Final Project- Machine Learning for Robotics (RBE 577)

Object Detection and Depth Estimation for Drone Camera Images

The goal of this project is to generate bounding box and depth for the cars in the images from the camera on-board a drone.

We are using the SynDrone synthetic drone dataset provided in the paper, G. Rizzoli, et. al, "SynDrone – Multi-modal UAV Dataset for Urban Scenarios", 2023 and in this repository.

Please only use **Town01** data.

The project has two parts:

Part 1: Object Detection

In this part you will be finetuning Ultralytics's Yolo v11 object detection that can be found here. The objective is to predict bounding boxes for the cars in the camera images.

Part 2: Depth Estimation

In this part, you will be finetuning the supervised depth estimation method proposed in R. Ranftl, et. al, "Vision Transformers for Dense Prediction", 2021.

The objective is to predict the image depth and essentially get the depth of each bounding box that's predicted in part 1 (you only need to provide depth for the whole image not a specific bounding box).

For the above depth estimation algorithm the authors provided their model in this GitHub page. However, the provided code does not have a training script.

Team Collaboration:

- Teams of two can collaborate on the homework.

Note on Grading:

Grades will be based on the quality of the submission, emphasizing effort, and problem-solving rather than just achieving perfect results. Submissions will be ranked based on the best work produced, and we will fairly consider whether the project was completed individually or as a team of two, with adjusted expectations accordingly.

Final Deliverables:

1. 10-minute presentation highlighting challenges, and results.
2. Submission of presentation slides, a presentation video, and well-documented, organized code.
3. For part one of the project, you must:
    a. Write a training dataloader for Yolo v11 and finetune the Yolo v11 model for the SynDrone dataset.
    b. provide plot of the loss function in training as a function of epoch.
    c. Provide 10 example images of predicted bounding box versus ground truth.
4. For part two of the project, you must:
    a. Write a training script and dataloader for the Vision Transformer model in Ranftl, et. al. and finetune the model for the SynDrone dataset.
    b. Provide plot of the loss function in training as a function of epoch.
    c. Provide 10 example images of predicted depth versus ground truth depth.