Temasek Defence Systems Institute

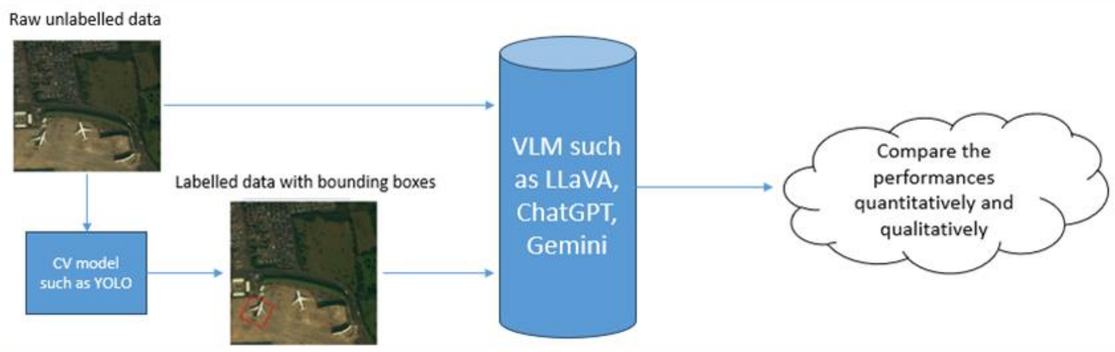
Autonomous Systems Applications of Large Language Models

Author: Chua Jia Yun
Thesis advisors: Dr Miguel Arana-Catania and Prof Argyrios Zolotas

INTRODUCTION

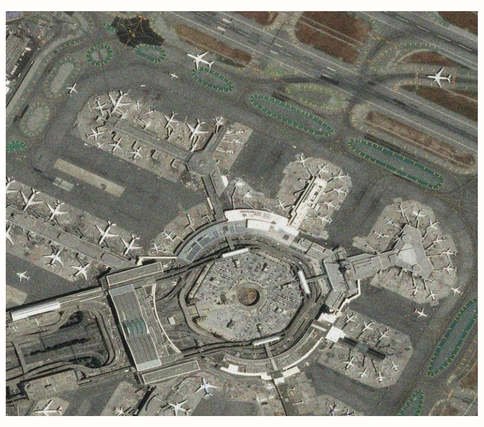
This thesis explores combining YOLO with VLMs, like Chat-GPT, LLaVA and Gemini, to improve accuracy and contextual understanding in remote sensing image analysis, addressing challenges in degraded image conditions and enhancing overall data interpretability.

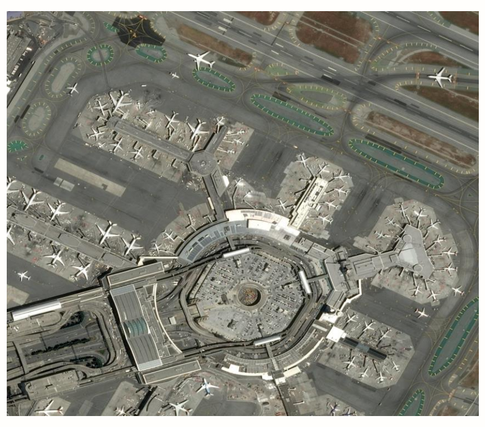
METHODOLOGY AND EXPERIMENT

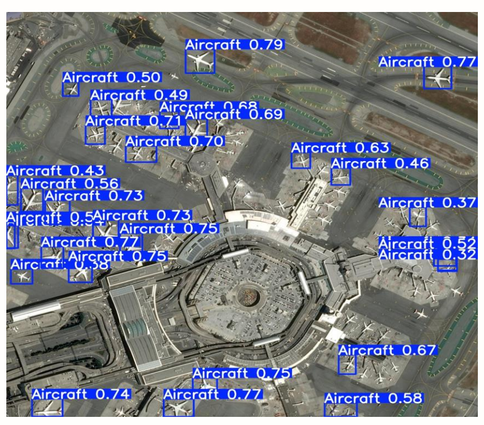


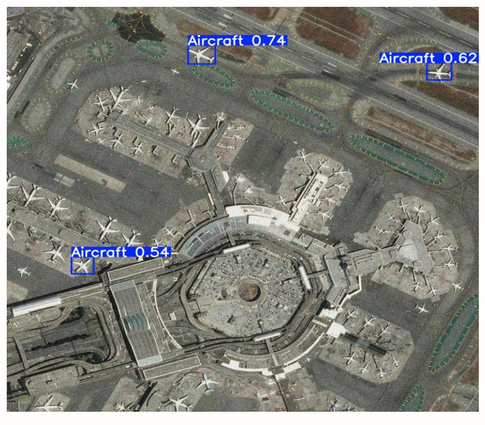
This methodology integrates YOLOv8's object detection with Vision-Language Models (VLMs) like Chat-GPT, LLaVA, and Gemini to enhance both identification and interpretation of objects in complex visual datasets.

Aircraft-Related Dataset



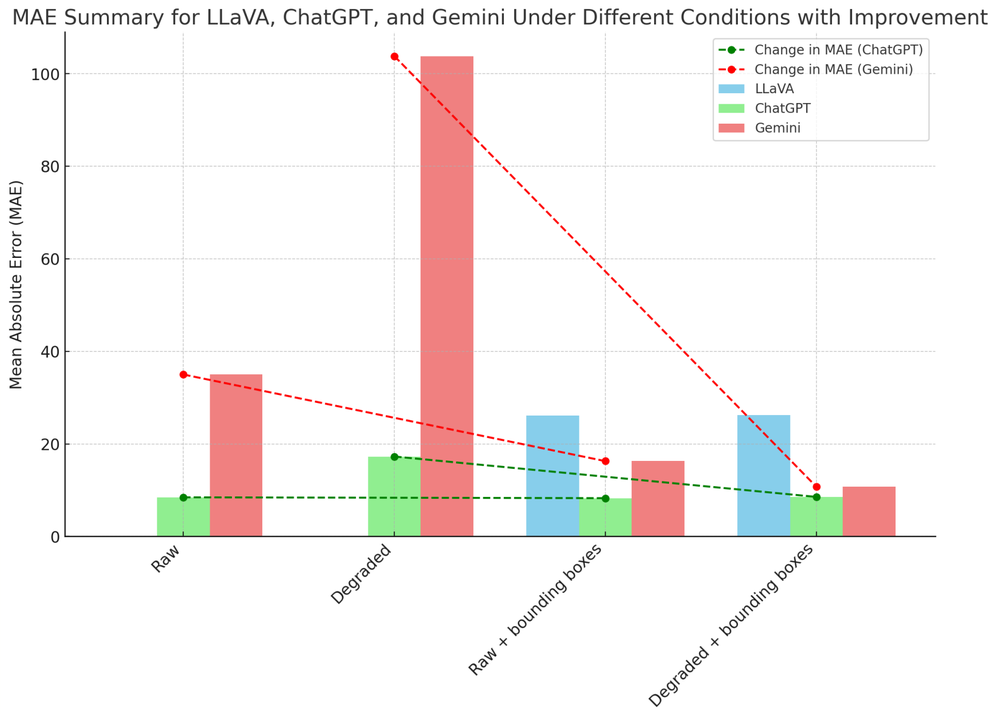


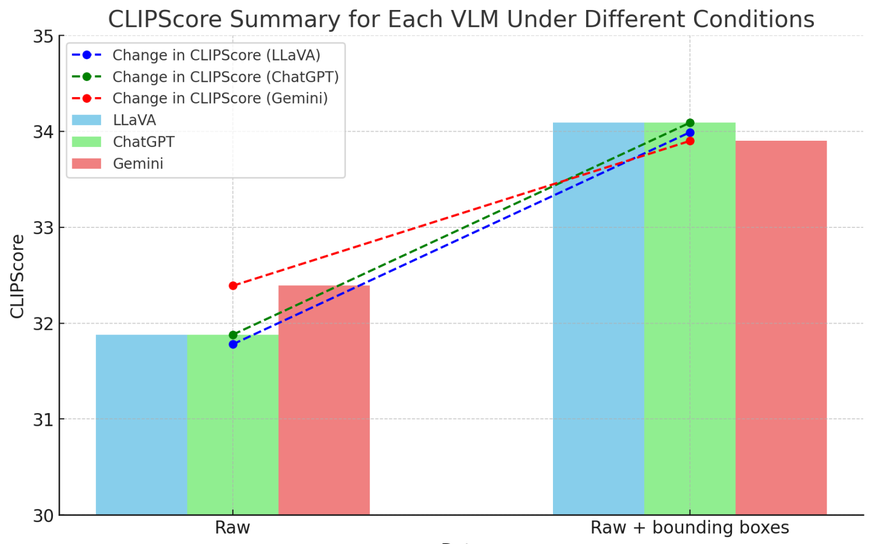




Disaster-Relief Dataset

RESULTS AND DISCUSSION





- **Object Counting**: Bounding boxes significantly reduced MAE, particularly for Gemini and LLaVA, enhancing object counting accuracy across models.
- **Improved CLIPScore**: Bounding boxes improved CLIPScores, leading to more accurate and contextually aligned descriptions from all models.
- **Improved Contextual Understanding**: Bounding boxes enhanced models' ability to provide detailed and contextually rich descriptions of complex scenes. With additional visual cues, it could provide more information such as the number and location of aircraft, identified additional elements within the dataset, and correctly deduced unobstructed routes in disaster scenarios.

CONCLUSION

In conclusion, integrating YOLOv8 with Vision-Language Models show potential in enhancing both detection accuracy and contextual understanding in remote sensing.

- Qualitative: An average MAE improvement of **48.46%** across models in both raw and degraded scenarios
- Qualitative: **6.17%** average improvement in CLIPScore

Date: Oct 2024

NUS National University of Singapore