# Real-Time Object Detection for Self-driving Cars

Cristiano Langner [1]   Kelmer M. Cunha [2]   Pedro A. F. Castro [1]

[1]Postgraduate Program in electrical engineering   [2]Postgraduate Program in Biology of Fungi, Algae and Plants

## Introduction

Since global industrialization, road transportation has been essential for many aspects of industrial and day-to-day life. However, high road dependence has created a global health issue, especially in Brazil. The implementation of emerging technologies, such as autonomous cars and driver assistance systems, could be a great opportunity to reduce human error in traffic, which is the leading cause of accidents. These systems have deeply evolved since the advent of Convolutional Neural Networks, which enabled safe implementations in real-time object detection in traffic contexts. This project aimed to develop a CNN-based application for object detection in real-time, focused on fine-tuning widely used architectures to the Florianópolis region traffic context.

## Methodology

This project made use of 2 datasets to test how the application would work in each scenario. All experiments were based on the You Only Look Once (YOLO) architecture versions 3 and 11, which are pre-trained based on the MS-COCO dataset. First, we fine-tuned the model on the public dataset KITTI and evaluated its performance using the model for real-time inference (camera stream inside the car). After this, we created and implemented a custom dataset for fine-tuning and again tested the model's performance in the real-time scenario.

- The **KITTI dataset** (Karlsruhe Institute of Technology and Toyota Technological Institute) is one of the most popular datasets for use in autonomous driving. It consists of 7481 training images and 7518 test images, comprising a total of 80.256 labeled objects. The classes that the dataset contains are: car, pedestrian, van, cyclist, truck, misc, tram, and person sitting. The images were split into training and test sets, converting the default COCO format to the YOLO format.
- For the **Custom Dataset**, we used a base of 9 videos recorded over the region of Florianópolis, in Santa Catarina (Brazil), from several conditions of illumination (rain, cloud, sunny, and night) to capture diverse scenarios. The videos were recorded using only a horizontal view, inside/out of a car, with a resolution of 720p (1280 x 720 pixels) at 30 Frames Per Second (FPS). Frames from all videos were randomly taken and manually curated. We used the auto-labeling tool for a standardized labeling, which was manually verified. The dataset is composed of 1447 images, that were splited into training (1110), validation (224), and test sets (93). There are 9570 annotations in total, representing the classes bicycle (62), bus (117), car (5419), motorcycle (528), person (827), sign (traffic signs — 1809), traffic light (370), and truck (438).

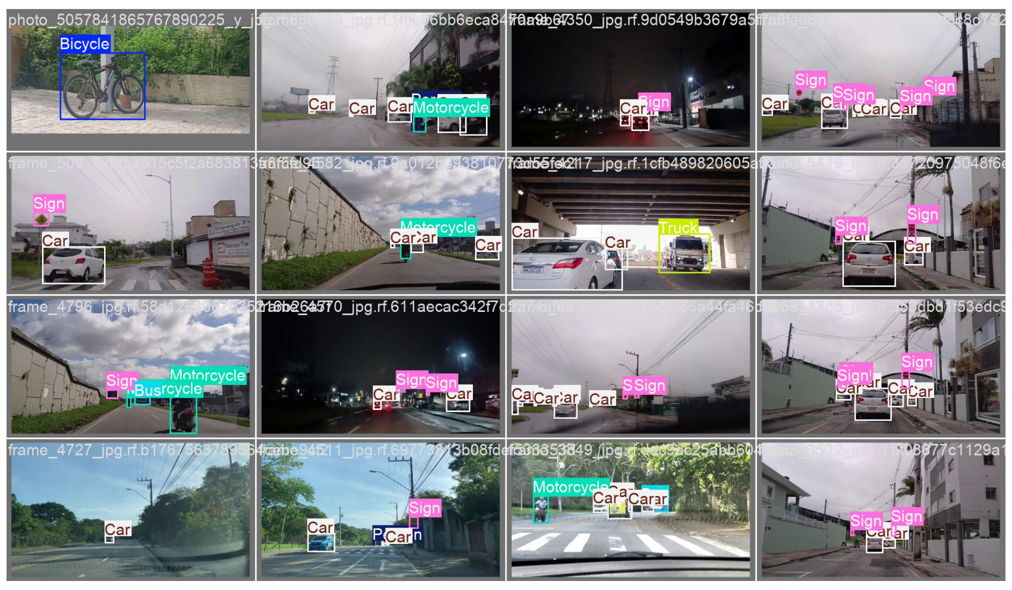

Figure 1. Sample images from the custom dataset, showing the diversity of its recordings. Source: CompVisionProjectUFSC (2024).

Before passing the images to the networks tested, a pre-processed step was taken. Data augmentation was applied to both the KITTI and the custom dataset, with the transformations and hyperparameters listed bellow:

| Transformation | Parameters and Values |
|---|---|
| Blur | Probability: $p = 0.01$<br>Blur limit: $(3, 7)$ |
| MedianBlur | Probability: $p = 0.01$<br>Blur limit: $(3, 7)$ |
| ToGray | Probability: $p = 0.01$<br>Output channels: $3$<br>Method: Weighted Average |
| CLAHE | Probability: $p = 0.01$<br>Clip limit: $(1.0, 4.0)$<br>Tile grid size: $(8, 8)$ |

Table 1. Transformations done on both datasets using the Albumentations library. Source: authors.

When evaluating a model's performance over different architectures or datasets, it is important to establish metrics to check the performance of each test on the proposed task. Five metrics were considered: precision, recall, F1-Score, mAP50, and mAP50-95, where the **mAP50** and **recall** were our goal metrics.

## Experiments and Results

Considering the YOLOv3 and YOLOv11 architectures for real-time object detection in the traffic context of Florianópolis, experiments were setup as follows:

1. **YOLOv3 — Used as default (pre-trained on the MS-COCO dataset).** Implemented as a baseline, this model performed poorly on the Florianópolis traffic context, despite being widely used in autonomous driving research. Misclassifications were frequent for bicycles and motorcycles, with a low mean Average Precision (mAP), highlighting the need for domain-specific adaptation.
2. **YOLOv11s (small) — Fine-tuned on the KITTI dataset for 20 epochs.** This implementation showed improvements over YOLOv3. It has an associated mAP50 of 0.90, and a mAP50-95 of 0.65. While with better performance, still struggled with distinguishing bicycles from motorcycles and accurately identifying buses under specific conditions.
3. **YOLOv11m (medium) — Fine-tuned on the custom CompVisionProjectUFSC dataset for 200 epochs.** Fine-tuning YOLOv11m on the custom dataset yielded reliable results, achieving a mAP50 of 0.714 and a mAP50-95 of 0.475, with recall values exceeding 0.6 for most classes.

## YOLOv11m performance on the custom dataset

The implementation showed high accuracy in detecting Signs (85%), Traffic Lights (77%), and Cars (78%). Significant confusion between Bicycles (33% accuracy) and other classes such as Motorcycles, highlights the model's difficulty in distinguishing visually similar objects (Figure 2). Also, Bus (60%) is often misclassified, possibly as cars or trucks, due to variations in angles or insufficient representation in the training dataset.

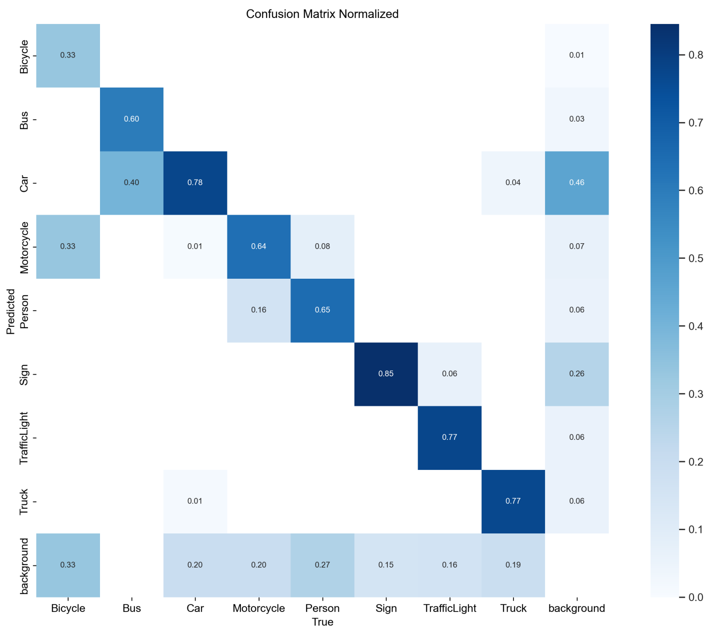

Figure 2. Normalized Confusion Matrix - YOLOv11m Fine-tuned on CompVisionProjectUFSC.

Next, the trade-off between precision and recall for each class was evaluated. Traffic Light (0.843 AUC), Car (0.804 AUC), and Motorcycle (0.777 AUC) were the best performing classes, whereas Bus (0.717 AUC) and Bicycle (0.389 AUC) were the most challenging ones (Figure 3).

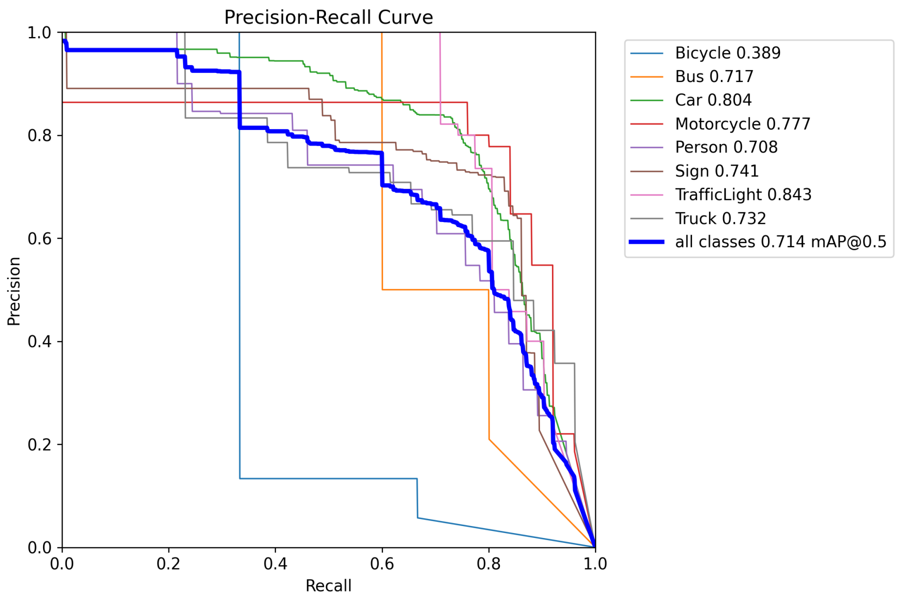

Figure 3. Precision-Recall Curve - YOLOv11m Fine-tuned on CompVisionProjectUFSC.

Overall, our model exhibits competitive precision and recall, with stable loss and metric curves. The slightly lower mAP50-95 (0.47 compared to 0.65 in KITTI) may indicate that your custom dataset has higher variability or fewer samples for specific classes. However, the model is well-fitted to your dataset, demonstrating good generalization and high accuracy. Possible next steps could include increasing the sample diversity for worst-performing classes, analyzing examples (false positives and negatives) for targeted adjustments, and performing more comprehensive data augmentation.

Code available at: http://bit.ly/49oXY4Z   NE410121-41000025DO/ME (20242) — Computer Vision   Dataset available at: https://bit.ly/3ZjyxNu