slicing Aided Hyper Inference (SAHI), by which it presents a novel approach for enhancing model performance toward the detection of smaller instances of damage, as evidence of the importance of real world data in high accuracy detection.

Yu (2024) has shown that the optimization of the RT-DETR model, building a lightweight iRMB backbone and improving the loss function with MPDIoU loss can tackle the small object detection problem. It reduced parameters but with slightly increased mAP and FPS values, hence showing effective improvement in small target detection tasks. Huang & Wang (2024) gives an overview of the real time performance and balance of the small object detection algorithms, focusing majorly on the limitations of the model RT-DETR. The paper emphasizes those techniques which capture multiscale information quite well, proposing fine granular path

augmentation and adaptive feature fusion algorithms to enhance the accuracy by the incorporation of detailed information of various feature levels within models.
Altogether, these studies shows that continuous improvement in techniques for small object detection in various applications and outline the importance of new developments combined with real world data integration in solving challenges in this field.

3. Methodology

This paper adopts an integrated approach for the better detection of small objects in aerial images using state-of-the-art deep learning models, namely YOLO and RT-DETR, with the Slicing Aided Hyper Inference (SAHI) technique. It involves data collection, preparation, model training, and performance evaluation, hence offering a sound framework to appraise the effectiveness of these models in detecting small objects from various classes. Figure 1 demonstrates the procedural flowchart of this study.

3.1 Dataset Description

The VisDrone dataset was the starting point of our investigation, providing a large amount of annotated aerial images that captured most of the intrinsic complexities of real-world object detection scenarios. Considering the inherent challenges within aerial imagery, such as scale variations, diverse orientations, and possible occlusions, a nuanced approach to dataset refinement was implemented. First, the original dataset contained twelve object classes, each of which introduced an initial complexity that could add noise and reduce the precision of the detection.

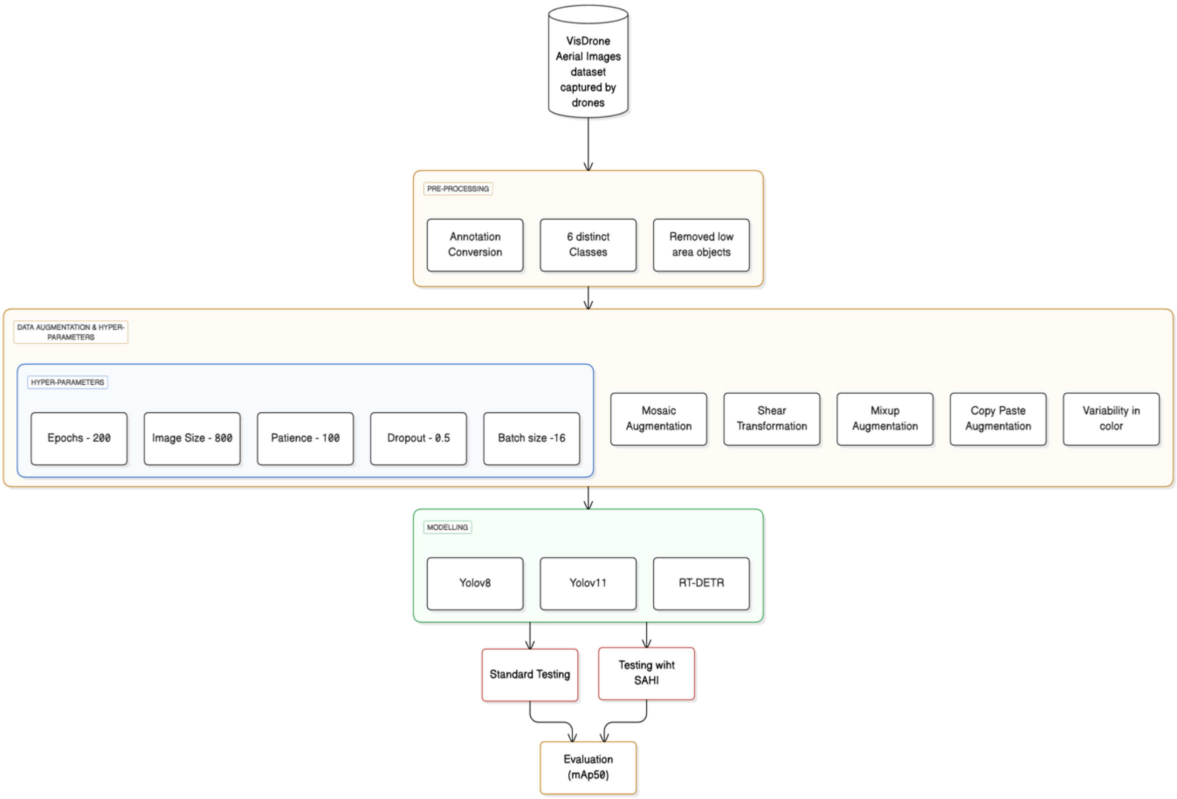
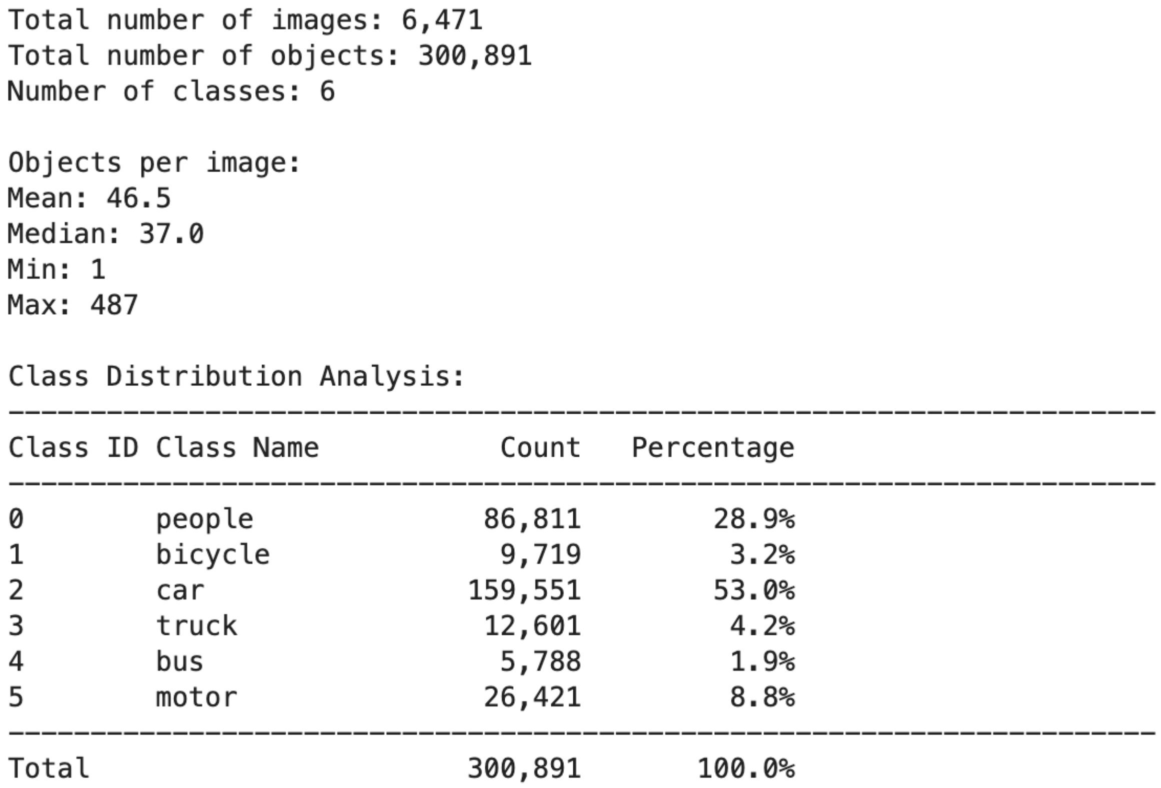


Figure 1. Methodology Flow Chart

3.2 Dataset Preprocessing

3.2.1 Annotation Transformation

Probably one of the most important methodological interventions was providing the transition of the data set from its original form of annotation to a structure compatible with YOLO. That entailed writing a lot of utility functions with complete sophistication, translating those annotations seamlessly without hurting semantic integrity. Transformations in this regard were much more strategic to obtain maximal compatibility with our deep learning architectures of choice.



```
Total number of images: 6,471
Total number of objects: 300,891
Number of classes: 6

Objects per image:
Mean: 46.5
Median: 37.0
Min: 1
Max: 487

Class Distribution Analysis:
----------------------------------------------------------------------
Class ID Class Name          Count   Percentage
----------------------------------------------------------------------
0        people             86,811        28.9%
1        bicycle             9,719         3.2%
2        car               159,551        53.0%
3        truck              12,601         4.2%
4        bus                 5,788         1.9%
5        motor              26,421         8.8%
----------------------------------------------------------------------
Total                      300,891       100.0%
```

Figure 2. Dataset description

3.2.2 Comprehensive Analysis

For maximum functionality, we performed class consolidation and reduction in a structured manner. We cleaned the classification framework by collapsing conceptually related classes such as "Pedestrian" or "People" and "Car" or "Van". Discontinuation of peripheral classes, like "Ignored region" and "Others," made our dataset fine-tune for a more particular and relevant training of deep learning models.

Figure 2. shows the resulting dataset presented a set of very interesting characteristics underlining the complexity of aerial imagery. This dataset, with 6,471 images containing 300,891 total objects, presented a great variation in object density. The statistical profile, with a mean of 46.5 objects per image and a median of 37.0, highlighted the rich informational landscape of the dataset. It might provide insight into the compositional dynamics of this dataset by realizing that one class, "Car," dominates at 53 percent. Figure 3. Shows the class distribution chart.

3.2.3 Precision Enhancement through Filtering

A very important part of our methodology was the introduction of an intensive filtering mechanism for the enhancement of data quality. Systematically removing all the objects that occupied less than 100 pixels, a principled approach to noise reduction was introduced. This strategic filtering was performed for multiple objectives that minimize the low-information data points, reduce potential ambiguities within object detection, and create an optimized training environment biased toward meaningful object representations.

4. Modeling

Three of the most advanced deep learning models have been strategically chosen for this research to address the small object detection challenges in aerial imagery. The methodological approach was thus to capitalize on the unique strengths of each model and hence provide a comprehensive evaluation of advanced

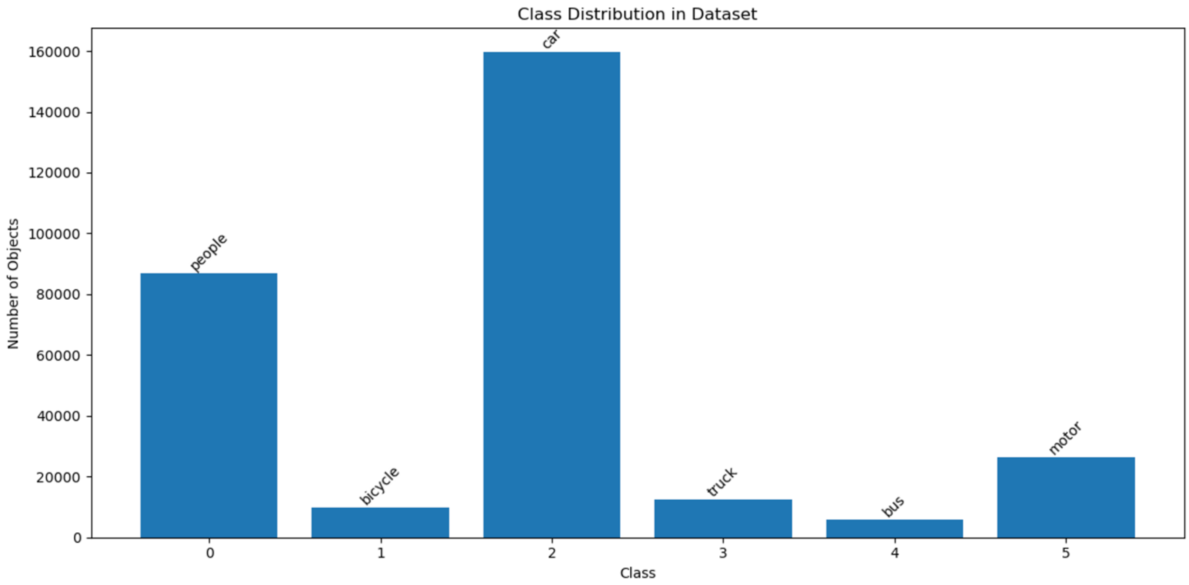


Figure 3. Class distribution chart

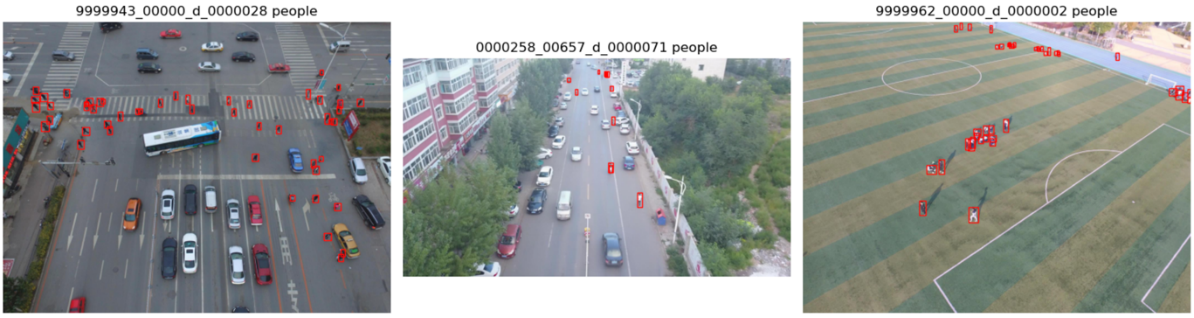

9999943_00000_d_0000028 people
0000258_00657_d_0000071 people
9999962_00000_d_0000002 people

Figure 4. People Annotation in dataset

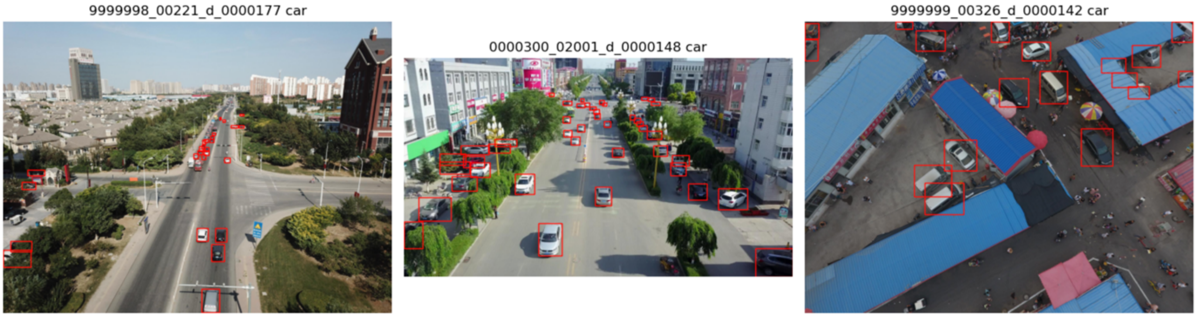

9999998_00221_d_0000177 car
0000300_02001_d_0000148 car
9999999_00326_d_0000142 car

Figure 5. Car Annotation in dataset

object detection techniques. These three models, YOLOv8, YOLOv11, and RT-DETR, are among the very leading edge of computer vision technologies, with different architectural innovations promising enhancements in detection capabilities under challenging imaging environments. Figures 4, 5, 6, and 7, as well as Figure 8 are the annotated objects in the random images from the dataset after dataset preprocessing.

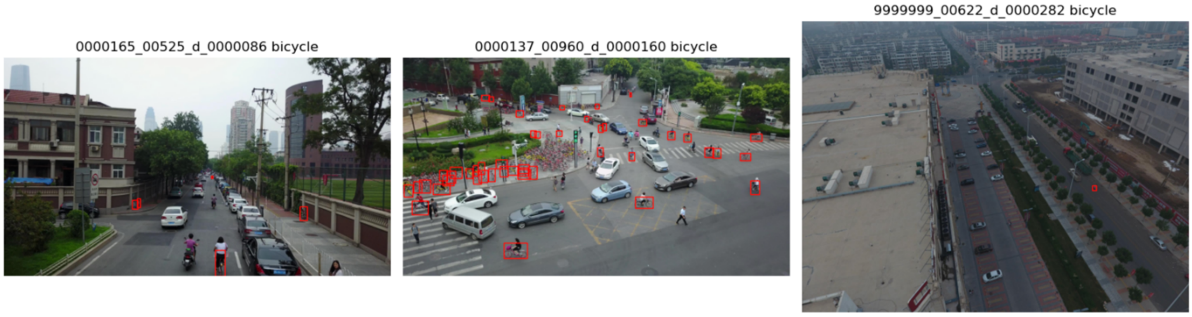

0000165_00525_d_0000086 bicycle
0000137_00960_d_0000160 bicycle
9999999_00622_d_0000282 bicycle

Figure 6. Bicycle Annotation in dataset

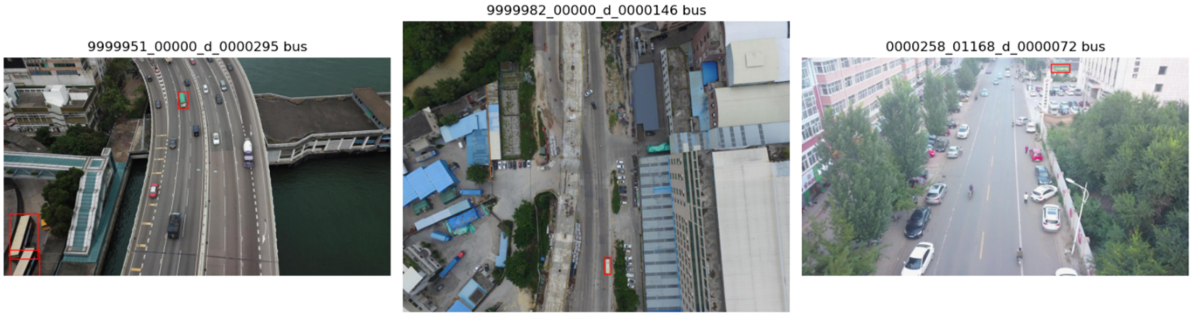

9999951_00000_d_0000295 bus
9999982_00000_d_0000146 bus
0000258_01168_d_0000072 bus

Figure 7. Bus Annotation in dataset

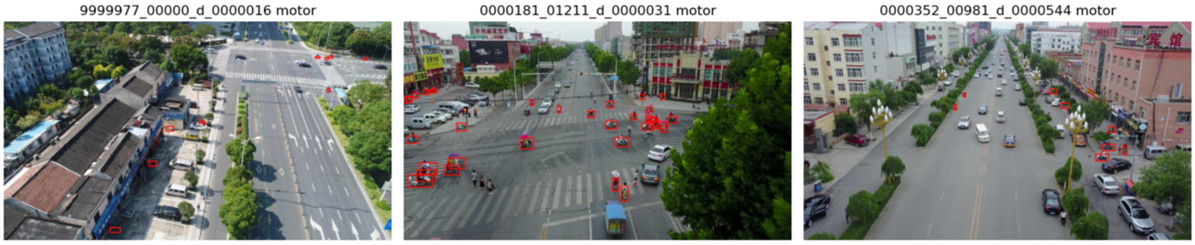

9999977_00000_d_0000016 motor
0000181_01211_d_0000031 motor
0000352_00981_d_0000544 motor

Figure 8. Motor Annotation in dataset

4.1 YOLOv8: Advanced Single Stage Detection Architecture

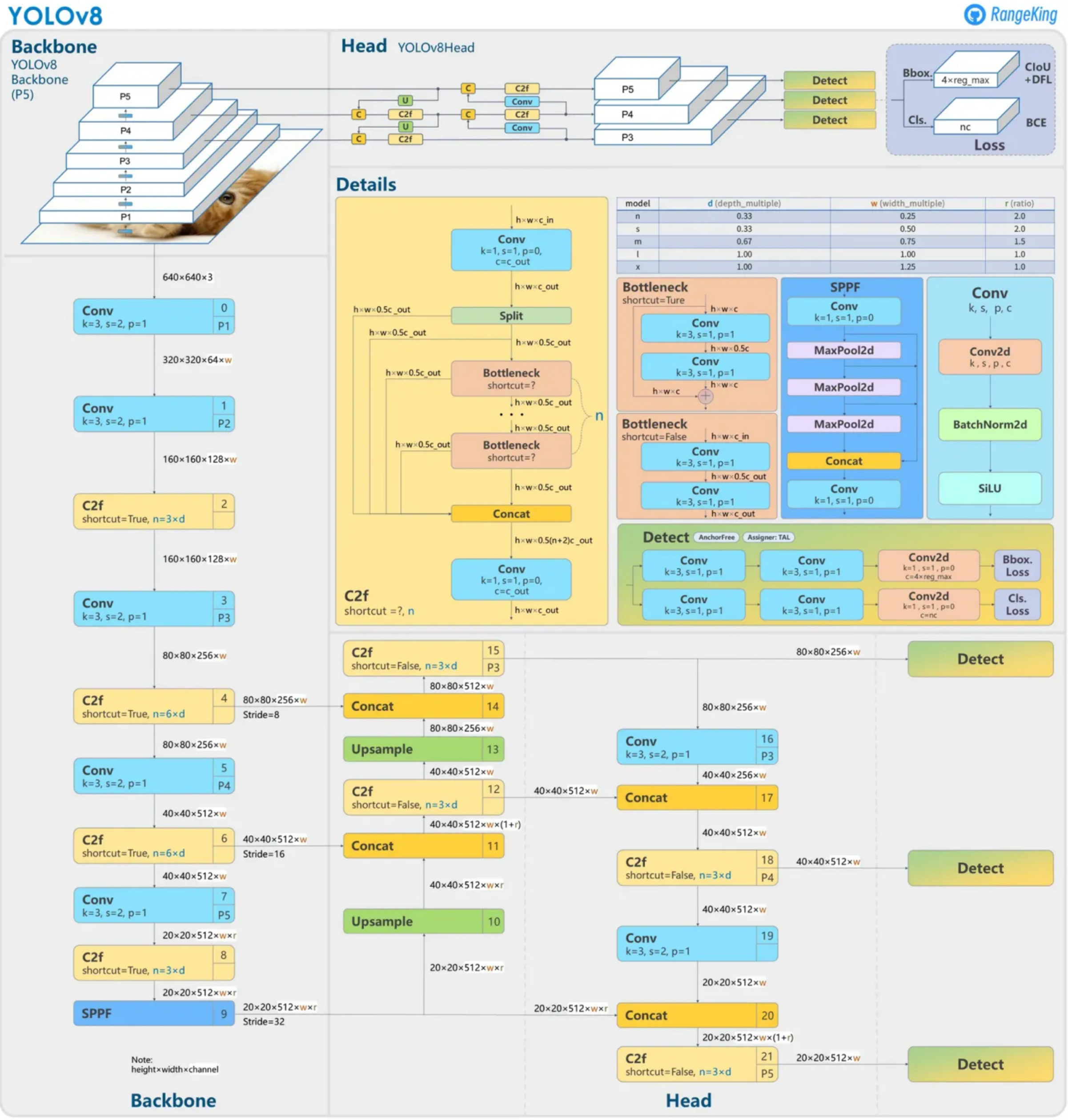
YOLOv8 is the extended development of the well known object detection system called You Only Look Once, outshining others in an incredibly promising way by balancing computing efficiency with accuracy in object detection. Being the great development of the real time object detection methodology, the model architecture will be highly targeted at the solution of specific aerial imagery challenges. By embedding an anchor-free technique with improved feature extraction, YOLOv8 demonstrates impressive performance in detecting small objects in complex visual scenes.

The model design is architectural with a series of innovative convolutional layers, which allow it to process images fast while allowing for object detection with high accuracy. Such efficiency is very critical in drone surveillance and aerial monitoring applications where immediate and exact identification of objects is called for. YOLOv8 is specifically good at processing complex visual landscapes that occur in aerial imagery by way of its ability to extract and prioritize relevant features.

4.2 YOLOv11: Enhanced Multi Scale Detection Architecture

YOLOv11 is a much more polished object detection methodology, building on the core competencies of previous YOLO versions with sophisticated detection improvements. In this model, there is a more advanced multiscale feature aggregation and a sophisticated attention mechanism that will allow the model to be much more sensitive and contextually informed about object detection. These architectural innovations make it possible for YOLOv11 to cope more gracefully with many challenging aspects of the aerial imagery such as variable object sizes, potential occlusions, and variable environmental conditions.

It balances detection accuracy and computational efficiency in modeling, thus rendering a robust framework for real-time object identification. Performing intelligent feature suppression and emphasizing more on relevant visual information, YOLOv11 presents fantastic performance in the detection of small objects that might not be easily possible with classic methodologies related to the detection. The adaptive architecture is one huge leap toward the drawbacks of earlier object detection approaches.



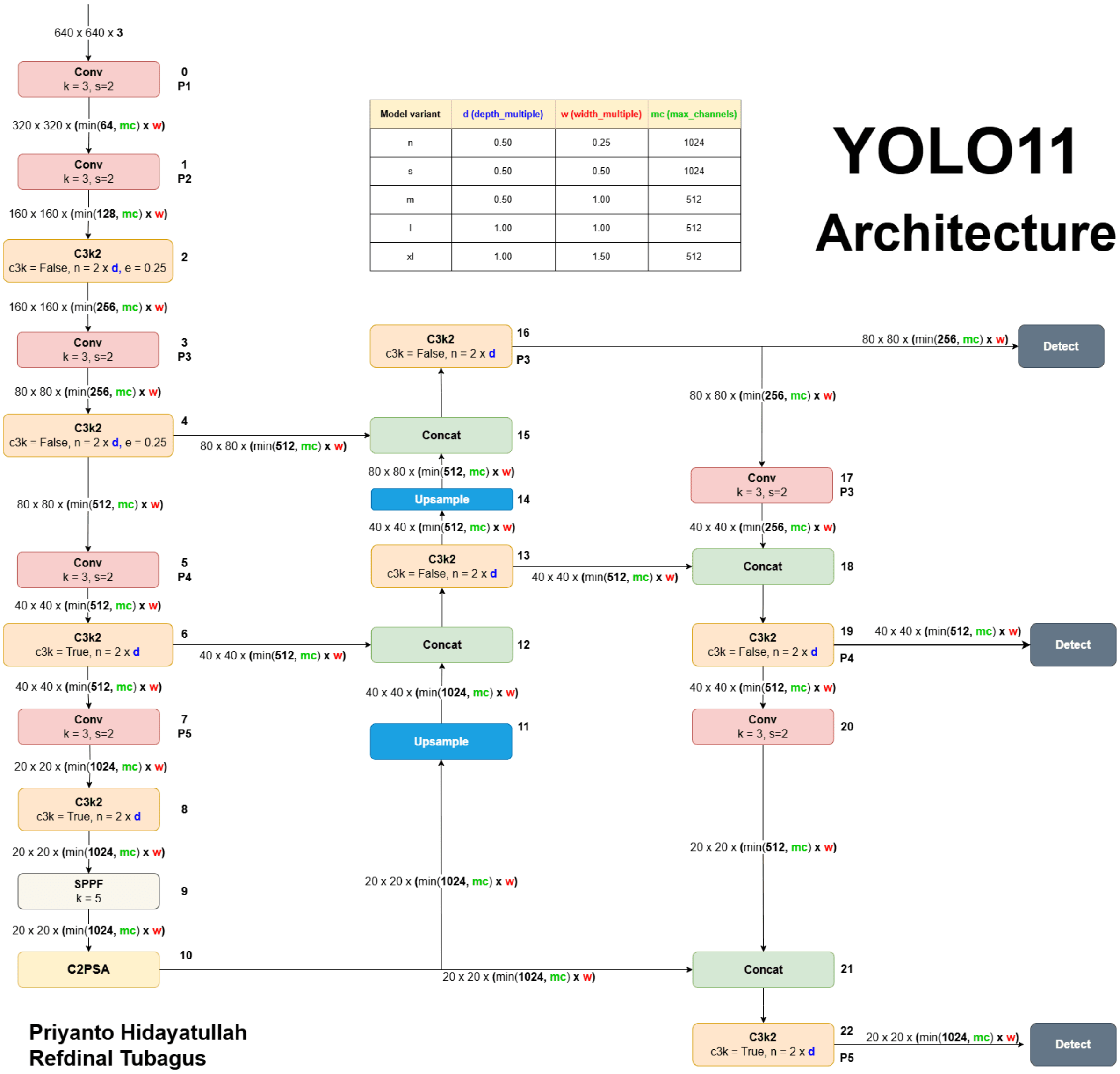
model	d (depth_multiple)	w (width_multiple)	r (ratio)
n	0.33	0.25	2.0
s	0.33	0.50	2.0
m	0.67	0.75	1.5
l	1.00	1.00	1.0
x	1.00	1.25	1.0

Figure 9. YOLO v8 architecture

4.3 RT-DETR: Transformer-Based Detection Paradigm

RT-DETR offers a new take on object detection by combining the strength of transformer architecture with traditional computer vision. The model’s self-attention mechanism deeply analyses input data for a better understanding of the complex images. It captures the long range objects within the images which makes the RT-DETR to break the limitations of traditional approaches using CNN.

The hybrid architecture of the model, combining convolutional layers with transformer blocks, represents the state-of-the-art strategy in small size object detection for challenging imaging environments. The design of RT-DETR enables real time, accurate object detection with computational efficiency for real time applications. It is especially appropriate for aerial imagery analyses where scale, orientation, and environment may affect the precise identification of objects owing to dynamic focusing on critical features.



Model variant	d (depth_multiple)	w (width_multiple)	mc (max_channels)
n	0.50	0.25	1024
s	0.50	0.50	1024
m	0.50	1.00	512
l	1.00	1.00	512
xl	1.00	1.50	512

YOLO11
Architecture
640 x 640 x 3
Conv k = 3, s=2 0 P1
320 x 320 x (min(64, mc) x w)
Conv k = 3, s=2 1 P2
160 x 160 x (min(128, mc) x w)
C3k2 c3k = False, n = 2 x d, e = 0.25 2
160 x 160 x (min(256, mc) x w)
Conv k = 3, s=2 3 P3
80 x 80 x (min(256, mc) x w)
C3k2 c3k = False, n = 2 x d, e = 0.25 4
80 x 80 x (min(512, mc) x w)
Conv k = 3, s=2 5 P4
40 x 40 x (min(512, mc) x w)
C3k2 c3k = True, n = 2 x d 6
40 x 40 x (min(512, mc) x w)
Conv k = 3, s=2 7 P5
20 x 20 x (min(1024, mc) x w)
C3k2 c3k = True, n = 2 x d 8
20 x 20 x (min(1024, mc) x w)
SPPF k = 5 9
20 x 20 x (min(1024, mc) x w)
C2PSA 10
20 x 20 x (min(1024, mc) x w)
Upsample 11
40 x 40 x (min(1024, mc) x w)
Concat 12
40 x 40 x (min(512, mc) x w)
C3k2 c3k = False, n = 2 x d 13
40 x 40 x (min(512, mc) x w)
Upsample 14
80 x 80 x (min(512, mc) x w)
Concat 15
80 x 80 x (min(512, mc) x w)
C3k2 c3k = False, n = 2 x d 16 P3
80 x 80 x (min(256, mc) x w)
Detect
80 x 80 x (min(256, mc) x w)
Conv k = 3, s=2 17 P3
40 x 40 x (min(256, mc) x w)
Concat 18
40 x 40 x (min(512, mc) x w)
C3k2 c3k = False, n = 2 x d 19 P4
40 x 40 x (min(512, mc) x w)
Detect
40 x 40 x (min(512, mc) x w)
Conv k = 3, s=2 20
20 x 20 x (min(512, mc) x w)
Concat 21
C3k2 c3k = True, n = 2 x d 22 P5
20 x 20 x (min(1024, mc) x w)
Detect
Priyanto Hidayatullah
Refdinal Tubagus

Figure 10. YOLO v11 architecture

5. Design Specification

Small object detection in aerial imagery has a research design that follows a two-layered architectural framework of an advanced level, having the Presentation Layer and the Business Layer. It will provide data exploration and visualization by mapping complex computational data into visual insights of dataset characteristics and model performance metrics. It provides scope for the researchers to examine class distributions, object counts, and other such detection nuances through advanced techniques.

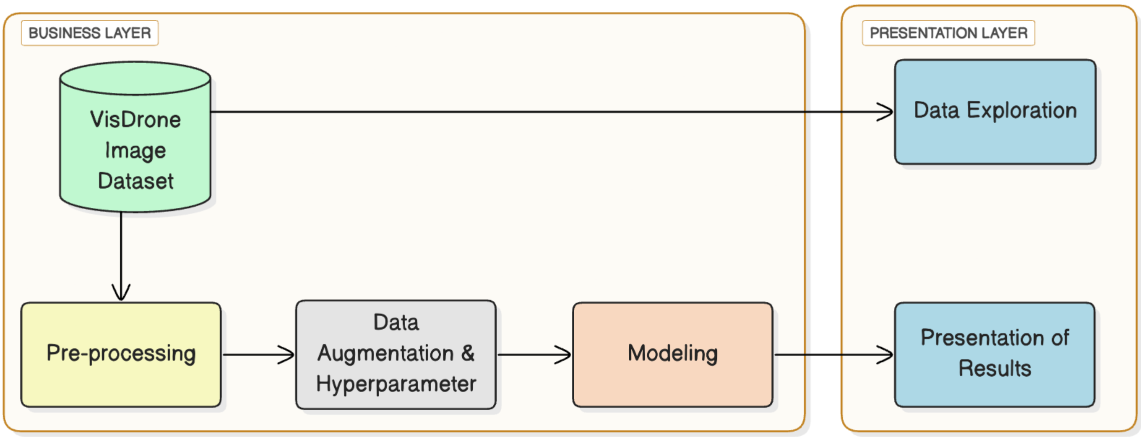


Figure 11. Car Annotation in dataset

The Business Layer has provided a computation infrastructure that does comprehensive data preparation, model training, and systematic evaluation. By performing extensive pre-processing on the VisDrone dataset, converting it to a suitable annotation format, further refining classes, and object filtering, a curated dataset consisted of 6,471 images and 300,891 objects was obtained in six classes. In this paper, three deep learning state-of-the-art models, namely YOLOv8, YOLOv11, and RT-DETR, were implemented, trained up to 200 epochs with performances measured in terms of mean Average Precision (mAP50), loss measurements and Slicing Aided Hyper Inference (SAHI) optimization techniques. This is where the methodology ensures that the research in small object detection capability in aerial imagery has been comprehensively investigated. Figure 9 explains about the design specification for this research.

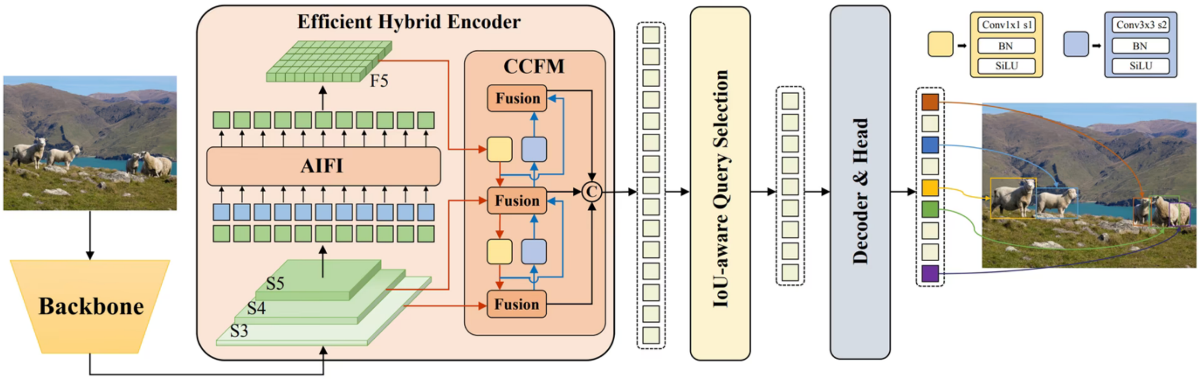


Figure 12. RT-DETR architecture

6. Implementation

This section of the report explains the execution of the model and the outcomes derived from the experiments. The implementation of the system is done by using Python programming language on an AWS EC2 GPU instance, which makes available 24 GB of GPU memory for training. To begin with, the VisDrone dataset is downloaded to the EC2 instance. It contains 6,471 images and 300,891 objects labeled in six classes.

In this work, a unified training strategy was performed, where a similar setting of hyperparameters was used for all the models and each was trained on 200 epochs with an input image size of 800 pixels to capture every minute detail in a computationally efficient manner. Batch size 16, dropout 0.5 and cosine learning rate schedule for allowing the models to reach the best generalization and performance were configured.

These well-designed augmentation strategies have pointed out some of the key problems in small object detection. This paper creates controlled variations by mosaic, shear, color, mixup and copy-paste to largely improve the generalization ability of models on diverse scenes in aerial imaging. This technique effectively utilizes synthetic data generation to enlarge the effective size of a training dataset artificially and thus exposes the models to a huge majority of challenges in detection without going through the process of collecting a large amount of data.

6.1 Data Augmentation and Hyper-parameter

6.1.1 Epochs

One epoch means one full cycle through the training dataset. Setting this to 200 will give ample opportunity for the model to learn from the data by iteratively adjusting weights for minimizing loss. More epochs can lead to better convergence but may also risk overfitting if not monitored properly.

6.1.2 Image Size

The size of the input image has a direct relationship with object detectability over variations in scale. An input size of 800 pixels is a good balance that can provide the model with sufficient detail and an acceptable computational burden.

6.1.3 Close Mosaic

This refers to the number of images used in mosaic augmentation. Mosaic augmentation merges several images into one and helps the model learn from different contexts, enhancing robustness against various object scales and occlusions.

6.1.4 Shear

Shearing in geometric transformation is the skew of an image along one axis. A shear value of 0.1 adds slight distortions that enable the model to generalize better by learning invariant features despite changes in object orientation.

6.1.5 Brightness Gain

This parameter randomly varies image brightness during training. Setting 0.1 allows relatively medium variations in brightness helping in robustness to various lighting conditions.

6.1.6 Mixup

Mixup is an augmentation technique wherein two different images mix with each other with certain random ratios, enhancing generalization of the model due to mixed combinations of different classes presented to the model during its training period.

6.1.7 Copy-Paste

This technique refers to pasting objects from one image onto another to create new training samples that can improve diversity in the dataset and increase the detection performance for rare objects.

6.1.8 Patience

Patience, an early stopping mechanisms, is the number of epochs to wait before concluding that further training is not beneficial based on an improvement in validation loss. Patience of 100 allows for significant fluctuations in performance before concluding that further training is not useful.

6.1.9 Cosine Learning Rate Scheduler

This technique changes the learning rate based on the cosine function in such a manner that its gradual decrease during training is achieved. This has the added benefit of better convergence by avoiding overshooting minima in the loss landscape.

6.1.10 Dropout

Dropout is a regularization technique where a fraction of input units is randomly dropped out during training. to prevent overfitting by making sure the model does not rely on any one feature.

6.1.11 Batch Size

The batch size is the number of samples that are assessed before the model's weights are adjusted. A batch size of 16 strikes a balance between stable gradient estimates and efficient memory usage.

6.2 YOLOv8

YOLOv8s is implemented with a standard architecture because it strikes the best balance between speed and precision for real-time object detection tasks. CSPDarknet53 was used as a backbone pre-training feature extractor to allow the extraction of both complex, small features with large variations in scale. The implementation thus conducts transfer learning by initializing this training using the pre-trained yolov8s.pt weights with an intention to have a very significant improvement in initial detection performances. Figure 10 shows the output of the training.

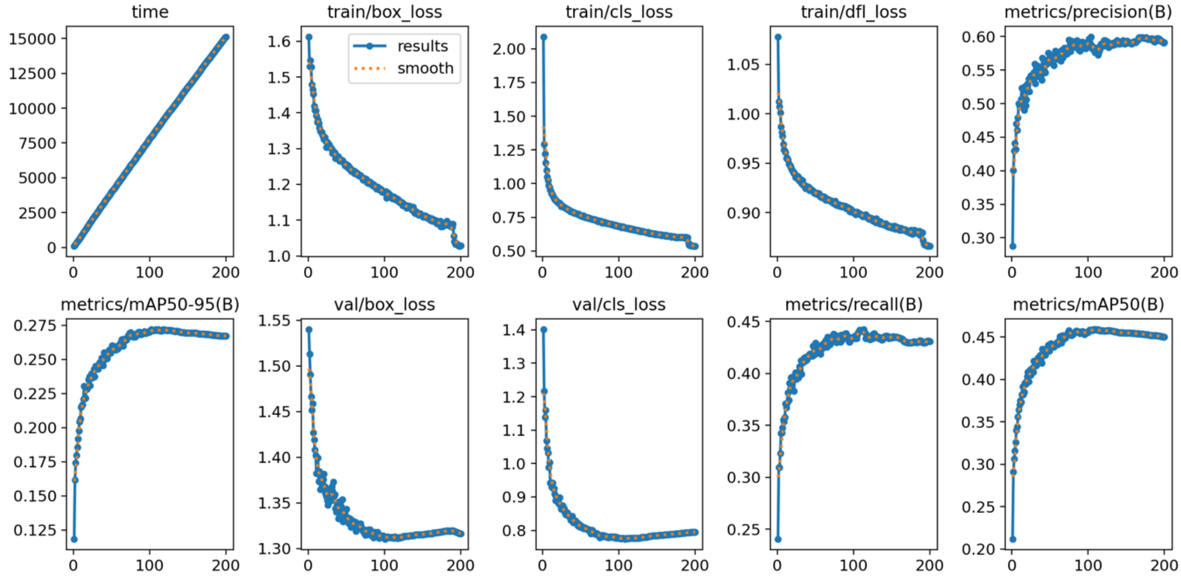


Figure 13. Yolov8 Training result

The model head predicts the coordinates of the bounding box, objectness score, and class probabilities all in one go. Data augmentation techniques were, therefore, aggressively used in this respect to make this model more robust and fine-tuning was performed for a total of 200 epochs. The model was trained for the small objects detection in the VisDrone image dataset and its performance was evaluated by mAP metric.

6.3 Yolov11

The YOLOv11 model updated its predecessors with enhanced features targeted at improving the detection performance and handling occlusion. In contrast to traditional usages, this research modified the top layers of the YOLOv11 architecture to implement custom layers more suitable for aerial imagery classification. Figure 11 shows the output of the training.

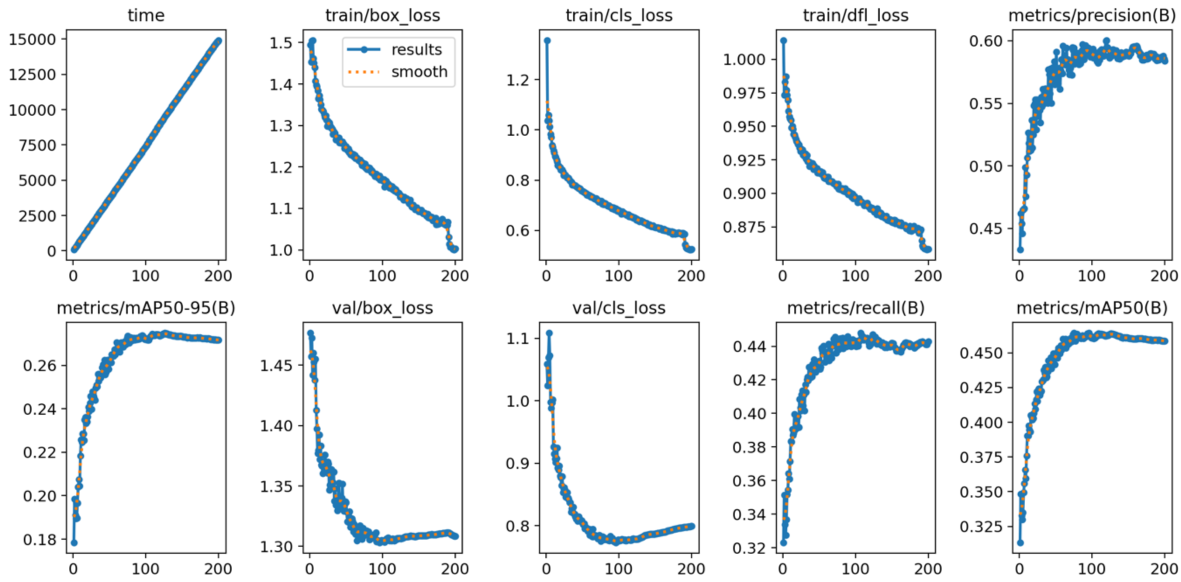


Figure 14. Yolov11 Training result

Extensive training was performed based on `yolo11s.pt` pre-trained weights. The hyperparameters were adjusted a few times through trial and error until the best results were realized in terms of validation accuracy. The key architectural novelties of this model were multi-scale feature aggregation and sophisticated attention mechanisms that made possible more refined object detection.

6.4 RT-DETR

RT-DETR pioneered a transformer-based approach to object detection, bringing with it the ability to capture long range dependencies across complex imagery. The implementation combined convolutional layers with transformer blocks into an elaborate architecture capable of processing intricate aerial scenes impressively efficiently. Figure 12 shows the output of the training.

Starting with the pre-trained model, namely `rtdetr-l.pt`, RT-DETR was trained following a unified training protocol of 200 epochs on VisDrone. Being specific to a transformer architecture, complete features and contextual understanding could be drawn out, thereby perhaps alleviating some challenges in small object detection.

7. Evaluation

7.1 Experiment 1: Performance of the models on standard testing

Three models are compared here for this experiment. RT-DETR was outstanding and showed the best performance, reaching the top in mAP with a wide margin over the YOLOv8 and YOLOv11 for both the 0.50 and 0.50:95 IoU thresholds, then further IoU verification since its final mAP was 0.2695. This would mean that RT-DETR showed a high advantage over the problem of small object detection, possibly due to being better at feature extraction and context understanding. While there is not much difference in performances between YOLOv8 and YOLOv11, the architectural innovations introduced by RT-DETR paid off with tangible performance improvements. The average precision and recall metric also seems to indicate that the performances of RT-DETR are consistently better than others. This work exemplifies the potential of transformer-based approaches in pushing the boundary of object detection capabilities and provides perspectives for further research and model development. Figure 12 shows the annotated images of all 3 models, the confidence score of each annotated object is also marked on those images. These information give us better idea on how to improve the model, especially during the training phase.

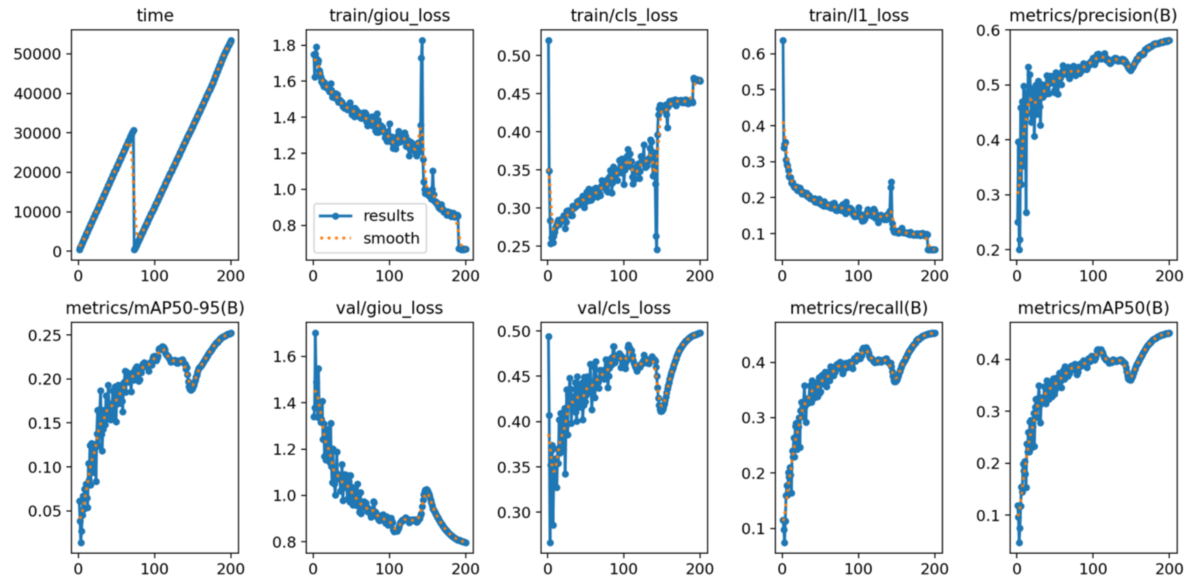


Figure 15. RT-DETR Training Result

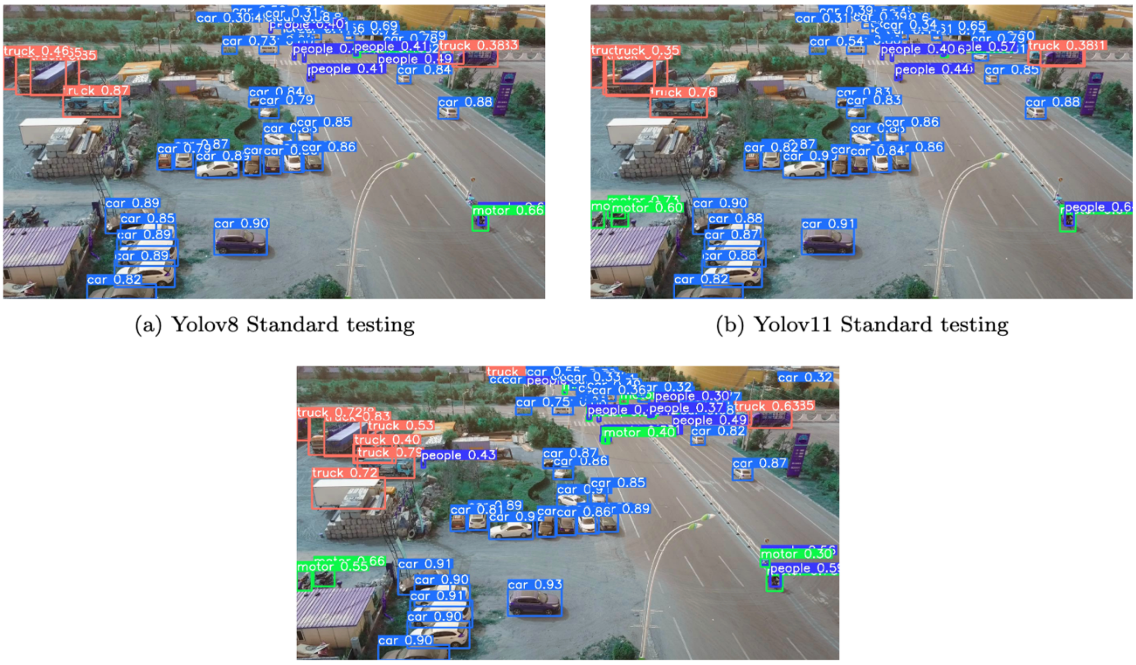
7.2 Experiment 2: Performance of the models on SAHI testing

SAHI has significantly improved the efficiency of small object detection in aerial images. This approach uses a sliding window, segmentation of high resolution images into smaller, possibly overlapping patches, thus helping to improve the identification of small objects. It fills several gaps found in the traditional object detection frameworks and allows models to analyze complex scenes with finer granularity. The SAHI technique is expected to provide a generally applicable approach to improving the detection of small objects across different architectural designs. A comparison of mean recall values reveals that YOLOv11 delivers the best recall of 0.3599, which is slightly better than the recalls achieved by RT-DETR and YOLOv8, which were logged at 0.3546 and 0.3570, respectively, thus establishing an insignificant improvement in the performance of object detection. The results demonstrate the capability of the SAHI sliding window method in tackling small object detection issues. The approach is not trivial in that it enhances deep learning model performance in interpreting intricate aerial images by allowing an image-wide investigation through systematic image partitioning. This study further accentuates the effectiveness of the methodology over different model architectures, hence it is a very important study that will give insight into an advanced strategy to improve the accuracy in object detection, especially in applications where small and complicated objects need to be recognized with care in large and visually complicated scenes. Figure 13 shows the annotated images of all 3 models, the confidence score of each annotated object is also marked on those images.

8. Discussion

The performance of deep learning models YOLOv8, YOLOv11 and RT-DETR were tested in two ways, standard testing and testing enhanced with the Slicing Aided Hyper Inference technique. The results show an improvement in metrics for all the models after the use of SAHI, therefore proving the efficiency of SAHI in enhancing small object detection in aerial imagery.

RT-DETR model obtained the best mAP50 0.4142 under the same testing condition, but YOLOv11 is 0.3719, YOLOv8 was 0.3665. This once more suggests that RT-DETR performed relatively well. Still, in general, the models attained a fair to moderate level of accuracy, given that small objects are challenging to detect within an imbalanced dataset.

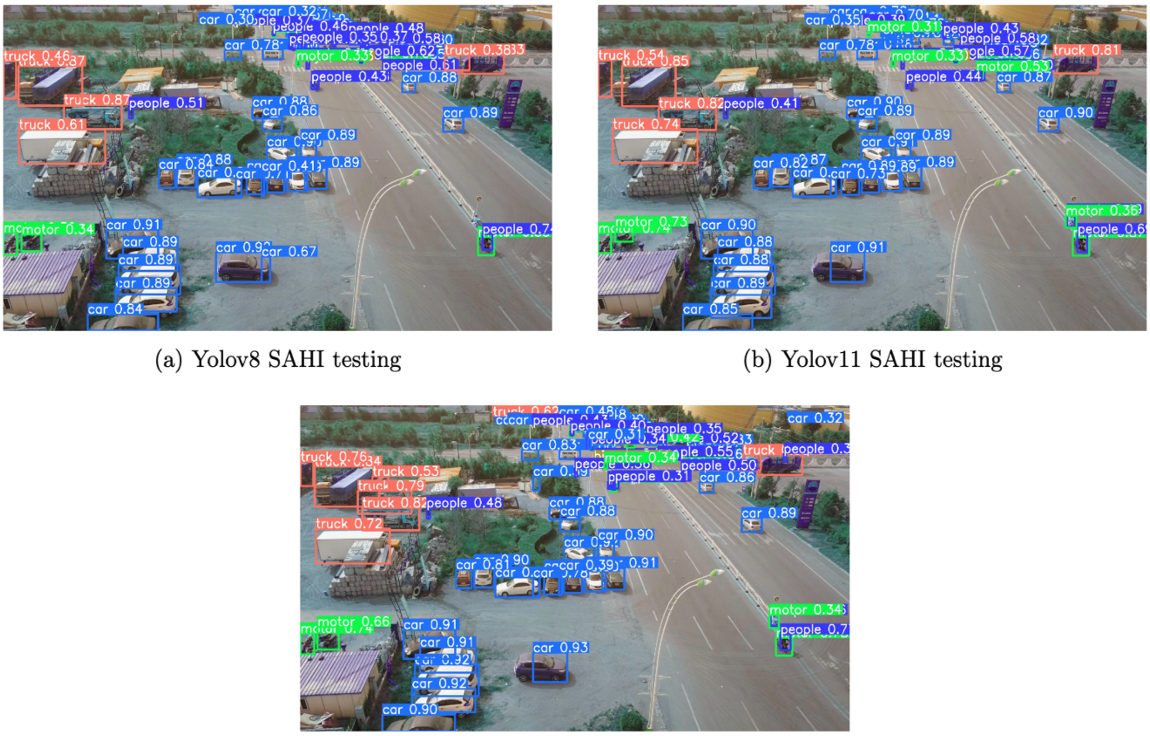


(a) Yolov8 Standard testing

(b) Yolov11 Standard testing

(c) RT-DETR Standard testing

Figure 16. Standard Testing Result



(a) Yolov8 SAHI testing

(b) Yolov11 SAHI testing

(c) RT-DETR SAHI testing

Figure 17. SAHI Testing Result

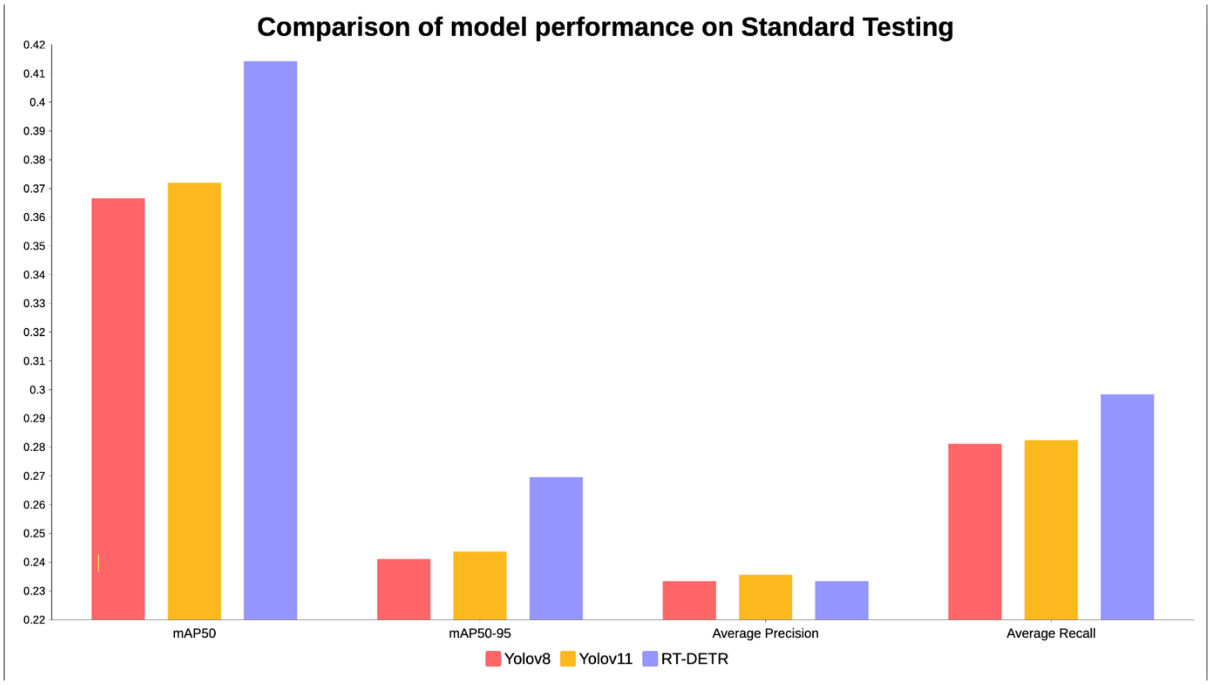


Figure 18. Comparison of Standard Testing Result Chart

Despite, after the introduction of the SAHI method, improvement was witnessed for all models. RT-DETR(SAHI) obtained an excellent mAP50 of 0.4716, which indicates a considerable gain in its small object detection ability. YOLOv11(SAHI) and YOLOv8(SAHI) gained mAP50 of 0.4631 and 0.4573 respectively. The high scores above testify to the fact that SAHI will efficiently overcome the defects of the existing object detection methods through better detail extraction using slicing.

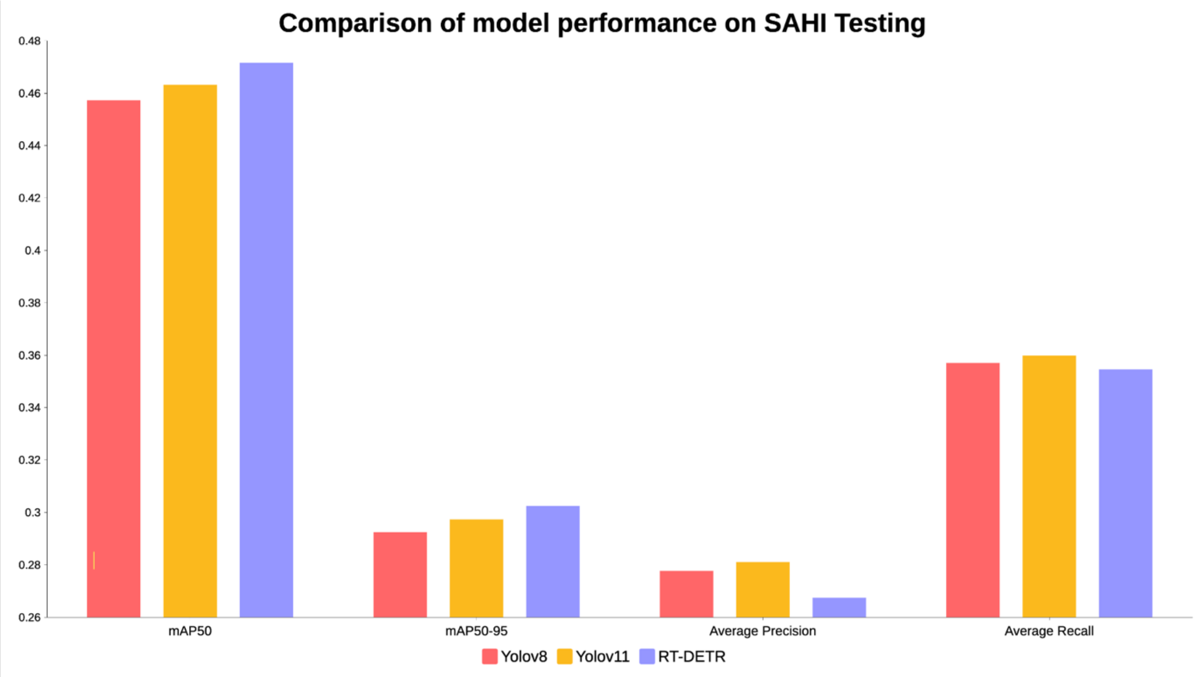


Figure 19. Comparison of SAHI Testing Result Chart

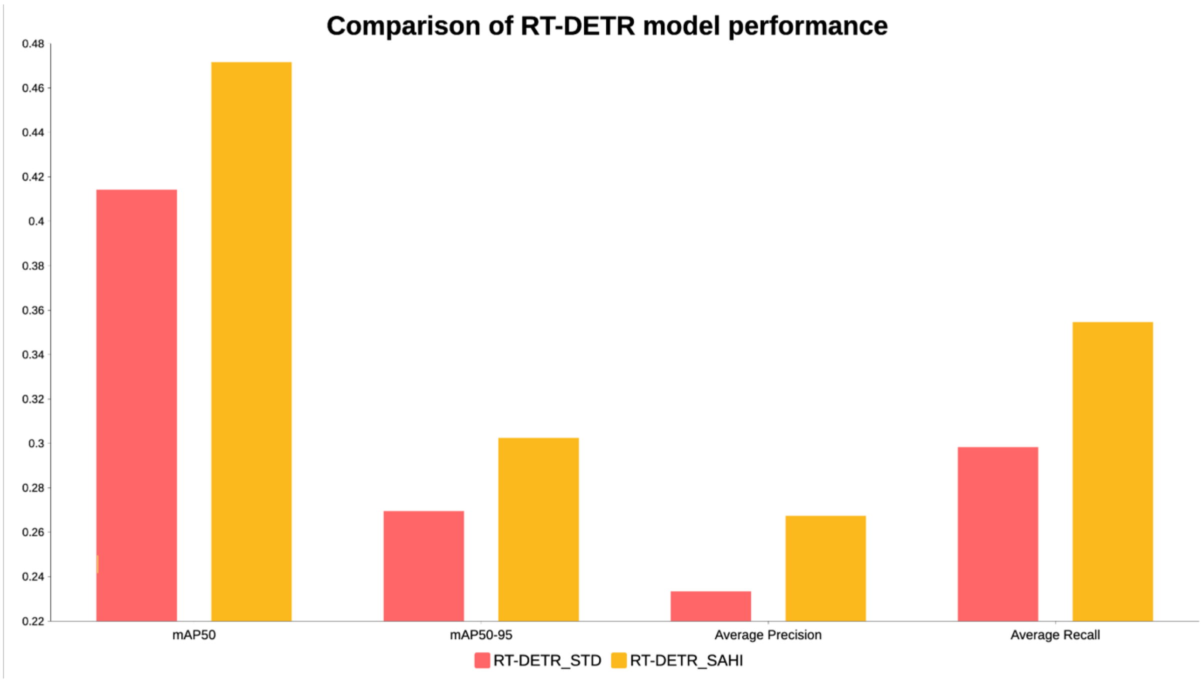


Figure 20. RT-DETR Standard vs SAHI result chart

It also puts forward that the dataset properties have a large influence on model performance, since models demonstrate different strengths and weaknesses depending on the dataset applied. For example, while YOLO models showed dramatic improvements when using SAHI, their baseline performance in normal conditions was relatively less competitive to RT-DETR. The hyperparameters had an impact on the results, even small changes led to visible differences during training. Figure 13 and Figure 14 shows the comparison chart for all the three models based on their performance in standard testing and testing enhanced with SAHI. Both charts clearly shows that RT-DETR performed better than YOLO. Figure 15 shows that RT-DETR along with SAHI drastically improved the mAP score of RT-DETR compared to standard testing.

9. Conclusion and Future Work

Small object detection in aerial imagery is one of the traditional problems within computer vision, for surveillance applications, urban structure planning and environmental concerns. The paper presents research into how the three deep learning models YOLOv8, YOLOv11, and RT-DETR perform for detecting small objects within the VisDrone dataset, powered by Slicing Aided Hyper Inference.

The RT-DETR model achieves its highest mAP50 score of 0.4716, which proves that transformer-based models can really handle small object detection tasks. YOLOv11 model with an mAP50 score of 0.4631, performed better than Yolov8 and the YOLOv8, which had a rise in mAP50 to 0.4573 with SAHI. These comments prove that SAHI is an effective way to improve the small object detection in complex aerial images.

It threw light on an important aspect as well, the dataset composition and the model selection are really key to reaching optimum performance in detection. Moreover, relative benefits from each model convey that some of them can be better suited to some specific applications.

9.1 Future Work

The deep learning models can also be developed further for the detection of small objects in aerial imagery to bring efficiency and effectiveness. Other research directions will be the extension of state-of-the-art methods using semi-supervised and unsupervised learning, which generalize better and accommodate the benefits from the unlabeled data common in aerial datasets. Integration into real time detection systems for surveillance, urban and environmental monitoring will afford greater operational efficiency. Larger and more diversified datasets should be required for experimentation with a view to improving model performance. Increasing the dataset with more object classes, environmental conditions and viewed conditions will enable the models to learn more robust features so that better detection accuracy may be achieved. Inclusion of data taken from different geographical locations and seasons may build a better foundation for training in general. This may open the models to real time processing on edge devices, once optimized for deployment on resource constrained devices and further extend their reach and applications in precision agriculture and smart city projects.

10. References

Aboyomi, D. D. & Daniel, C. (2023), '*A comparative analysis of modern object detection algorithms: Yolo vs. ssd vs. faster r-cnn*', ITEj (Information Technology Engineering Journals) 8(2), 96–106

Akyon, F. C., Altinuc, S. O. & Temizel, A. (2022), '*Slicing aided hyper inference and fine-tuning for small object detection*', Remote sensing pp. 966–970.

Best, O., Khan, A., Gianni, M., Sharma, S. & Collins, K. (2024), '*Wind turbine surface damage detection using yolov7 with slicing aided hyper inference (sahi)*', CoRR pp. 564–576.

Doloriel, C. T. C. & Cajote, R. D. (2023), '*Improving the detection of small oriented objects in aerial images*', CoRR p. 176–185.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014), '*Rich feature hierarchies for accurate object detection and semantic segmentation*'.

Huang, H., Wang, B. & Zhu, T. (2024), '*Improved small-object detection using yolov8: A comparative study*', Applied and Computational Engineering.

Huang, J. & Wang, H. (2024), '*Small object detection by detr via information augmentation and adaptive feature fusion*', CoRR.

Lin, F., Zheng, X. & Wu, Q. (2020), '*Small object detection in aerial view based on improved yolov3 neural network*', CoRR.

Mentari, M., Asmara, R. A., Arai, K. & Oktafiansyah, H. S. (2023), '*Detecting objects using haar cascade for human counting implemented in openmv*', CoRR.

Moskalenko, V. V., Moskalenko, A., Korobov, A. & Zaretsky, M. O. (2019), '*A model and training algorithm of small-sized object detection system for a compact aerial drone*', Radio Electronics, Computer Science, Control (1).

Narayanan, P., Borel, C. C., Lee, H., Lee, H., Kwon, H. & Rao, R. M. (2018), '*A real-time object detection framework for aerial imagery using deep neural networks and synthetic training images*', CoRR 10646, 25–33.

North, S. (2017), '*Salient features, combined detectors and image flipping: an approach to haar cascades for recognising horses and other complex, deformable objects*', CoRR.

Ohn-Bar, E. & Trivedi, M. M. (2016), '*To boost or not to boost? on the limits of boosted trees for object detection*', CoRR pp. 3350–3355.

Ohn-Bar, E. & Trivedi, M. M. (2017), '*To boost or not to boost? on the limits of boosted trees for object detection*', Computer Vision and Pattern Recognition.

Pham, M.-T., Courtrai, L., Friguet, C., Lef`evre, S. & Baussard, A. (2020), '*Yolo-fine: One-stage detector of small objects under various backgrounds in remote sensing images*', Remote Sensing 12(15).

Ramli, M. F. B., Shamsudin, S. S. & Legowo, A. (2017), '*Obstacle detection technique using multi sensor integration for small unmanned aerial vehicle*', Indonesian Journal of Electrical Engineering and Computer Science 8(2), 441–449.

Reddy, S., Pillay, N. & Singh, N. R. (2024), '*Comparative evaluation of convolutional neural network object detection algorithms for vehicle detection*', Journal of Imaging 10(7), 162–162.

Wang, J., Jiang, S., Song, W. & Yang, Y. (2019), '*A comparative study of small object detection algorithms*', CoRR pp. 8507–8512.

Wu, Z., Chen, X., Gao, Y. & Li, Y. (2018), '*Rapid target detection in high resolution remote sensing images using yolo model*', ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences pp. 1915–1920.

Xu, C., Zheng, Y.-G., Zhang, Y., Li, G. & Wang, Y. (2022), '*A method for detecting objects in dense scenes*', Open Computer Science 12(1), 75–82.

Xue, Q. (2024), '*Advancements in object detection: From machine learning to deep learning paradigms*', Applied and Computational Engineering 75(1), 154–159.

Yao, J. & Zhang, Z. (2005), '*Object detection in aerial imagery based on enhanced semi-supervised learning*', CoRR 2, 1012–1017.

Yu, Z. (2024), '*Rt-detr-irmb: A lightweight real-time small object detection method*', CoRR.

Zhang, H., Hao, C., Song, W., Jiang, B. & Li, B. (2023), '*Adaptive slicing-aided hyper inference for small object detection in high-resolution remote sensing images*', Remote sensing 15(5), 1249–1249.