Aerial Image-based Vehicles and Objects Detection

Rabpreet Singh
W.Booth School of Engineering McMaster
University, Hamilton, Canada
singr78@mcmaster.ca

Arshpreet Singh
W.Booth School of Engineering McMaster
University, Hamilton, Canada
singa322@mcmaster.ca

Sizhe Guan
W.Booth School of Engineering McMaster
University, Hamilton, Canada
guans9@mcmaster.ca

***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.**

Keywords—Aerial image, Aerial objection detection, Vehicle detection, Image segmentation, object detection, Feature extraction.

I. Introduction (*Heading 1*)

The task of detecting vehicles within aerial images holds significant importance due to its potential to contribute crucial information to a multitude of valuable applications, including urban vehicle management, intelligent transportation systems, and traffic surveillance. However, vehicle detection within aerial images is complex. Vehicles typically occupy relatively fewer pixels within the images, challenging their detection. Moreover, these vehicles are often impeded by various obstructions, such as trees and road signs, complicating their accurate identification. The intricate urban landscape further exacerbates the challenge, with the background exhibiting high complexity, particularly in densely populated urban areas. False alarms can arise from various objects situated atop buildings, adding to the intricacies of vehicle detection within aerial images.

To address this challenge, machine learning methods have garnered substantial attention, particularly within the realm of object detection. These methods involve extracting relevant features from the images, which serve as the basis for diverse detection techniques. A substantial body of research has been dedicated to vehicle detection, exploring a range of methodologies and strategies. This includes the incorporation of various features, each intended to enhance the detection accuracy. Notably, researchers have even explored novel approaches that involve the fusion of multiple features, creating composite features to bolster detection performance. Pursuing effective vehicle detection within aerial images necessitates a comprehensive understanding of these machine learning-based techniques and their potential to enhance detection accuracy in complex real-world scenarios.

The advent of Deep Learning neural networks has significantly transformed the landscape of vehicle detection in aerial images. These sophisticated algorithms leverage intricate layers of interconnected nodes to automatically learn and extract complex features from the images, thus enhancing their ability to accurately identify vehicles amidst challenging environments. Deep Learning networks, such as Convolutional Neural Networks (CNNs) and their advanced variants, have exhibited remarkable capabilities in discerning intricate patterns, shapes, and structures associated with vehicles, even within cluttered and occluded scenes. By learning hierarchical representations from the data, these networks can capture low-level visual features and high-level contextual information, enhancing detection accuracy. Furthermore, their adaptability to accommodate large datasets and the ability to generalize across diverse scenarios make Deep Learning neural networks a potent tool for addressing the intricacies of vehicle detection in aerial imagery. This transformative influence holds the promise of improving the accuracy of vehicle detection and unlocking new avenues for broader applications in remote sensing, urban planning, and transportation management.

II. Problem Statement And Background

In contrast to images captured from a first-person perspective, aerial views present distinct challenges that impact the effectiveness of object detection. These challenges encompass various aspects:

Diminutive Object Dimensions: Objects within aerial images, particularly vehicles, are characterized by small dimensions, often spanning only 15 to 30 pixels. The encompassing aerial images encapsulate extensive geographical expanses, yet the comparatively inconspicuous proportions of vehicles in relation to the comprehensive image domain engender a substantial impediment. Consequently, these vehicular entities frequently manifest as diminutive constructs, potentially commingling with the ambient milieu or neighboring objects in an imperceptible manner. The genesis of this quandary can be traced back to the restrictions imposed by pixel resolution, which, in turn, obfuscates the meticulous depiction of diminutive entities, thereby encumbering the efficacy of established object recognition algorithms. Effectuating a resolution to the Diminutive Object Dimensions challenge mandates the adroit deployment of sophisticated methodologies, encompassing cutting-edge deep learning paradigms, procurement of imagery characterized by elevated resolutions, and the judicious amalgamation of contextually relevant cues. Such endeavors are oriented towards augmenting the precision and dependability of vehicle identification within the framework of aerial-oriented object detection systems.

Varied Object Orientation: The spatial orientation of objects is not fixed, and they can appear rotated in the image. Given that these images provide an overhead perspective of extensive geographical terrains, vehicles may assume diverse angular positions, orientations, and tilts influenced by roadway topography, parking arrangements, and vehicular locomotion. This diversity in orientation engenders a notable impediment for conventional object detection algorithms, which are inherently tailored to recognize objects predominantly in conventional, upright orientations. Consequently, vehicles characterized by unconventional tilts or orientations may be subject to misclassification or even evaded detection, culminating in compromised accuracy throughout the detection process. Addressing the Varied Object Orientation challenge necessitates the formulation of sophisticated methodologies capable of adeptly handling objects with divergent orientations. Such strategies may encompass the utilization of advanced deep learning architectures endowed with the capacity to discern invariant features, coupled with the assimilation of multi-angle training data and the refinement of training paradigms to bolster the algorithms' aptitude in reliably identifying vehicles with variegated orientations within the realm of aerial-based object detection frameworks.

High Image Resolution: Aerial images can possess substantial resolution, often extending into hundreds of megapixels. The sheer size of these images increases computational demands and necessitates robust processing strategies. Although high-resolution aerial images offer an exhaustive and nuanced depiction of terrestrial surfaces, they concurrently introduce intricacies into the process of vehicle detection due to the profusion of intricate attributes and finely-detailed textures inherent to these images. This richness of detail encompasses a diverse array of objects, including thoroughfares, edifices, flora, and shading, thereby fostering a milieu of visual intricacy. Consequently, vehicles of modest dimensions may become subsumed within the visual complexity engendered by these intricate details, consequently impeding the accurate and dependable identification of vehicles. The amelioration of the High Image Resolution challenge mandates the formulation of sophisticated methodologies in image processing and object detection, capable of discriminating vehicles within visually intricate surroundings while adeptly winnowing out extraneous information. Additionally, the assimilation of contextually-sensitive approaches, multifaceted scale analysis, and feature abstraction emerges as imperative in alleviating the predicaments introduced by high image resolution, ultimately enriching the precision and effectiveness of vehicle detection within the domain of aerial-based object detection scenarios.

Limited Availability of Datasets: The availability of suitable datasets for training and validation purposes is restricted in the context of aerial view object detection. This scarcity of annotated data can hinder the development of accurate detection models. However, the creation of all-encompassing datasets, comprising a broad spectrum of real-world scenarios encompassing various environmental circumstances, vehicle typologies, orientations, and scales, presents itself as a formidable undertaking. The shortage of such multifaceted and high-caliber datasets impairs the capacity to train algorithms endowed with the ability to generalize adeptly across various scenarios, ultimately resulting in compromised performance when confronted with unprecedented or unrepresented situations. Addressing the Limited Availability of Datasets predicament mandates a collaborative endeavor involving researchers, providers of aerial imagery, and domain specialists, to meticulously assemble and expand datasets that encompass representative instances. This endeavor aims to facilitate the construction of object detection models that aptly discern vehicles across a spectrum of real-world settings in aerial image-based detection applications.

However, aerial views offer advantages in terms of the availability of real distances and known image dimensions, facilitating straightforward dimension calculations. Additionally, the consistent observation angle in aerial imagery simplifies the analysis.

III. Methodology

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 labeled 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

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 aerial 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 aerial imagery. The network parameters were tuned to optimize detection precision while ensuring computational efficiency during real-time inference.

Datasets: VisDrone Dataset t[1]

The VisDrone Dataset t[1] is a pivotal resource within aerial view vehicle detection through aerial 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, and contains over 7000 images.

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

this dataset is foundational in advancing research and innovation in computer vision, particularly about object detection and tracking from aerial viewpoints.

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.

IV. Results and Discussion

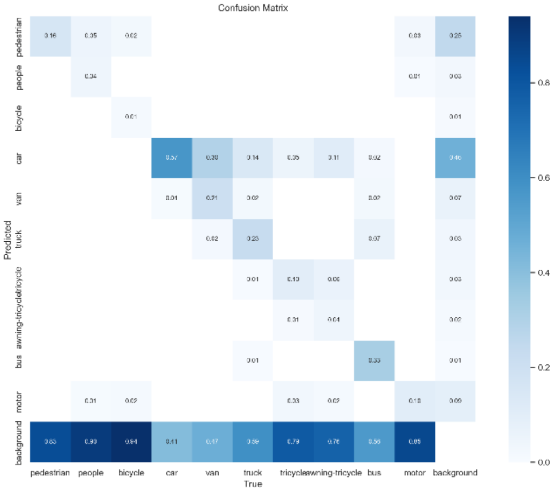

Figure 1: Confusion Matrix of model performance on Training dataset

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.

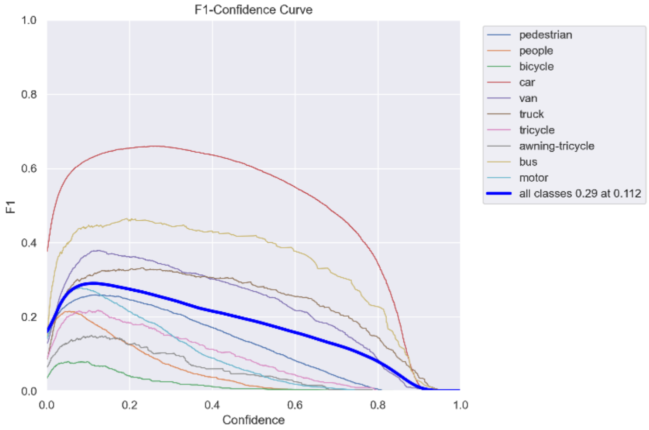

Figure 2: F1-Confidence Curve for YOLOv8 model on Training dataset.

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.

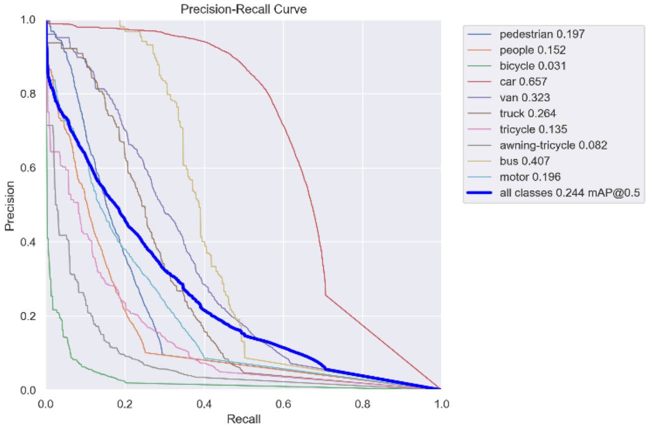

Figure 3: Precision and Recall curve on Training dataset

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.

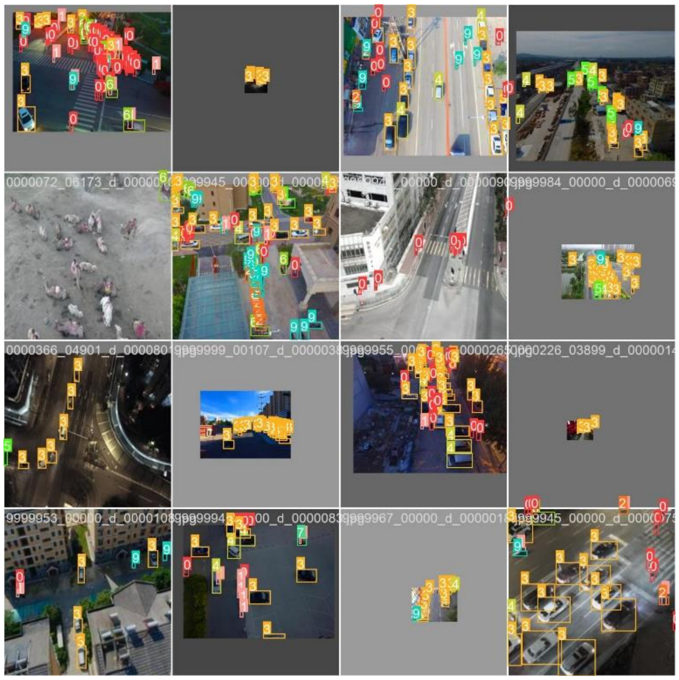

Figure 4: Sample results of prediction on training images

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.

V. 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.

Acknowledgment

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.

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