HALO: High Autonomous Low-SWaP Operations

Sloan Hatter, Blake Gisclair

Faculty Advisor: Dr. Ryan White, Dept. of Mathematics and Systems Engineering, Florida Institute of Technology

MOTIVATION

Orbital object detection is a vital aspect of space operations, particularly for identifying satellite components. Convolutional Neural Networks (CNNs) are typically used for such operations by running on-board models directly on satellite systems. However, a new neural network architecture, known as Vision Transformers (ViTs), have shown greater effectiveness due to their ability to capture global context. One main issue of deploying systems with such capabilities is resource allocation. One solution is to run models on a Low-SWaP system; however, this results in inefficient performance. To enable efficient ViT operations on Low-SWaP systems, the model must be scaled down through quantization, enabling the achievement of High Autonomous Low-SWaP Operations.

GOALS

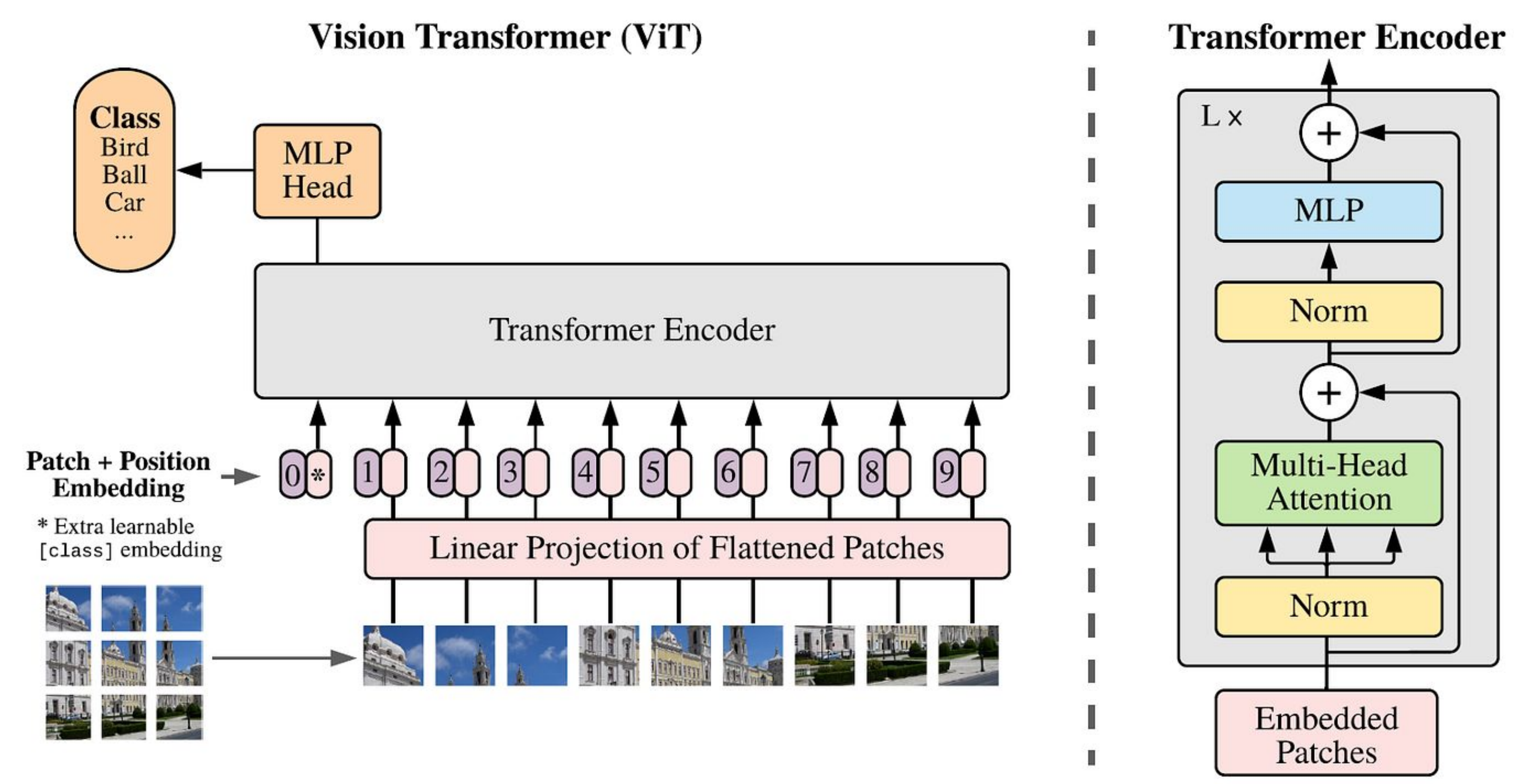
- Train types of CNN and ViT models
- Quantize a ViT to FP16 representation
- Quantize a ViT to INT8 representation
- Evaluate & Compare CNN and ViT performance

METHODS/RESOURCES

- NVIDIA TensorRT SDK
- ONNX Quantization
- Quantize/Dequantize (Q/DQ) Node Insertion
- YOLOv5 Repository
- NVIDIA Jetson AGX Orin
- Dr. White's NEural TransmissionS (NETS) Lab GPU servers

REFERENCES

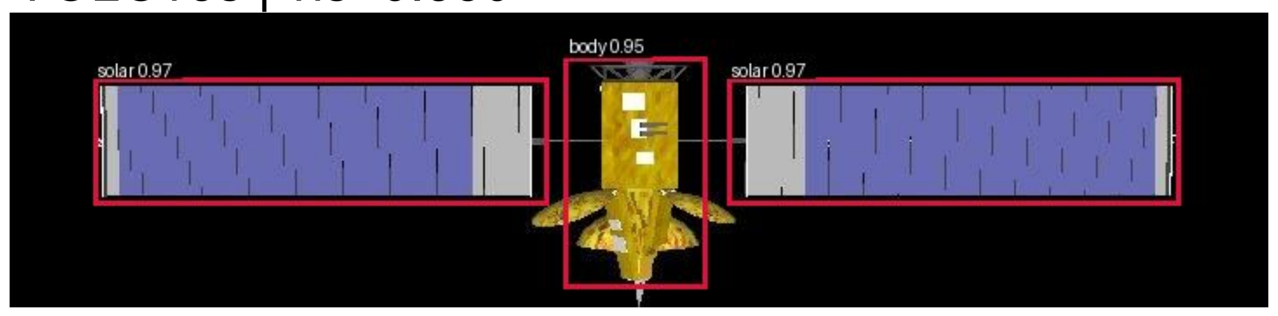
- Dosovitskiy, Alexey, et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv.Org*, 3 June 2021.
- Issitt, Arianna, et al. "Reliable onboard 3D reconstruction of unknown spacecraft via data acquisition guided by Orbital Geometry and lighting." AIAA SCITECH 2026 Forum, 8 Jan. 2026.
- Mahendrakar, Trupti, et al. "Unknown non-cooperative spacecraft characterization with lightweight convolutional neural networks." Journal of Aerospace Information Systems, vol. 21, no. 5, May 2024, pp. 455–460.
- Mukherjee, Shaoni. "Quantizing Vision Transformers for Efficient Deployment: Strategies and Best Practices." *Paperspace by DigitalOcean Blog*, Paperspace by DigitalOcean Blog, 15 Dec. 2023.
- NVIDIA. "Working with Quantized Types." *Working with Quantized Types - NVIDIA TensorRT*, 7 Apr. 2026.
- Ultralytics. "Comprehensive Guide to Ultralytics YOLOv5." *Ultralytics YOLO Docs*, 7 Apr. 2026.



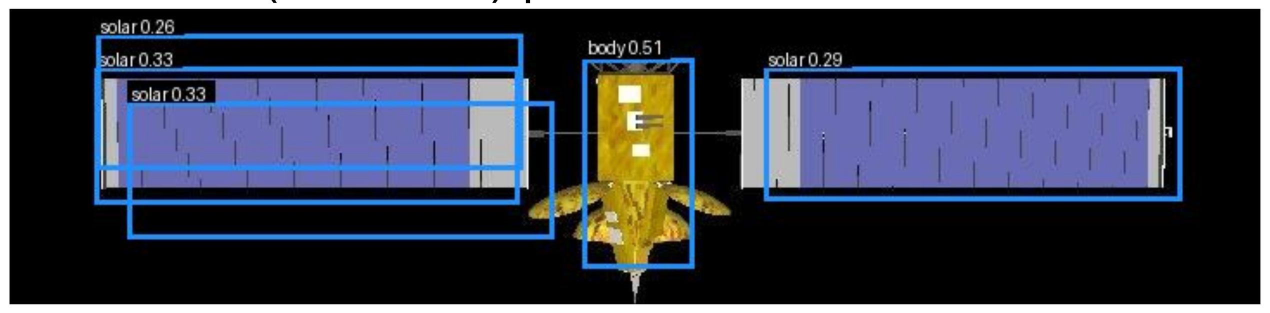
Model	mAP@ 0.5	mAP@ [.50:.95]	GFLOPs	Model Size (MB)	Runtime Memory (MB)	FPS
YOLOv5s	0.853	0.667	15.8	26.82	230.38	16.066
YOLOv5s (Reduced)	0.701	0.491	10.4	14.24	184.93	12.9063
YOLOv5n	0.808	0.57	4.1	6.75	111.31	10.8393
FP32 ViT	0.848	0.53	18.6	24.69	288.74	6.3679
FP16 ViT	0.848	0.577	18.6	12.34	144.37	1.249

RESULTS

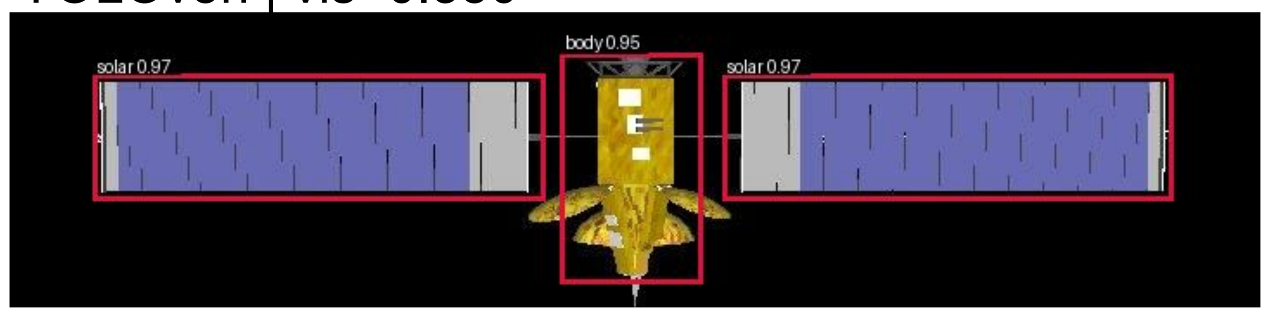
YOLOv5s | vis=0.350



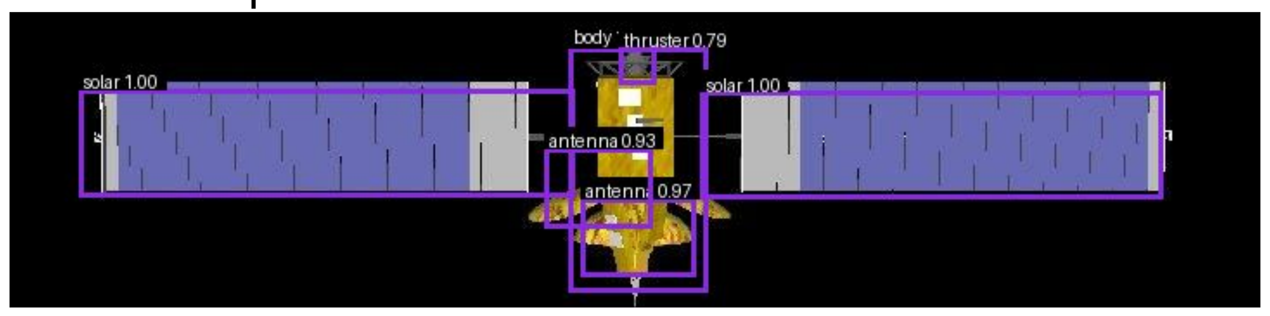
YOLOv5s (Reduced) | vis=0.350



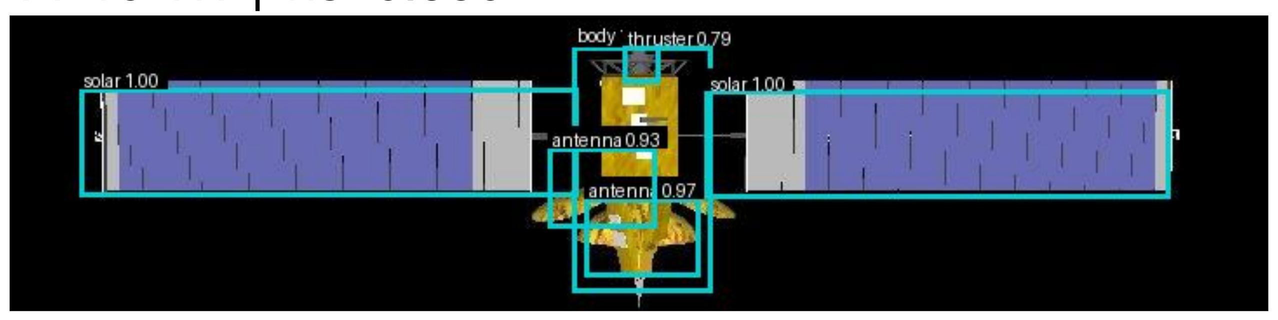
YOLOv5n | vis=0.350



FP32 ViT | vis=0.350



FP16 ViT | vis=0.350



FUTURE WORK

- Achieve true 8-, 4-, 2-, and 1-bit representations of ViTs for orbital object detection tasks
- Efficient satellite component object detection for:
  - Diagnosing broken components
  - Docking onto satellites
  - Identifying space debris
- Further analysis of CNN vs ViT comparisons

ACKNOWLEDGEMENTS

- Dr. White, for his guidance and foundational research insights.
- Arianna Issitt, for her help with the Jetson computer, datasets, and initial CNN model