HALO:
High-efficiency Autonomous Low-SWaP Operations

Sloan Hatter, Blake Gisclair
Advisor: Dr. Ryan T. White

MILESTONE 4 PROGRESS MATRIX

Task	Completion %	Sloan	To Do
Implement FP16 engine from FP32 ONNX model	**100%**	Compile the FP32 ONNX model into a TensorRT engine using FP16	**None**
Implement INT8 engine	**100%**	Build an INT8 engine using the TensorRT framework from the Jetson SDK	**None**
Write an Inference Script for the INT8 Model	**100%**	Write a Python inference script for the INT8 engine	**None**
Record Metrics for 8-bit Model	**100%**	Get inference metrics from the INT8	**None**

TASKS COMPLETED

- Implement FP16 engine from FP32 ONNX model
  - Used TensorRT to compile FP32 ONNX into a device-specific engine for FP16 engine using trtexec

- Implement an INT8 Pipeline on Jetson Orin
  - Utilized TensorRT's Polygraphy tools to run calibration on 256 representative images

- Write an Inference Script for the INT8 Model
  - Produces detections and draws bounding boxes

- Compute/Record Metrics for 8-bit Model
  - Get mAP (mean Average Precision) scores for each model

MODEL COMPARISON

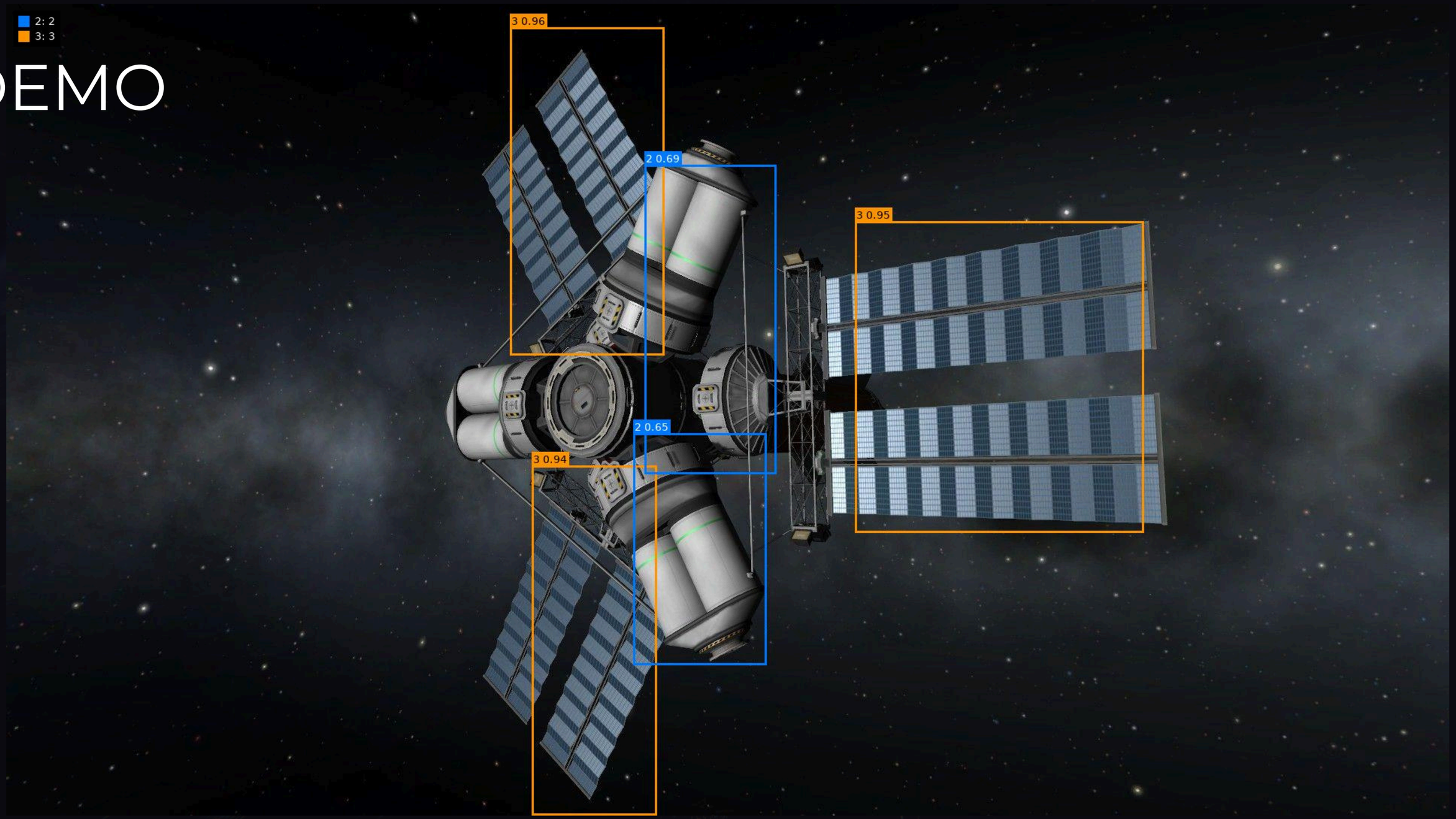
Precision	mAP@[.50:.95]	AP@0.50	AP@0.75	AR@100
FP32	0.0958	0.2589	0.0586	0.1826
FP16	0.0724	0.2274	0.0274	0.1577
INT8	0.0735	0.2285	0.0237	0.1575

FP32 → FP16

- mAP: ≈ –24%
- AP50: ≈ –12%
- AP75: ≈ –53%
- AR100: ≈ –13%

FP16 → INT8

- mAP: ≈ +1.5%
- AP50: ≈ +0.5%
- AP75: ≈ –0.0037
- AR100: ≈ –0.0002


2: 2
3: 3
DEMO
3 0.96
2 0.69
3 0.95
2 0.65
3 0.94

MILESTONE 4 TASKS

- Fix the 32-bit Model
  - Use TensorRT on FP32
- Implement 4-bit Model
  - Run INT4 inference engine using TensorRT
- Begin Progress on 2-bit Model
  - No specific support for INT2 inference engine, implement manually
- Formal Compilation of Results
  - Faculty advisor’s request, compile results thus far to begin publication

MILESTONE 5 TASK MATRIX

Task	Sloan
Fix the 32-bit Model	100%
Implement 4-bit Model	100%
Begin progress on 2-bit Model	100%
Formal Compilation of Results thus far for Publication	100%

THANK YOU!
QUESTIONS?