Multimodal Perception

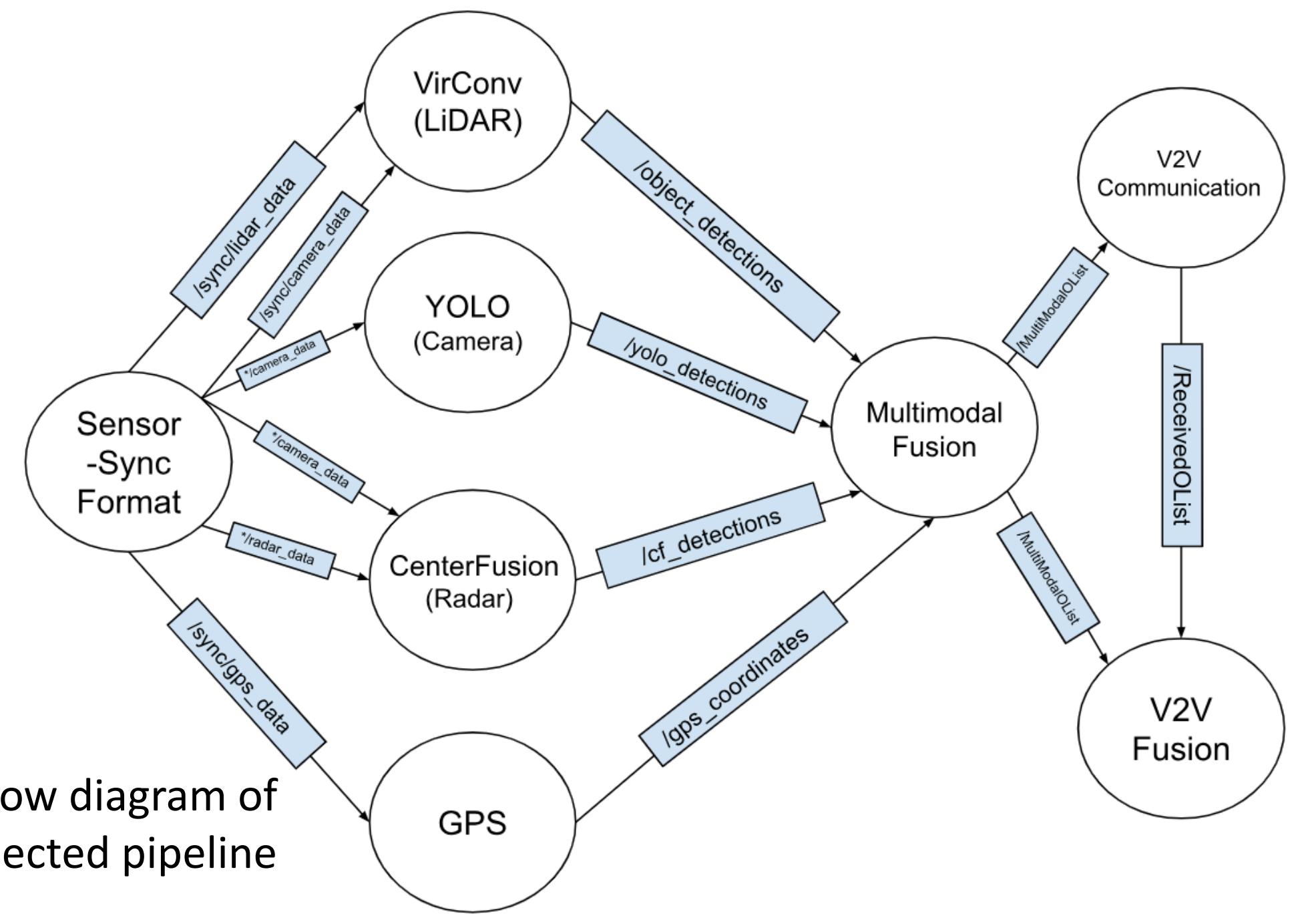
Fulmi Chang – ajchangwa@gmail.com, Sydney Dukelow - sydneydukelow@gmail.com

Abstract

Our project aims to enhance vehicular environment perception by implementing multimodal and vehicle-to-vehicle (V2V) fusion methods. We use state-of-the-art machine learning algorithms to apply object detection tailored to our datasets.
Multimodal fusion integrates data from multiple sensors:

- LiDAR
- Radar
- Camera
- GPS

V2V fusion integrates local multimodal data with another vehicle.



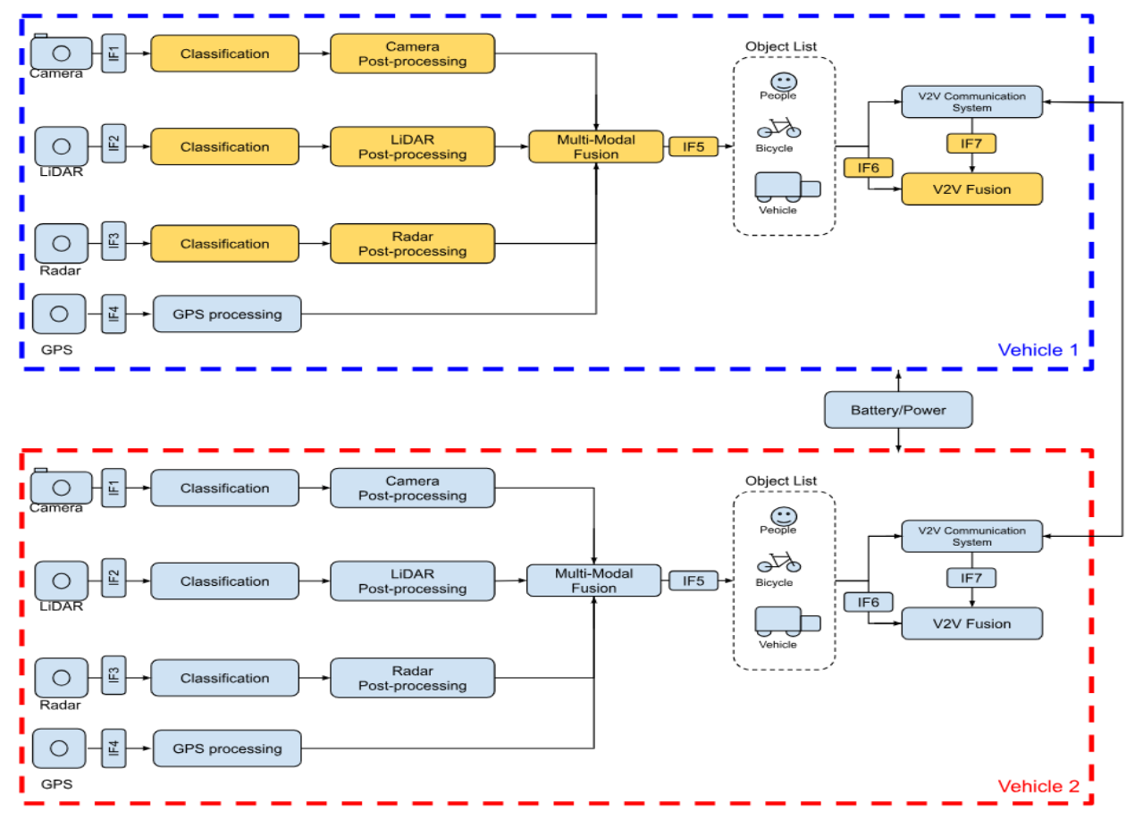
Complete flow diagram of our expected pipeline

Background

Connected Autonomous Vehicles (CAVs) rely on accurate environmental perception to ensure safe operation.
With a focus on:

- Accuracy
- Energy efficiency
- Timeliness

We developed a data processing pipeline that performs object detection and fusion with onboard sensors.



Hardware level diagram of our system

References

[1] H. Wu, C. Wen, S. Shi, X. Li and C. Wang, ”Virtual Sparse Convolution for Multimodal 3D Object Detection,” 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 21653-21662, doi: 10.1109/CVPR52729.2023.02074.
[2] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.
[3] R. Nabati and H. Qi, "CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 1526-1535, doi: 10.1109/WACV48630.2021.00157.
[4] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research. 2013;32(11):1231-1237. doi:10.1177/0278364913491297
[5] H. Caesar et al., "nuScenes: A Multimodal Dataset for Autonomous Driving," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 11618-11628, doi: 10.1109/CVPR42600.2020.01164.

Design Methods and Tools

We utilized a Linux system with the ROS Noetic framework to contain our data processing system. Within the ROS framework, we created nodes for each sensor object detection and fusion. We trained each model to detect cars, cyclists, and pedestrians. We would then send raw data through our data processing pipeline to each model and have it output its respective object lists.
The algorithms used for each node in our system depends on their respective data type. For LiDAR, we used “Virtual Sparse Convolution for Multimodal 3D Object Detection” (VirConv), camera used “You Only Look Once version 8” (YOLOv8), and radar used “Center-based Radar and Camera Fusion for 3D Object Detection” (CenterFusion).
We utilized the KITTI dataset for all modalities except radar, which used the nuScenes dataset.

Hardware Specifications

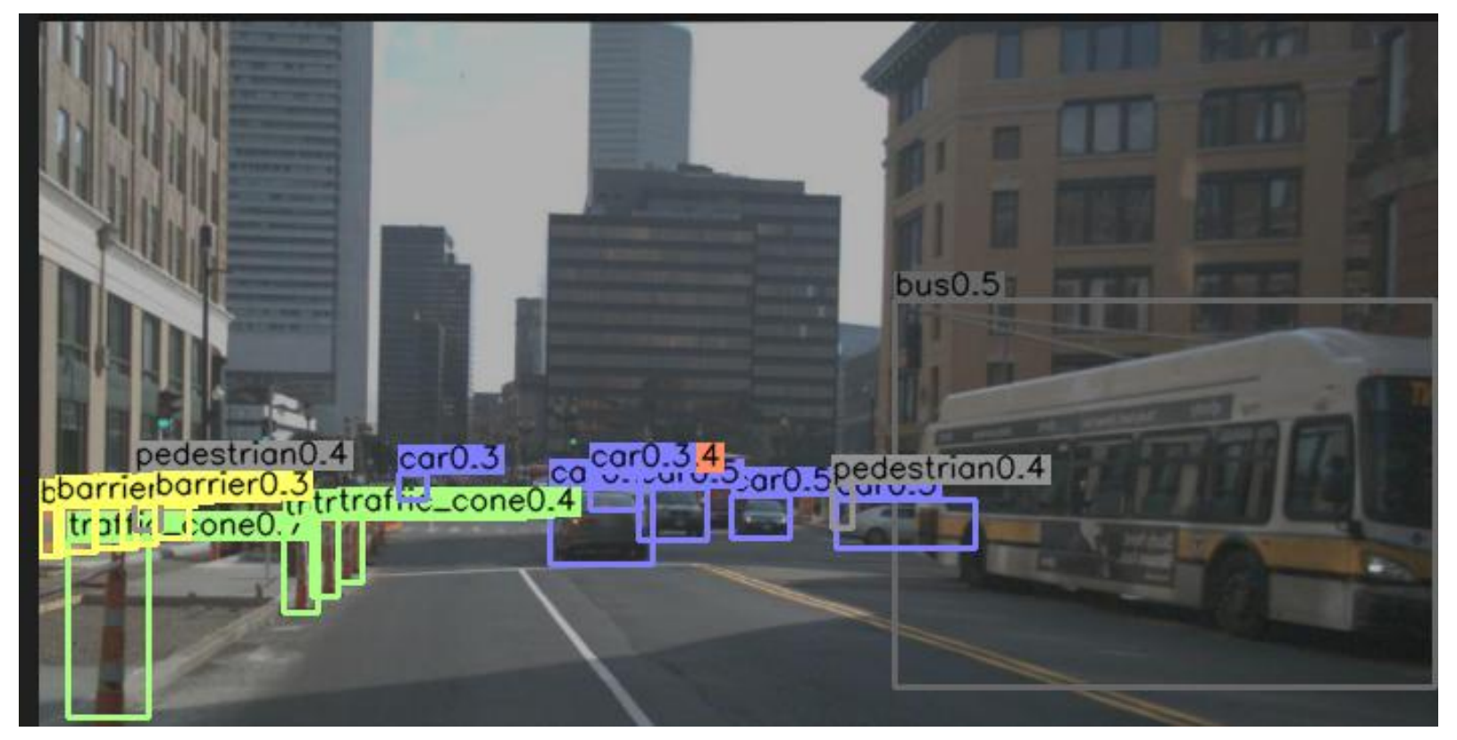
	Lenovo Legion Pro 7	Nvidia Jetson AGX Orin
Manufacturer:	Lenovo	Nvidia
Model:	16IRX8H/16IRX9H	945-13730-0050-000
System Frequency:	CPU: 5.4GHz GPU: 2040MHz	2.2GHz
Memory:	64GB	64GB
Storage:	4TB	64GB + 1TB
Environment:	Windows Subsystem for Linux 2 ROS Noetic	Ubuntu 20.04 LTS ROS Noetic

Lenovo Legion Pro 7

Nvidia Jetson AGX Orin

ROS Node Inputs

Node	Inputs
VirConv	• LiDAR • Camera
YOLO	• Camera
CenterFusion	• Radar • Camera
GPS	• GPS
Multimodal Fusion	• LiDAR • Camera • Radar • GPS



CenterFusion Validation Samples

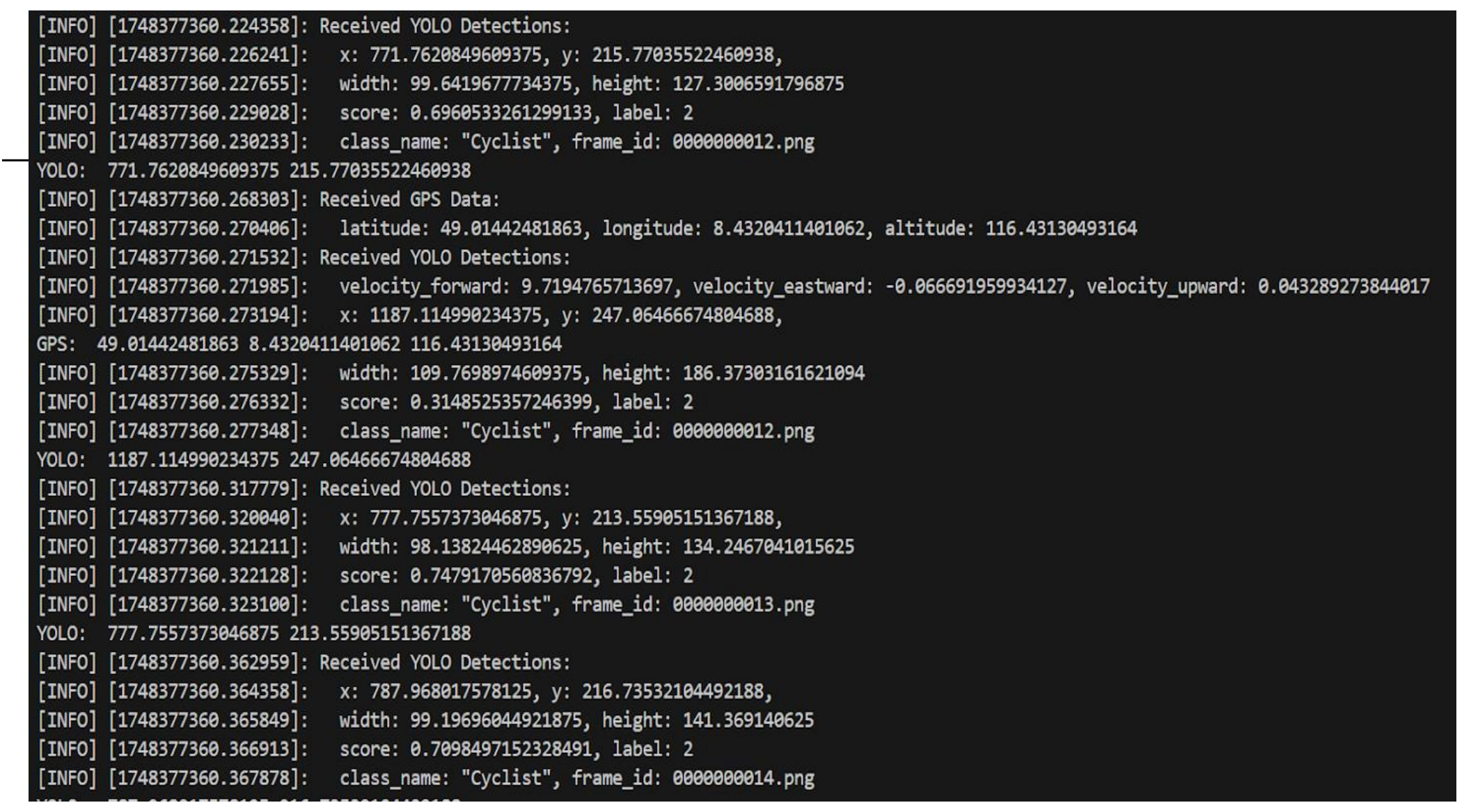
YOLOv8 Validation Samples

Results

- Successfully created the pipeline that does individual modality object detection
- These tasks are sending their respective object lists to the multimodal fusion node
- Integrated YOLO with Real Time Cooperative Perception Using C-V2X team
- Merged each modality’s object list to create a singular comprehensive object list

Task Metrics

Node Task	Execution Time per file (milliseconds)	Energy Efficiency (G-FLOPs)
VirConv	950ms	90.85 GFLOPs
YOLO	2ms	151.01 GFLOPs
CenterFusion	50ms	306.76 GFLOPs
GPS	50ms	0

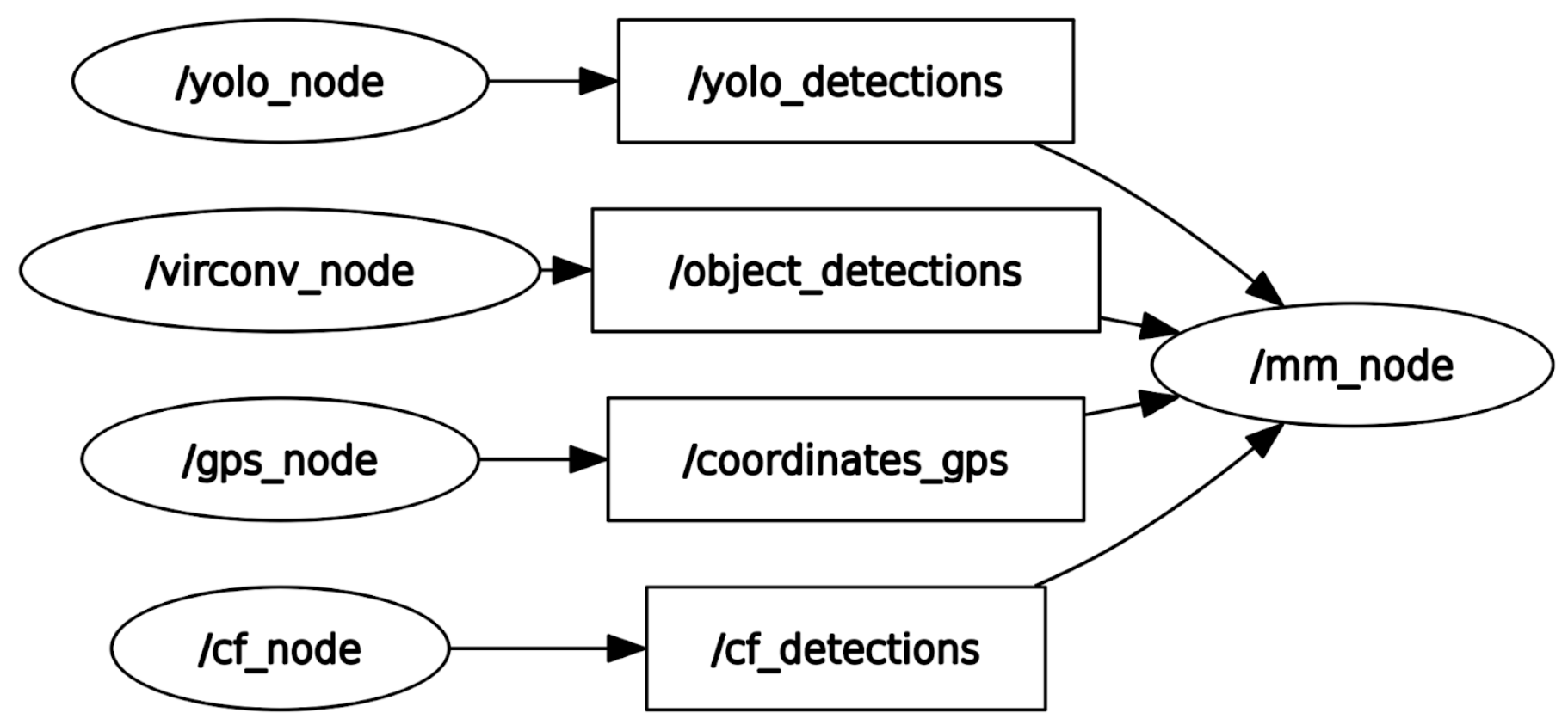


```
[INFO] [1748377360.224358]: Received YOLO Detections:
[INFO] [1748377360.226241]:   x: 771.7620849609375, y: 215.77035522460938,
[INFO] [1748377360.227655]:   width: 99.6419677734375, height: 127.3006591796875
[INFO] [1748377360.229028]:   score: 0.6960533261299133, label: 2
[INFO] [1748377360.230233]:   class_name: "Cyclist", frame_id: 0000000012.png
YOLO:  771.7620849609375 215.77035522460938
[INFO] [1748377360.268303]: Received GPS Data:
[INFO] [1748377360.270406]:   latitude: 49.01442481863, longitude: 8.4320411401062, altitude: 116.43130493164
[INFO] [1748377360.271532]: Received YOLO Detections:
[INFO] [1748377360.271985]:   velocity_forward: 9.7194765713697, velocity_eastward: -0.066691959934127, velocity_upward: 0.043289273844017
[INFO] [1748377360.273194]:   x: 1187.114990234375, y: 247.06466674804688,
GPS:  49.01442481863 8.4320411401062 116.43130493164
[INFO] [1748377360.275329]:   width: 109.7698974609375, height: 186.37303161621094
[INFO] [1748377360.276332]:   score: 0.3148525357246399, label: 2
[INFO] [1748377360.277348]:   class_name: "Cyclist", frame_id: 0000000012.png
YOLO:  1187.114990234375 247.06466674804688
[INFO] [1748377360.317779]: Received YOLO Detections:
[INFO] [1748377360.320040]:   x: 777.7557373046875, y: 213.55905151367188,
[INFO] [1748377360.321211]:   width: 98.13824462890625, height: 134.2467041015625
[INFO] [1748377360.322128]:   score: 0.7479170560836792, label: 2
[INFO] [1748377360.323100]:   class_name: "Cyclist", frame_id: 0000000013.png
YOLO:  777.7557373046875 213.55905151367188
[INFO] [1748377360.362959]: Received YOLO Detections:
[INFO] [1748377360.364358]:   x: 787.968017578125, y: 216.73532104492188,
[INFO] [1748377360.365849]:   width: 99.19696044921875, height: 141.369140625
[INFO] [1748377360.366913]:   score: 0.7098497152328491, label: 2
[INFO] [1748377360.367878]:   class_name: "Cyclist", frame_id: 0000000014.png
```

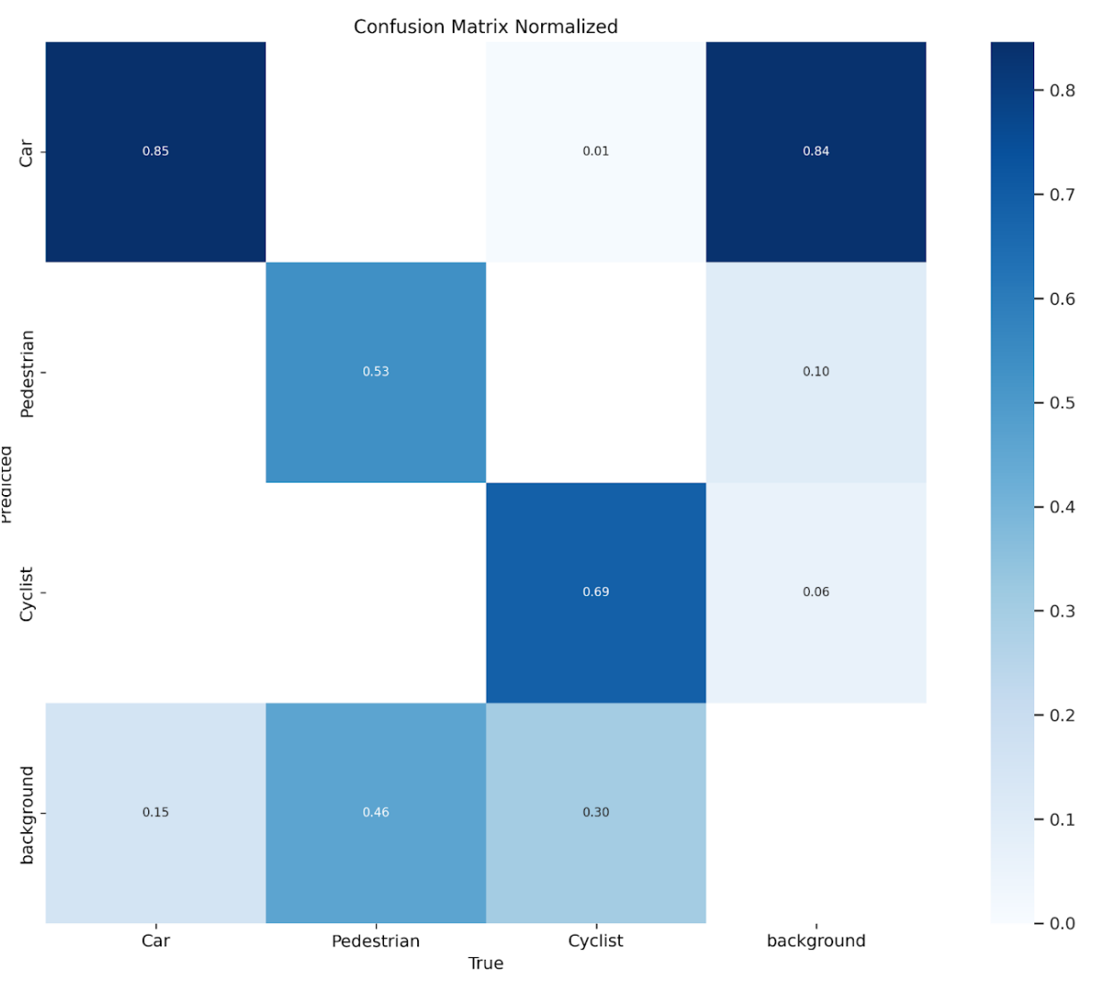
Unprocessed multimodal fusion object list outputs

Mean Average Precision Scores per Algorithm

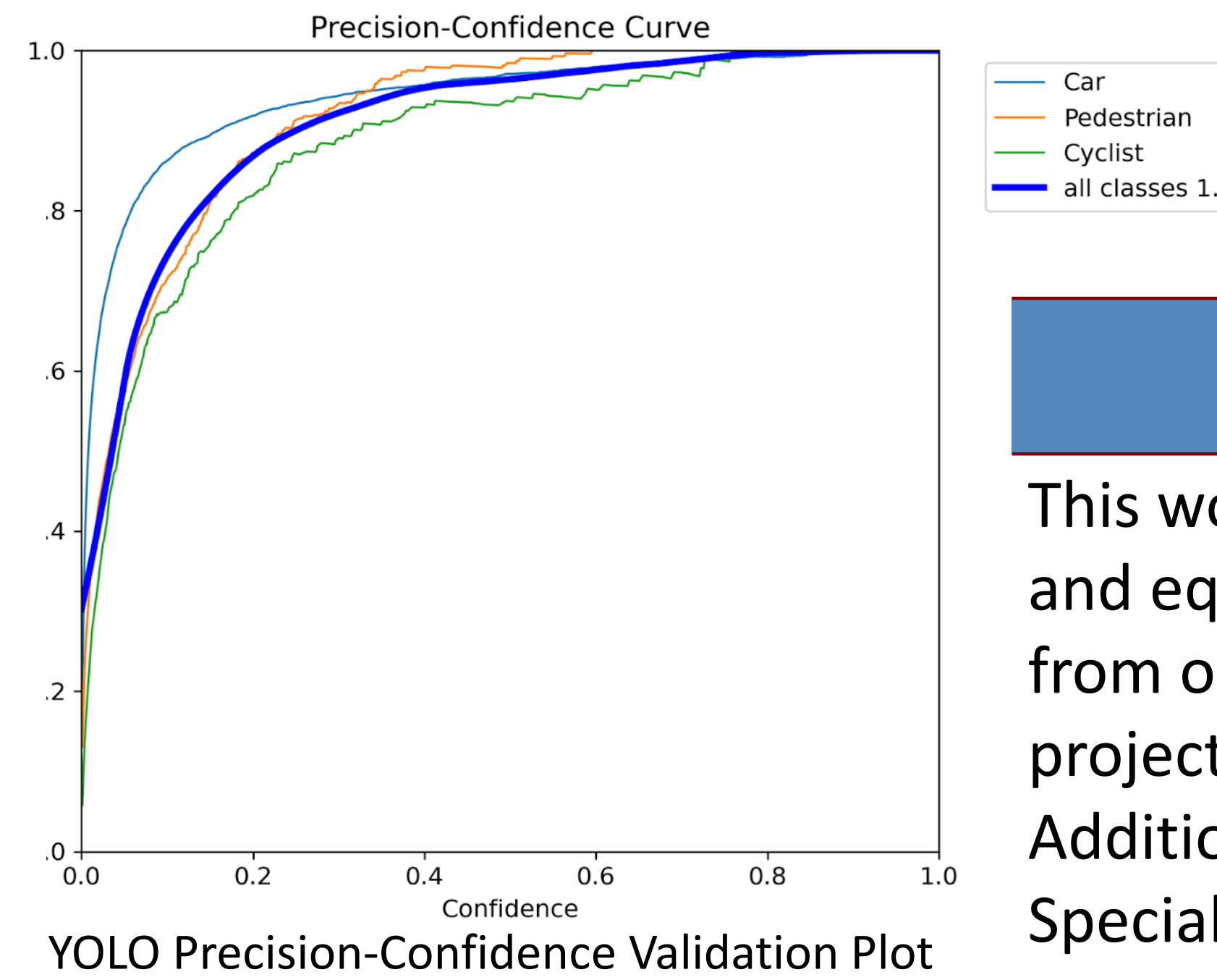
Algorithm	mAP Score
VirConv	0.9623
YOLO	0.8090
CenterFusion	0.3298



Node connections as seen through ROS



YOLO Confusion Matrix Output



YOLO Precision-Confidence Validation Plot

Future Direction

- Train VirConv-S model for better timeliness
- Train a model of CenterFusion on the WWU cluster for better overall performance
- Integrate V2V fusion
- Integrate multimodal fusion in real-time
- Improve multimodal object list with a Kalman Filter and Multi-Object Tracking (MOT)
- Produce the energy consumption and flops of the VirConv and CenterFusion algorithms

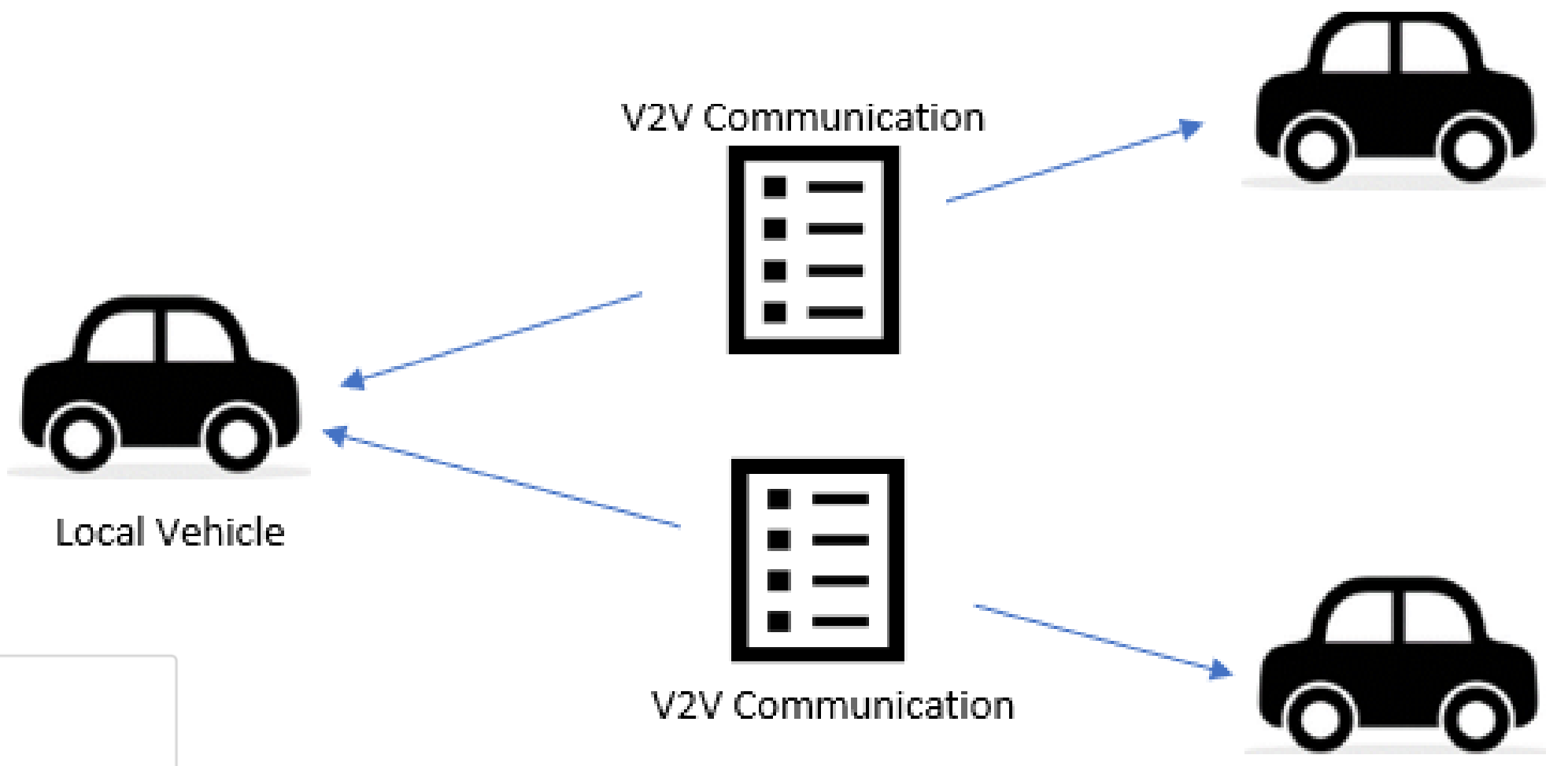


Diagram of V2V communication

Acknowledgements

This work is supported by PACCAR, who sponsored the parts and equipment for this project. We also received assistance from our PACCAR engineer of contact, James Nelson, and project mentor, Junaid Khan.
Additionally, Reza Afshari helped us with parts acquisition.
Special thanks to our professors Ying Lin and Yuzhang Zang.

2025 Capstone Project
Electrical & Computer Engineering Program