Problem Statement and Goals
Equivariant Sensor Fusion

Alaap Grandhi

Table 1: Revision History

Date	Developer(s)	Change
Jan. 21	Alaap	Initial Commit

1 Problem Statement

[You can change the section headings, as long as you include the required information. —SS]

Over the past few years, we have seen more and more autonomous vehicles being allowed onto the road. With this increase, the need for perception methods that can effectively use readings from sensor modalities like Camera and LiDAR to inform these vehicles of their surroundings has similarly increased.

1.1 Problem

While sensor modalities like LiDAR and Camera can capture meaningful information about road conditions, they are inherently prone to failure. LiDAR sensors can fail in scattering media like snow while cameras are easily occluded. Thus, learnt methods combining the readings from these modalities have become the state-of-the-art for this task due to their ability to overcome individual sensor modality failures. Still, these techniques suffer from local misalignment between the features obtained for these modalities. This motivates the need for new approaches that can robustly combine readings from these modalities.

1.2 Inputs and Outputs

For the purpose of this document, the process by which the given network/model is trained is out of scope. Given this, the inputs in this problem would be a set of camera images taken from different poses (multi-view) alongside a pointcloud obtained from a lidar sensor. Subsequently, the outputs would be a set of labelled bounding box predictions corresponding to pedestrians and vehicles observed in the robot's surroundings.

1.3 Stakeholders

1.3.1 Academic Stakeholders

Academic stakeholders for this task would primarily include Computer Vision and Robot Path Planning researchers in the field of Autonomous Driving. With the sheer volume of research being conducted in this field using public datasets, this work would aim to be an easily integratable and easily adaptable approach to alignment-aware sensor fusion for autonomous driving. Thus, it could serve as a basis for future research in the field by academic institutions.

1.3.2 Industrial Stakeholders

Industrial stakeholders for this task would primarily include companies like Waabi, Waymo and Tesla that produce autonomous vehicles.

1.4 Environment

1.4.1 Software

The software should be compatible with any up-to-date operating system.

1.4.2 Hardware

Given that the solution space will revolve around learned methods, any computer possessing a machine-learning capable GPU with at least 8GB vram will be suitable for training and inference of models. However, this vram requirement would be significantly lower for inference-only settings using pre-trained models.

2 Goals

1. Given a set of camera images $\{C_i\}$ and a lidar pointcloud P obtained from an autonomous vehicle at a given point in time, determine the set of bounding boxes for dynamic entities in its surroundings $\{B_i\}$

2. Given a common standard training dataset (i.e. Waymo, NuScenes, Kitti), the trained model should achieve comparable accuracy (mAP) to existing state-of-the-art methods

3. In the presence of disturbances in one of the two given modalities, the trained model should achieve comparable accuracy (mAP) to a network solely trained on and processing the undisturbed modality.

3 Stretch Goals

1. Given a set of disturbed camera images $\{C_{d,i}\}$ and a disturbed lidar pointcloud P_d, achieve better accuracy (mAP) than a network solely trained on either modality.

2. Produce a range of architectures that can each meet the aforementioned requirements and thus be chosen from based on the specific dataset used.

4 Challenge Level and Extras

This project would be considered an advanced research project as it will include the development of newer state-of-the-art perception architectures (a task that typically can lead to published paper). Additionally, although beyond the scope of this document, the approaches considered will utilize group theoretic principles and Equivariant architectures to improve alignment between sensor modalities. This would in turn introduce the need for custom layers (rather than just combining a set of off-the-shelf components).