YOLOv8n – Custom Dataset Performance

Ayaanulla Khan (YOLOv8n)

1. Scope and Goal

This report compiles training results for YOLOv8n on the project's custom dataset. The objective is to run multiple trainings, compare metrics, and identify the best-performing setup. Roboflow dataset was not accessible at training time, therefore, this document focuses on custom-only training and provides clear, reproducible commands.

2. Environment and Training Config

Two YOLOv8n training runs were available:

• Run A: train (100 epochs)

epochs: 100
batch: 32
imgsz: 640

• Run B: train2 (250 epochs)

epochs: 250
batch: 16
imgsz: 768

lr0: 0.003
patience: 50

workers: 2
mosaic: 0.8

close_mosaic: 10

mixup: 0.15

hsv_h: 0.015

hsv_s: 0.7

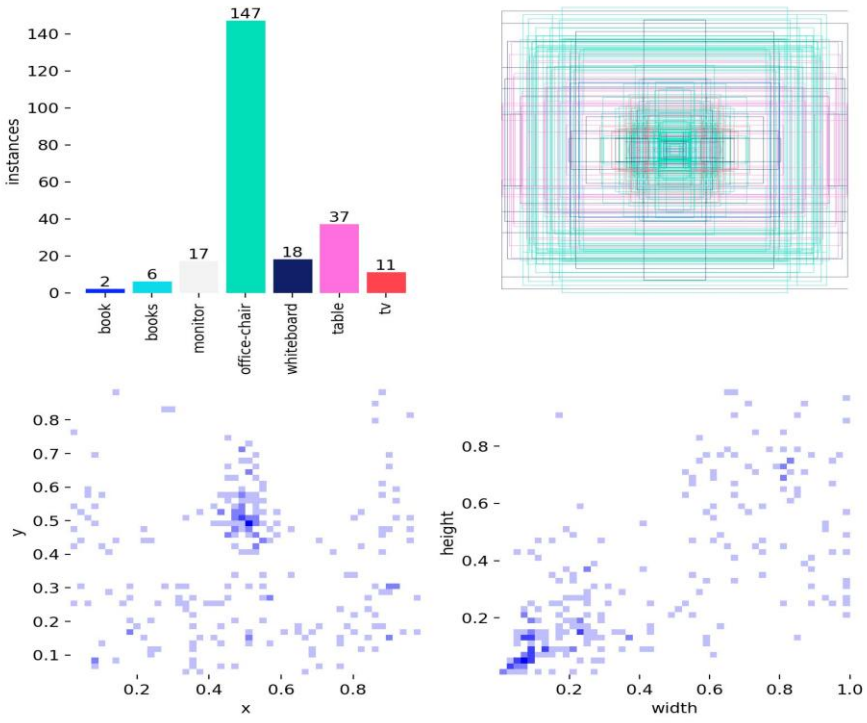
hsv_v: 0.4

scale: 0.5

shear: 2.0

perspective: 0.01

3. Understanding Custom Dataset



- Based on the above insights we can observe heavy class imbalance as office chair – 147 dominates.
- Center bias, many box centers cluster around the image middle, fewer near borders.

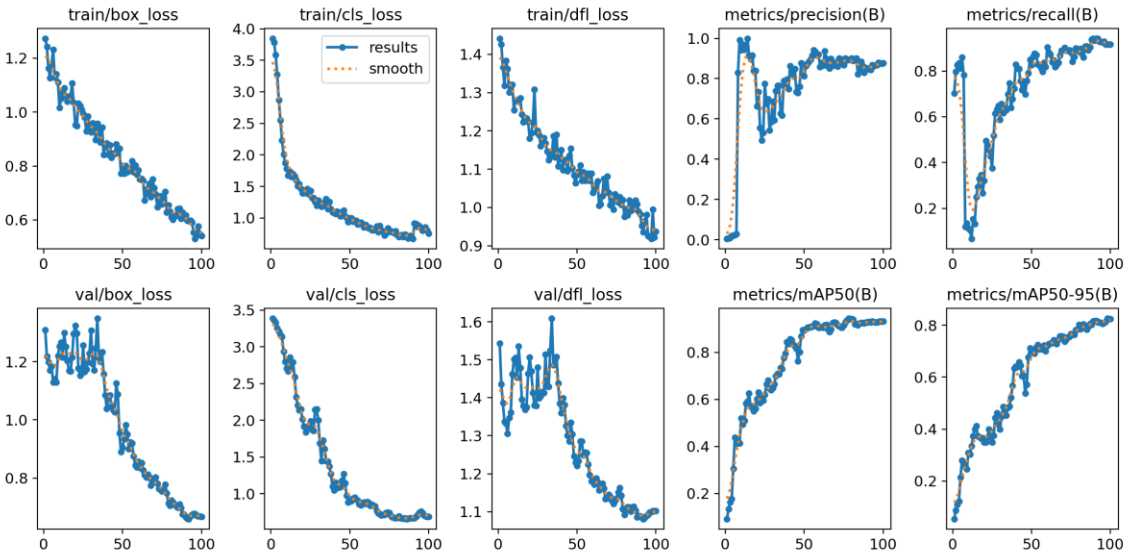
- Imbalance and center bias favor conservative detectors. That explains why Run-B shows higher precision but lower recall and lower mAP50-95, its missing minority cases.
- Run-A generalizes better under these data quirks suggested by its higher mAP50-95 and recall.

4. Results Summary

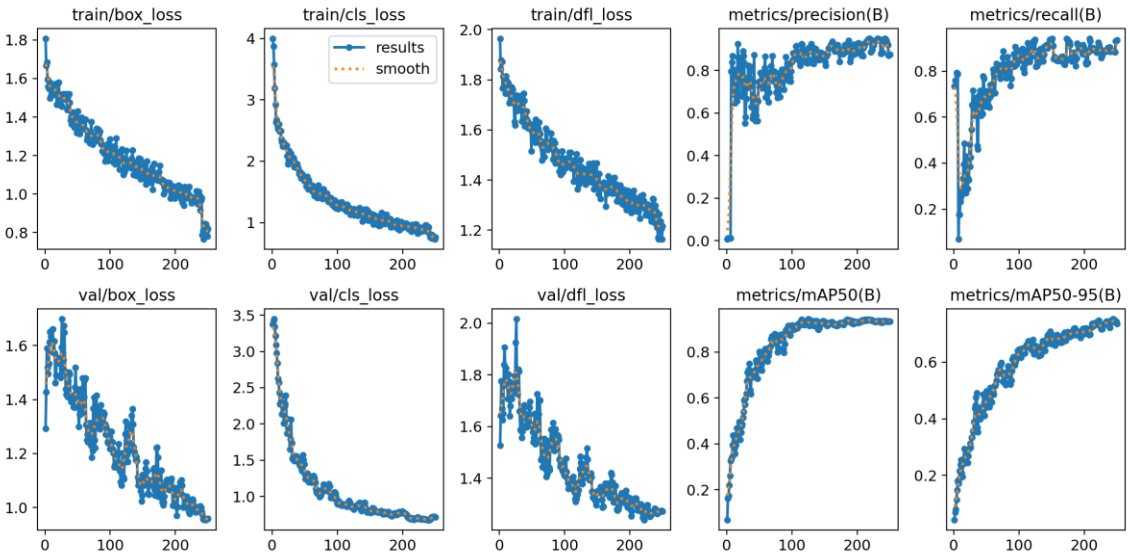
Run	Epochs	Best @	Precision	Recall	mAP50	mAP50-95	tr/cls_loss	val/cls_loss	epochs(set)
Run A - train (100)	100	98.0	0.876	0.915	0.931	0.824	0.752	0.689	100
Run A - train (100)	98	98.0	0.877	0.917	0.932	0.826	0.854	0.697	100
Run B - train2(250)	250	244.0	0.875	0.935	0.933	0.735	0.762	0.712	250
Run B - train2(250)	244	244.0	0.943	0.887	0.934	0.752	0.748	0.713	250

5. Curve Comparison

RUN-A



RUN-B



- Run-A reached its plateau around epoch 90-100 and achieved higher mAP50-95, while Run B improves slowly up to 240 epochs but stays lower overall.
- Validation classification loss stabilizes a bit lower in Run A which aligns with its better generalization.
- mAP50 is similar for both, the gap appears at mAP50-95, meaning Run-A localizes better at stricter IoUs.

- Run B's precision is higher but with lower recall and lower mAP50-95, its more conservative detector that misses more obstacles, not what we want for navigation safety.

6. Analysis

Run A (imgsz 640, batch 32, 100 epochs) achieved a best mAP50-95 of 0.8264 at epoch 98.

Run B (imgsz 768, batch 16, 250 epochs) achieved a best mAP50-95 of 0.7524 at epoch 244.

Trend (Run A): early mean mAP50-95 ~ 0.1944 → mid 0.6741 → late 0.8154; best at epoch 98.

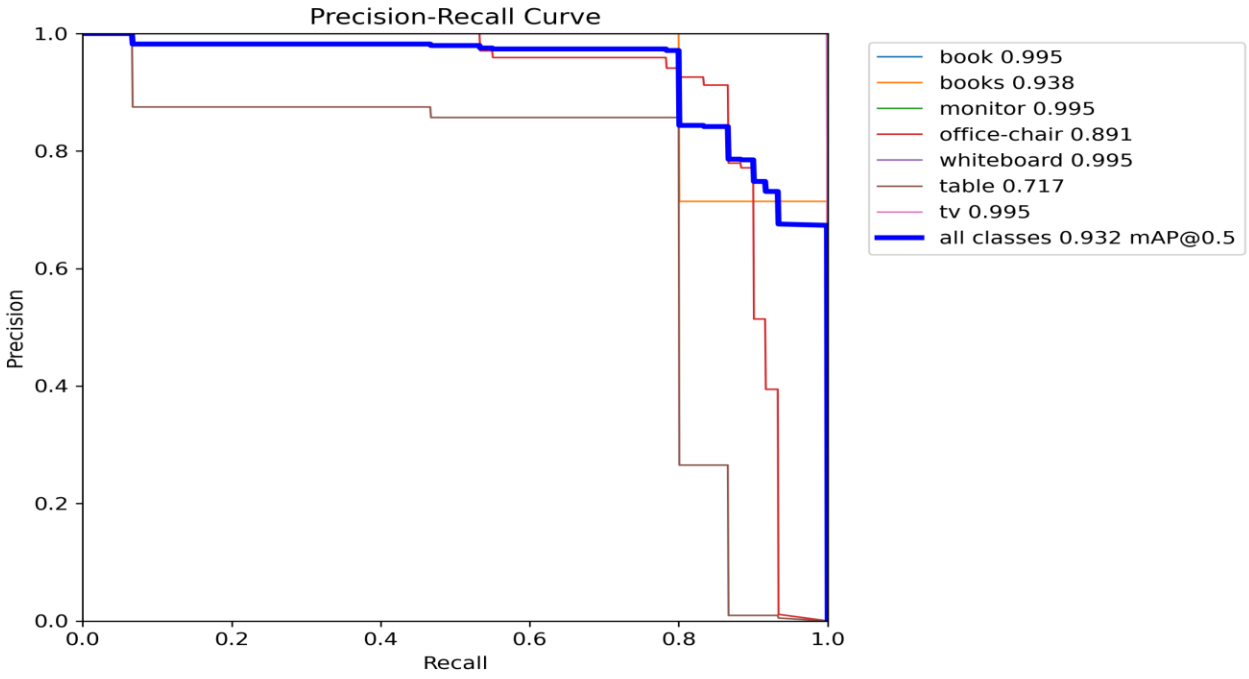
Trend (Run B): early mean mAP50-95 ~ 0.2159 → mid 0.6449 → late 0.7453; best at epoch 244.

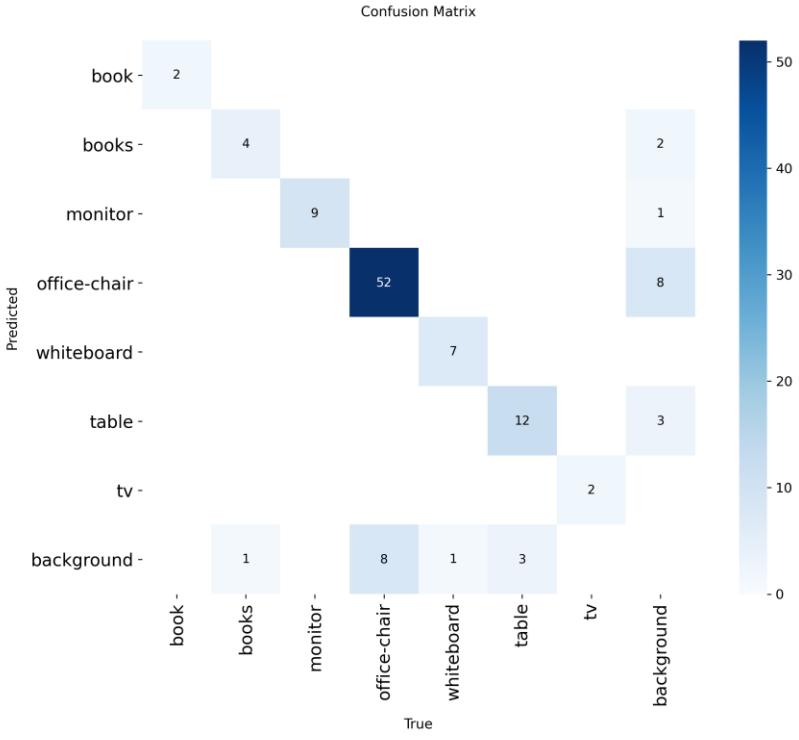
The 100-epoch configuration (smaller image size, larger batch) produced higher mAP50-95 with good recall and also with a comparable mAP50 in comparison to the 250-epoch run. Longer training with smaller batch and larger image size did not improve generalization, suggesting either overfitting or suboptimal optimization for this dataset and compute budget.

7. Key Plots

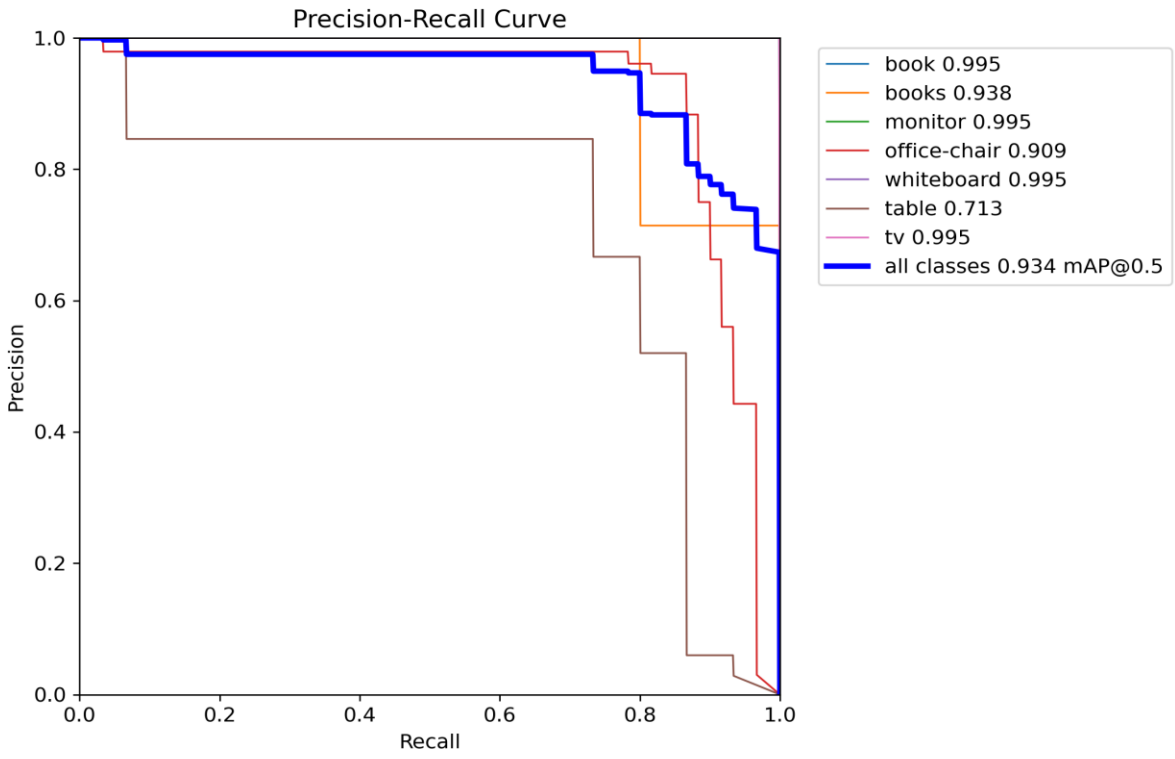
Precision–Recall and Confusion Matrix images are as follows.

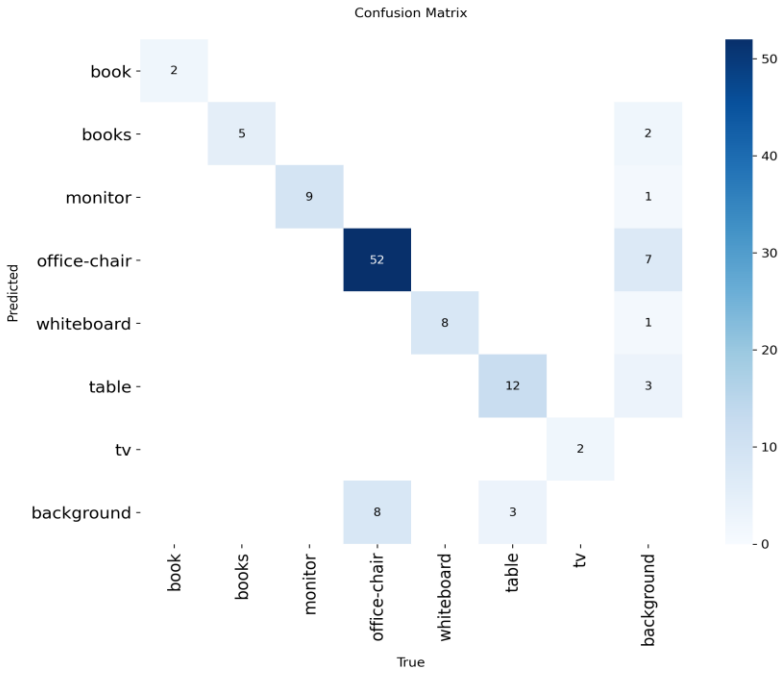
RUN-A plots:


Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
book 0.995
books 0.938
monitor 0.995
office-chair 0.891
whiteboard 0.995
table 0.717
tv 0.995
all classes 0.932 mAP@0.5


Confusion Matrix
Predicted
True
book
books
monitor
office-chair
whiteboard
table
tv
background
2
4
2
9
1
52
8
7
12
3
2
1
8
1
3
50
40
30
20
10
0

RUN-B plots:


Precision-Recall Curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0
book 0.995
books 0.938
monitor 0.995
office-chair 0.909
whiteboard 0.995
table 0.713
tv 0.995
all classes 0.934 mAP@0.5


Confusion Matrix
Predicted
True
book
books
monitor
office-chair
whiteboard
table
tv
background
2
5
2
9
1
52
7
8
1
12
3
2
8
3
50
40
30
20
10
0

8. Reproducibility

Commands used:

Run A:

yolo detect train model=yolov8n.pt, data=data.yaml, imgsz=640 epochs=100, batch=32, name=train.

Run B:

yolo detect train model=yolov8n.pt, data=data.yaml, imgsz=768 epochs=250, batch=16, name=train2, patience=50.

9. Conclusion

Run A delivers the best balance of accuracy and safety for our use case. Its higher mAP50-95 and recall mean it finds more obstacles and localizes them better at stricter IoUs, which matters more than a marginal precision gain. It also converges faster and with fewer resources, making it the sensible baseline for iteration.

Overall, for the custom dataset, YOLOv8n delivered the best performance using the configuration: imgsz=640, batch=32, and around 100 epochs, achieving an mAP50-95 of approximately 0.826. Until I gain access to the Roboflow dataset, the recommended approach is to move forward with this setup referred to as *Run A*. If time allows, I can re-run this configuration on the combined dataset to support a final decision.

Appendix: mAP50-95 Curves