M UN ER
MOTORVEHICLE
UNIVERSITY OF
EMILIA-ROMAGNA

Università di Parma

3D Perception, Learning-Based Data Fusion Exam

3D Object Detection With Transformers

Francesco Marotta, Simone Maravigna

November 27, 2023

Project Report

Goal of the Project

The goal of this project is to develop an **object detection** system for autonomous vehicles operating in urban environments using the nuScenes dataset[1]. Specifically, the system can **detect**, **localize** and **classify** urban objects between 10 classes: car, truck, construction vehicle, bus, trailer, barrier, motorcycle, bike, pedestrian, traffic-cone.

Dataset

The **nuScenes** dataset is a comprehensive and diverse dataset designed for autonomous vehicle research and development. It provides a rich collection of sensor data, including lidar, radar, and camera information, captured from a variety of urban driving scenarios. It includes **1000 scenes**, of the duration of **20 seconds**, divided in 700, 150, 150 respectively for training, validation and testing.

Sensors

In the pursuit of generating a high-quality multi-sensor dataset, the calibration of both extrinsic and intrinsic parameters for each sensor is imperative. Extrinsic coordinates are defined in relation to the ego frame, specifically the midpoint of the rear vehicle axle.
To achieve effective cross-modality **data alignment** between LIDAR and cameras, the camera's exposure is synchronized when the top LIDAR sweeps across the center of the camera's field of view (FOV). Since the operating frequencies of the cameras and LIDAR are different not every LIDAR scan corresponds to a camera frame. Therefore, within each scene, frames are annotated at a rate of **2 Hz** to ensure data alignment, resulting in 40 frames per scene.
Calibrating sensors, such as cameras and a LIDAR involves establishing a precise relationship between their respective coordinate systems in relation to a common ego reference frame. The ego reference frame represents the motion and orientation of the vehicle or system to which the sensors are attached. In this case, the ego reference frame is located at the midpoint of the rear vehicle axle, precisely where the IMU is positioned.
Using mathematical techniques to estimate the transformation matrix that relates the coordinate systems of the camera and the LIDAR to the ego reference frame. This matrix includes information about **translation** and **rotation**. The ego reference frame is connected to the camera reference frame or LIDAR reference frame through:

- A rotation around the optical center, expressed by a **rotation matrix R**
- A translation expressed by a **translation vector T**

All of these data are provided by the dataset.

Inference

A **pre-trained** model on nuScenes training subset (850 scenes) of the **Transfusion**[2] was utilized in this setup, which involved configuring the following parameters:

- **Image size**: 448x800
- **Data augmentation** like random flipping along both X and Y axes

The optimizer used for training was specified as follows:
`optimizer = dict(type='AdamW', lr=0.0001, weight_decay=0.01).`
Subsequently, **inference** on nuScenes test subset (150 scenes) was conducted to generate **predictions**. For visualization purposes, each box is generated, translated to the ego vehicle coordinate system, moved to the sensor coordinate system and ultimately rendered onto both the images and the LiDAR point cloud.

Results

After the inference, a **67.4%** mAP was achieved, significantly **higher** than the results of the previous project, which had a mAP of 48% for camera images and a 52% mAP for the LIDAR point cloud.
In conclusion, the model demonstrates **excellent performance** and exhibits robustness under challenging image conditions. It is important to observe that, in comparison to the LiDAR-based model from the previous project, the fusion of Camera-LiDAR has yielded a generally **improved** mean Average Precision and better results for small objects such as traffic cones, bikes, and motorcycles.

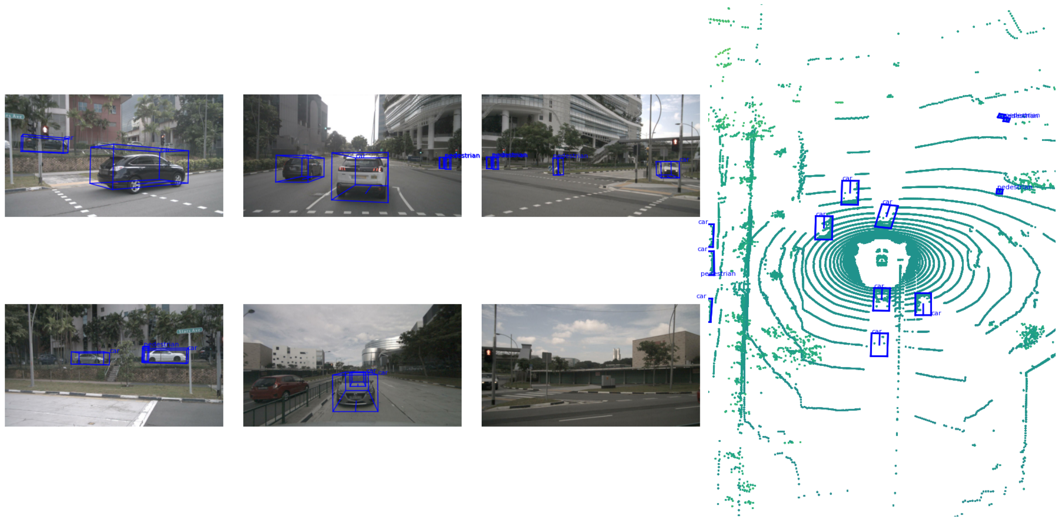

Figure 1: Example of plotted prediction

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, 2022.