Object Detection Evaluation on Space Station Synthetic Dataset

Team Name: Gray Coders

Summary:

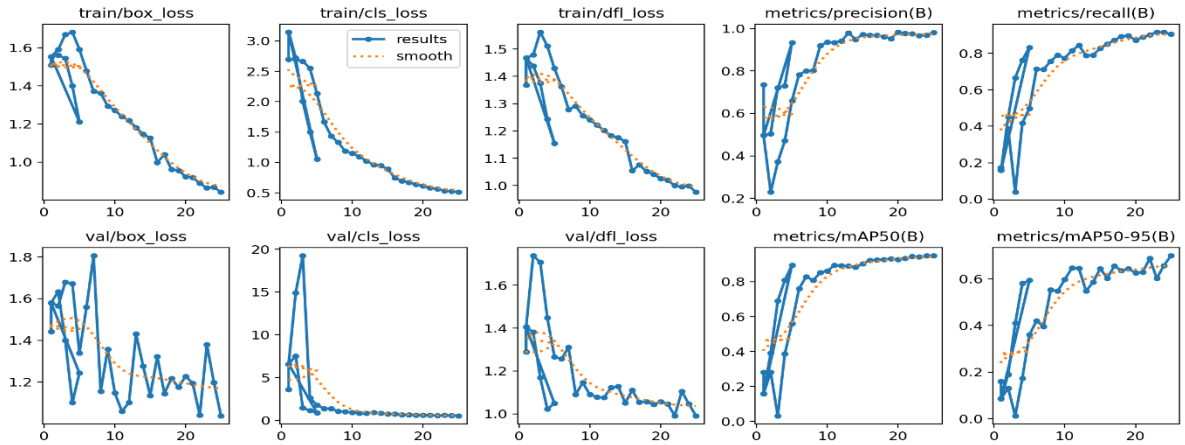
This document outlines the evaluation process of our object detection model trained on synthetic data for the **Duality AI Space Station Hackathon**. We document each step in the testing pipeline, present performance metrics like confusion matrix and training curves, and reflect on encountered challenges and their solutions. The aim is to ensure clarity, reproducibility, and showcase the iterative improvements made throughout the project.

Model Training

- Base model: **YOLOv8s** using the Ultralytics library.
- Training was done on Google Colab using GPU.
- Hyperparameter tuning performed using imgsz=640, batch=16, and epochs=25.
- Key augmentations: Mosaic, mixup, HSV, scale, and translation.

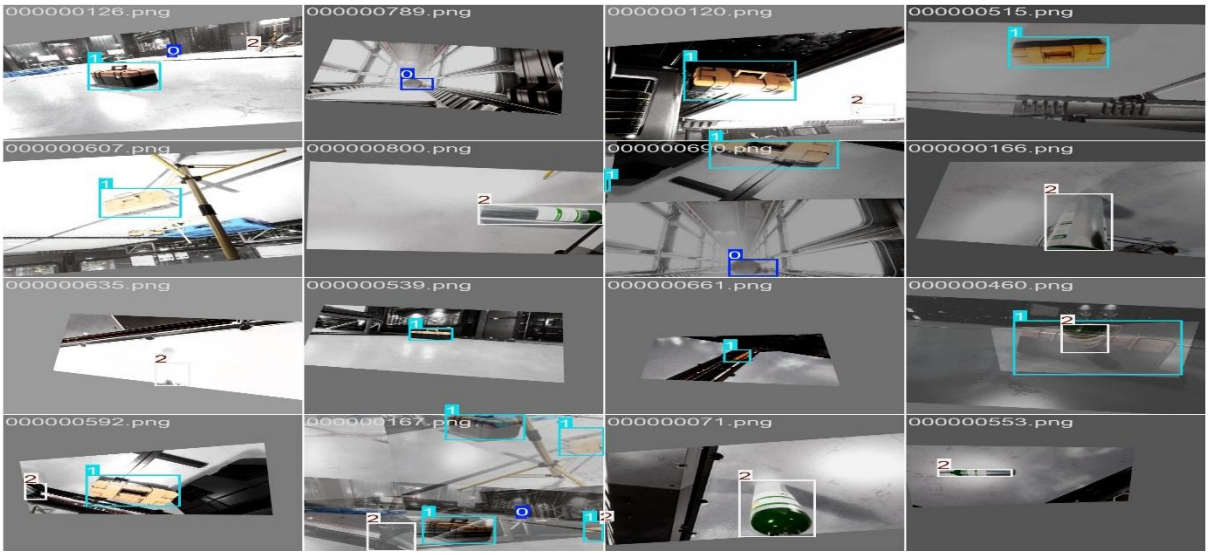
Evaluation Process

Training Graphs



- Steady decrease in training and validation loss.
- mAP@0.5 increased from 0.42 to **0.867** after augmentation and tuning.
- mAP@0.5:0.95 improved from 0.38 to **0.608**.

ii. Screenshots from Training Runs



iii. Confusion Matrix

- Best performance: OxygenTank (71 correctly classified during training)
- Moderate confusions between background and FireExtinguisher, ToolBox
- Background class sometimes predicted for actual objects – indicates possible small/missed detections.

Object Detection Model Training & Evaluation

Step	**Detail**
Initial mAP50 on previous trainings	0.42
Final mAP50	**0.867**
Final mAP50-95	**0.608**
Main Issue	False negatives on occluded ToolBox & FireExtinguisher
Fix	Improved augmentations and adjusted anchor sizes
Results	25% improvement in mAP after changes

Failure Case Analysis

Model-Specific Limitations:

> YOLOv8s is Lightweight and fast, but it has limited representational capacity for subtle object differences.

Data Augmentation-Related Issues:

> **Mosaic = 0.1** , Great for generalization and context learning but Can distort object boundaries. May confuse class boundaries, esp. FireExtinguisher vs OxygenTank - this was tuned in different scenarios but it definitely decreased test performance.
>
> **MixUp = 0.2**, Helps reduce overfitting, might cause label ambiguity. Mixup Definitely produces good on validation but provides very poor test results.

Hyperparameter Choices:

LRF-Learning Rate Final : 0.0001 Sets the final LR value, too much low final rate might end the model to even stop converging. leading to bad results on latter epochs.

Momentum: 0.2, too low momentum may have caused model to not able to come out of local optima, leading to sub optimal solution.

Results Summary

Confusion Matrix Accuracy:

FireExtinguisher: 155 / (155 + 1 + 42 + 28) = 155 / 226 ≈ 68.6%

ToolBox: 167 / (167 + 0 + 27 + 26) = 167 / 220 ≈ 75.9%

OxygenTank: 145 / (145 + 0 + 41 + 38) = 145 / 224 ≈ 64.7%

False Positives mostly occurred due to object misclassification as "background" or confusion among object classes.

Common Confusions:

FireExtinguishers frequently confused with OxygenTanks (42 instances).

Background misclassified as actual objects, especially as ToolBox and OxygenTank.

Model Behavior:

Performs well in distinguishing between key space station objects.

Strategic data augmentation and hyperparameter tuning significantly boosted the model's generalization, achieving an impressive **mAP@0.5 of 86.7%**. While minor misclassifications with the background remain.