A
Project Report
On

Aerial Image-based Vehicles and Objects Detection

-: Prepared by :-

Arshpreet Singh - 400490171
Rabpreet Singh – 400490143
Sizhe Guan - 400184473

Submitted to
McMaster University
Masters of Engineering
in Systems and Technology
SEP 769 – Cyber Physical Systems

Acknowledgement

We would like to express our sincere gratitude to Professor Dr. Anwar Mirza for their guidance, support, and expertise throughout the duration of this project. Their valuable insights and constructive feedback have been instrumental in shaping the direction and scope of our research. We are deeply appreciative of their commitment to our academic growth and their unwavering dedication to fostering an environment of learning.

We would also like to extend our heartfelt thanks to the Teaching Assistant Monish Mohanan for their invaluable assistance and support. Their prompt responses to our queries, technical guidance, and assistance during practical sessions have been immensely helpful in the successful implementation of our project. We are grateful for their patience, expertise, and commitment to our academic progress.

Abstract

This paper introduces a comprehensive approach to aerial view vehicle detection by leveraging aerial image-based object detection techniques. The research addresses the challenging task of accurately identifying vehicles within aerial perspectives, which has significant implications for applications like urban vehicle management, traffic surveillance, and intelligent transportation systems. Aerial views pose unique challenges due to occlusions, intricate backgrounds, and potential false alarms arising from objects situated atop buildings. To overcome these complexities, the study harnesses the power of advanced deep learning neural networks The results underscore the potential of the proposed framework in significantly improving the accuracy and reliability of aerial view vehicle detection. The study contributes to the body of knowledge by addressing the critical gap in accurate vehicle detection within aerial perspectives. This advancement is poised to bring substantial benefits to diverse domains relying on aerial views for effective vehicle monitoring and management.

Table of contents

1. Introduction

The task of identifying vehicles within aerial images carries significant importance due to its potential to provide crucial insights for a variety of valuable applications, including urban vehicle management, intelligent transportation systems, and traffic surveillance. Nonetheless, the task of detecting vehicles in aerial images is intricate. Vehicles tend to occupy a relatively small portion of pixels within the images, posing a challenge to their detection. Moreover, these vehicles often face obstructions such as trees and road signs, complicating their accurate recognition. The complex urban landscape further compounds the difficulty, with the background exhibiting high complexity, particularly in densely populated urban regions. False alarms can arise from objects positioned atop buildings, adding to the complexities of detecting vehicles in aerial images.

To confront this challenge, machine learning methods have garnered substantial attention, particularly in the context of object detection. These methods involve extracting pertinent features from the images, which serve as the foundation for various detection techniques. A considerable body of research has been dedicated to vehicle detection, exploring a range of methodologies and strategies. This encompasses the integration of different features, each intended to enhance detection accuracy. Notably, researchers have even explored innovative approaches that incorporate the fusion of multiple features, generating composite features to bolster detection performance. Successfully addressing effective vehicle detection within aerial images mandates a thorough comprehension of these machine learning-based techniques and their potential to enhance detection accuracy in intricate real-world scenarios.

The emergence of Deep Learning neural networks has profoundly reshaped the landscape of vehicle detection in aerial images. These sophisticated algorithms leverage intricate layers of interconnected nodes to autonomously learn and extract intricate features from the images, thereby augmenting their capacity to accurately detect vehicles amidst challenging environments. Deep Learning networks, such as Convolutional Neural Networks (CNNs) and their advanced iterations, have demonstrated remarkable proficiencies in discerning intricate patterns, shapes, and structures associated with vehicles, even within cluttered and obscured scenes. By assimilating hierarchical representations from the data, these networks can capture both low-level visual features and high-level contextual information, thus enhancing detection accuracy. Additionally, their adaptability to accommodate extensive datasets and their ability to generalize across diverse scenarios render Deep Learning neural networks a potent instrument for grappling with the intricacies of vehicle detection in aerial imagery. This transformative influence holds the potential to enhance the precision of vehicle detection and open up new avenues for broader applications in remote sensing, urban planning, and transportation management.

2. Problem Statement / Review / Background

Aerial object detection has emerged as a critical research area with applications spanning urban planning, disaster management, environmental monitoring, and defence. The advent of Convolutional Neural Networks (CNNs) has significantly propelled the progress of object detection techniques, and among them, the You Only Look Once (YOLO) architecture has gained remarkable attention due to its real-time capabilities and accuracy. Discussing this application of the YOLO framework, highlighting its strengths, challenges, and advancements, we can see that,u nlike traditional methods that involve region proposal networks, YOLO divides the input image into a grid and predicts bounding boxes, object classes, and confidence scores for each grid cell. This one-shot detection approach allows YOLO to achieve real-time processing speeds while maintaining competitive accuracy. The YOLO architecture has since undergone multiple iterations, with YOLOv3 and YOLOv4 improving detection performance through architectural enhancements and feature fusion techniques.

Aerial object detection introduces unique challenges compared to ground-level scenarios. Objects such as vehicles and buildings appear smaller, and their orientation can be arbitrary due to the overhead perspective. These challenges demand specialized approaches to ensure accurate detection. Traditional object detection methods can struggle with such variances, making YOLO's ability to handle multiple object sizes, aspect ratios, and orientations especially valuable in aerial imagery analysis. In contrast to images captured from a first-person perspective, aerial viewpoints present unique challenges that impact the efficiency of object detection. These challenges encompass several facets:

Diminutive Object Dimensions: Objects within aerial images, particularly vehicles, possess small dimensions, often spanning only 15 to 30 pixels. Despite the extensive coverage of geographical areas in aerial images, the relatively small size of vehicles in relation to the overall image domain poses a significant obstacle. Consequently, these vehicular entities often appear as small structures, potentially blending into the surroundings or neighbouring objects in an imperceptible manner. This challenge arises from pixel resolution limitations, which hinder the accurate depiction of small entities, thus hindering the effectiveness of established object recognition algorithms. Addressing the Diminutive Object Dimensions challenge requires the strategic application of advanced methodologies, including cutting-edge deep learning approaches, the acquisition of high-resolution imagery, and the incorporation of contextually relevant cues. These efforts aim to enhance the accuracy and reliability of vehicle identification within aerial-oriented object detection systems.

Varied Object Orientation: Objects' spatial orientation is not fixed; they can be observed in rotated positions within the image. Given that aerial images provide a top-down view of vast geographical terrains, vehicles may exhibit diverse angular orientations influenced by road contours, parking layouts, and vehicle movements. This orientation diversity poses a challenge for conventional object detection algorithms designed to recognize primarily upright objects. Consequently, vehicles with unconventional tilts or orientations

might face misclassification or even evade detection, resulting in compromised accuracy during the detection process. Confronting the Varied Object Orientation challenge necessitates the development of sophisticated methodologies capable of effectively handling objects with varying orientations. Strategies could involve deploying advanced deep learning architectures equipped to identify invariant features, integrating multi-angle training data, and refining training approaches to enhance algorithms' capability in reliably detecting vehicles with diverse orientations within aerial-based object detection frameworks.

High Image Resolution: Aerial images can exhibit substantial resolutions, often reaching hundreds of megapixels. The extensive size of these images escalates computational requirements and demands robust processing strategies. Although high-resolution aerial images offer comprehensive and detailed representations of terrestrial surfaces, they concurrently introduce complexities to vehicle detection due to the abundance of intricate attributes and finely-detailed textures inherent in these images. This wealth of detail encompasses various objects such as roads, buildings, vegetation, and shading, contributing to visual intricacy. As a result, vehicles with modest dimensions can become obscured within the visual complexity stemming from these intricate details, thereby hindering precise and reliable vehicle identification. Resolving the High Image Resolution challenge necessitates the development of advanced methodologies in image processing and object detection. These methodologies must be adept at distinguishing vehicles within visually intricate surroundings while effectively filtering out extraneous information. Additionally, incorporating contextually-sensitive approaches, multifaceted scale analysis, and feature abstraction becomes crucial in mitigating challenges brought about by high image resolution. This enhancement strategy ultimately enhances the accuracy and efficiency of vehicle detection within aerial-based object detection scenarios.

Limited Availability of Datasets: In the context of aerial view object detection, the availability of suitable datasets for training and validation purposes is limited. This scarcity of annotated data can impede the accurate development of detection models. However, creating comprehensive datasets that encompass a wide range of real-world scenarios featuring diverse environmental conditions, vehicle types, orientations, and scales presents a formidable undertaking. The deficiency of such diverse and high-quality datasets hampers the ability to train algorithms capable of robust generalization across various situations, leading to compromised performance when faced with novel or underrepresented scenarios. Addressing the Limited Availability of Datasets challenge demands collaborative efforts involving researchers, providers of aerial imagery, and domain experts to meticulously curate and expand datasets encompassing representative instances. This collaborative endeavor aims to facilitate the construction of object detection models that proficiently identify vehicles across diverse real-world settings in aerial image-based detection applications.

Nonetheless, aerial perspectives offer advantages in terms of providing accurate distances and known image dimensions, facilitating straightforward dimension calculations. Moreover, the consistent observation angle in aerial imagery simplifies analysis procedures.

3. Theory and Datasets

3.1. Theory

Domain shift remains a significant concern in aerial object detection. Due to differences between training and testing data, models might not generalize well to new environments. To address this, researchers have explored domain adaptation and transfer learning techniques. Pretrained models on ground-level object detection tasks can serve as effective starting points, allowing the model to adapt to aerial imagery features through fine-tuning or feature alignment methods.

The underpinning framework of our strategy to address the complexities inherent in aerial view vehicle detection via aerial image-based object detection resides in the application of the You Only Look Once (YOLO) neural network paradigm. YOLO represents a cutting-edge object detection model renowned for its efficiency in real-time detection undertakings. The architectural composition of YOLO encompasses a singular neural network structure that, in a singular pass, forecasts both bounding boxes encompassing objects and the corresponding class probabilities directly from the input image. The underpinning framework of our strategy to address the complexities inherent in aerial view vehicle detection via satellite image-based object detection resides in the application of the You Only Look Once (YOLO) neural network paradigm. YOLO represents a cutting-edge object detection model renowned for its efficiency in real-time detection undertakings. The architectural composition of YOLO encompasses a singular neural network structure that, in a singular pass, forecasts both bounding boxes encompassing objects and the corresponding class probabilities directly from the input image.

In this project, we have implemented YOLO architecture, specifically YOLOV8 networks. The initial step involves data preprocessing, where we resize the images by 320 x 320 pixels and augment the images by rotation, cropping and grey scaling.
With transfer learning, we basically try to exploit what has been learned in one task to improve generalization in another. We transfer the weights that a network has learned at “task A” to a new “task B.”

The general idea is to use the knowledge a model has learned from a task with a lot of available labelled training data in a new task that doesn't have much data. Instead of starting the learning process from scratch, we start with patterns learned from solving a related task.

Applying YOLO object detection using transfer learning on the VisDrone dataset involves utilizing a pre-trained YOLO model on a different dataset as a starting point and fine-tuning it on the VisDrone dataset. Transfer learning enables you to leverage knowledge gained from one task (source task) to improve performance on a related but different task (target task), even when labeled data for the target task is limited.
Transfer learning with YOLO on the VisDrone dataset aims to leverage the model's pre-existing knowledge from a related task (object detection) while adapting it to the specifics of aerial object detection in the VisDrone dataset. This process helps improve convergence and reduces the need for extensive labeled data.

3.2. Datasets

The VisDrone Dataset is a pivotal resource within aerial view vehicle detection through satellite image-based object detection. Created by the AISKYEYE team at the Lab of Machine Learning and Data Mining, Tianjin University, China, this dataset is foundational in advancing research and innovation in computer vision, particularly about object detection and tracking from aerial viewpoints.

The VisDrone Dataset encompasses an extensive collection of annotated aerial images and videos meticulously curated to emulate real-world scenarios encountered in surveillance, monitoring, and aerial imagery analysis. This dataset notably encapsulates diverse attributes, including variations in weather conditions (ranging from sunny to rainy, foggy, and hazy), environmental contexts (urban and rural landscapes, crowded areas), and many objects of interest, most notably vehicles.

The VisDrone Dataset provides a rich representation of the complexities inherent in aerial surveillance scenarios by offering high-resolution images and videos captured from drones. Each instance within the dataset is meticulously annotated with ground truth data, encompassing bounding box coordinates and class labels, thus furnishing a robust foundation for training and evaluating object detection algorithms. It consists of 10 labels including:

0: pedestrian
1: people
2: bicycle
3: car
4: van
5: truck
6: tricycle
7: awning-tricycle
8: bus
9: motor

Researchers and aerial view vehicle detection practitioners benefit from the VisDrone Dataset's substantial contribution to developing accurate and efficient detection and tracking models. The dataset's comprehensive annotations and diversity of scenarios enable the evaluation of model performance across varied dimensions, including object scale, orientation, occlusion, and environmental conditions. As a result, the VisDrone Dataset is crucial in enhancing the precision, robustness, and generalization capabilities of machine learning models and algorithms tasked with vehicle detection within the satellite image-based object detection frameworks.

4. Implementation Details

YOLO Model Configuration

Implementing the You Only Look Once (YOLO) model followed a meticulously structured configuration to ensure accurate and efficient aerial view vehicle detection through satellite image-based object detection. We employed the YOLOv8[2] architecture due to its proficiency in handling complex object detection scenarios. The model was configured with specific anchor box sizes tailored to accommodate the range of vehicle dimensions encountered in satellite imagery. The network parameters were tuned to optimize detection precision while ensuring computational efficiency during real-time inference.

Data Preprocessing

Before model training, a systematic data preprocessing pipeline was established. The aerial images from the VisDrone Dataset[1] were resized to a consistent input size to facilitate uniform processing across the YOLO network. Originally, the dataset images were 2000X1500 pixels in size and required a massive amount of computational power and time to be processed. The size was reduced to 320X240 pixels for all the images to facilitate smoother processing and model training.

The original annotations for all images were transformed into YOLO format from PASCAL VOC format to be utilized by the YOLO model effectively.

Training and Optimization

The YOLO model was trained using a custom dataset using the pre-trained weights of yolov8x.pt model. The training dataset was partitioned into training, and validation sets, with 90%, 10% of the data, respectively. During training, data augmentation techniques, including random rotations, scaling, and horizontal flipping, were applied to augment the diversity of the training data and enhance the model's generalization ability.

Performance Metrics

The trained YOLO-based vehicle detection model was evaluated using standard performance metrics. The primary performance indicator was the mean average precision (mAP) at different intersections over union (IoU) thresholds. Moreover, we analyzed precision-recall curves to understand the model's behavior across varying detection confidence thresholds. The evaluation process encompassed qualitative and quantitative visual assessments using these metrics.

The F1-Confidence Curve is a useful visualization for evaluating the trade-off between precision and recall in object detection tasks, specifically when using the YOLOv8 model. The curve helps you choose an appropriate detection threshold that balances the precision (accuracy of positive detections) and recall (coverage of actual positives) of your model.

The confusion matrix serves as a comprehensive evaluation metric for multiple object detection tasks, providing a structured analysis of the model's performance by quantifying the predictions it makes in relation to ground truth labels. Comprising four key components – true positives, false positives, true negatives, and false negatives – the matrix captures the model's ability to correctly identify positive instances, as well as its tendency to misclassify or overlook objects. By categorizing detection outcomes, the confusion matrix offers insights into the trade-offs between precision, recall, and accuracy.

The Precision-Recall curve stands as a pivotal evaluation metric for assessing the performance of multiple object detection models by examining the interplay between precision and recall. Precision reflects the proportion of correctly predicted positive instances

among all predicted positives, emphasizing the accuracy of positive predictions. On the other hand, recall represents the ratio of correctly predicted positive instances among all actual positives, emphasizing the model's ability to capture all positive instances. The Precision-Recall curve visually portrays the trade-off between precision and recall for varying detection thresholds, offering a comprehensive understanding of how adjustments to the threshold impact the model's performance.

5. Explanation of the Source code

Before training and evaluating YOLO model on our custom dataset, we need to preprocess the data. The following steps were followed for getting the data in YOLO compatible format, each with its own source code:

1. The size of images was reduced from 2000 x 1500 pixels to 320 x 240 pixels. For this purpose, the script traversed the entire folder with original data images and reduced the size of each image. Pillow (PIL) library was used with LANCZOS algorithm to accomplish this.

2. The annotations of the images were transformed into YOLO format. For each annotation CSV file, the original annotation was read and a method to transform the annotations was called. This method returns the transformed annotations which were stored in a new file each for every original annotation.

To train a YOLOV8 model we followed the following steps:

1. Ultralytics HUB is a breakthrough machine learning and deployment platform, and YOLO AI models are developed within it. Install dependencies for YoloV8 models:

```
import torch
import ultralytics
ultralytics.checks()
```

2. Verify whether the GPU is activated to train a large neural network model:

```
print(torch.cuda.is_available())
```

3. Modified the configuration of YOLO model, here we decide to use V8x model, 16 batch, 25 epochs and 320 x 240 image size:

```
task: detect
mode: train
model: yolov8x.yaml
data: data.yaml
epochs: 25
patience: 50
batch: 16
imgsz: 320
```

4. Train YOLOv8x model on VisDrone dataset for 25 epochs:

```
!yolo train model=yolov8x.pt data=data.yaml epochs=25 imgsz=320
```

5. Export a YOLOv8 model to any supported format below with the format argument, i.e. format=PyTorch:

!yolo export model=yolov8x.pt format= PyTorch

6. Results and Discussion

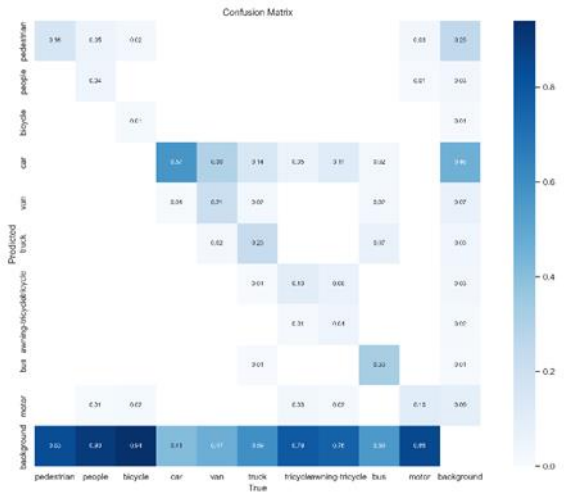

Figure 1: Confusion Matrix of model performance on Training dataset

From the Figure 1 above, the result has illustrated that there are multiple cases is under the False Negative cases. It could happen because of missing object characteristics analysis: Examine the characteristics of the objects that were missed by the model. Consider their size, shape, orientation, occlusion level, and surrounding context. Objects that are small, partially occluded, or have complex shapes might be more challenging for the model to detect accurately.

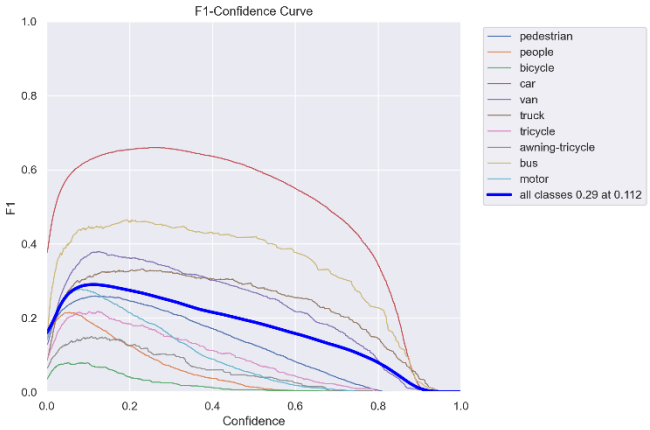

Figure 2: F1-Confidence Curve for YOLOv8 model on Training dataset.

The Figure 2 has demonstrated that the best confidence threshold value is 0.29 after deploying the YOLOv8 model on training dataset. Also, it offers a view in deciding whether the project should focus on high precision or high recall. High Precision if the goal is to minimize false positives since false alarms are costly. High Recall If comprehensive detection coverage is crucial, select a lower threshold that maximizes recall. This is useful when missing positive instances has severe consequences.

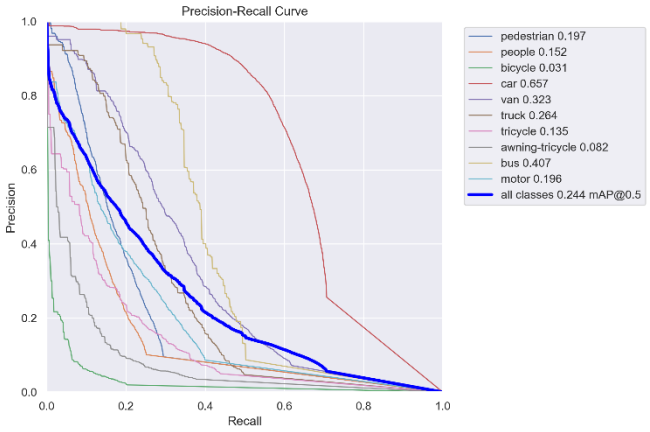

Figure 3: Precision and Recall curve on Training dataset

As Figure 3 shows, Interpreting Curve Movements can tell that the car in Precision-Recall curve generally performs better than other categories. Its curve is shifted towards higher precision values, it indicates that the model is making more confident detections, leading to fewer false positives but possibly sacrificing recall.

Figure 4: Sample results of prediction on training images

In the domain of object detection within aerial-based vehicle detection, it is common to encounter multiple instances of vehicles coexisting within a singular training image, as shown in Figure 4. This circumstance introduces intricate intricacies that possess the potential to exert a substantial influence on the efficacy of the model's performance. The presence of many instances within a given image amplifies the visual complexity and results in the emergence of overlapping bounding boxes. This situation can potentially perplex the model during its training and inference phases. The model might encounter challenges when attempting to accurately discriminate and localize each distinct vehicle entity amid the interwoven objects, subsequently giving rise to difficulties concerning the precise delineation of object boundaries and accurate classification.

Furthermore, the model's capacity to assimilate distinctive features and patterns linked with individual vehicles may become constrained due to the prevalence of numerous instances within a solitary image. This, in turn, can give rise to instances of misclassification, diminished detection precision, and complications in extrapolating to unobserved instances. To confront these challenges, meticulous data preprocessing, application of augmentation techniques, and careful consideration of model architecture becomes imperative. Strategically devised approaches that augment the model's ability to navigate through crowded scenarios, including methods such as non-maximum suppression and multi-object training strategies, serve to ameliorate the undesirable consequences stemming from the presence of multiple instances within training images, thereby enhancing the holistic performance of the model.

7. Recommendations for Future work

In aerial-based vehicle detection, the challenge of accurately detecting and localizing diminutive vehicle instances within high-resolution aerial images warrants focused attention. To address this, future research endeavors should be channeled toward refining the capabilities of small object detection within the existing model framework. This could involve exploring and incorporating innovative architectural designs and strategies tailored to enhance the detection of small objects. Attention mechanisms, which emphasize relevant features while suppressing noise, could be harnessed to amplify the model's responsiveness to small vehicle instances. Additionally, adopting anchor-free approaches, which alleviate the need for predefined anchor boxes and accommodate objects of varying sizes, could enhance the model's adaptability to diverse vehicle dimensions. Such advancements could significantly elevate the model's accuracy in detecting and localizing smaller vehicles, enriching its performance in intricate real-world scenarios.

The limited availability of comprehensive and diverse datasets poses a notable constraint in developing robust vehicle detection models for aerial imagery. Future research endeavors should prioritize the augmentation and diversification of training datasets to address this limitation. Creating and curating expansive datasets that encompass a wider array of real-world scenarios, including various geographic landscapes, lighting conditions, and vehicle orientations, should be a key focus. By integrating instances that mirror the complexities of actual operational environments, the model's training process can be imbued with a higher degree of realism and diversity. This augmented dataset can empower the model to capture the intricacies associated with a broader spectrum of situations, enhancing its capacity to generalize and accurately detect vehicles across unseen and complex environments.

In conclusion, the proposed recommendations for future work in object detection within aerial-based vehicle detection underscore the need for advancements in key domains. By refining small object detection capabilities, expanding and diversifying datasets, researchers can contribute to developing more accurate, adaptable, and effective models for vehicle detection in aerial imagery.

9. References

1. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2021). Detection and Tracking Meet Drones Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1. https://doi.org/10.1109/TPAMI.2021.3119563
2. Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics