ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

3D SEMANTIC OCCUPANCY PREDICTION FROM MONOCULAR CAMERA

MILESTONE 1

DEEP LEARNING FOR AUTONOMOUS VEHICLES

Max Henking & Baptiste Poffet

7th April 2023

CHAPTER 1

INTRODUCTION

1.1 PROBLEM DEFINITION

The problem defined in the article is to predict a dense semantic scene within a certain volume in front of a vehicle using only RGB images as input. The goal is to train a neural network that can generate a semantic voxel grid that is as close to the ground truth as possible. This voxel grid represents the 3D space in front of the vehicle, with each voxel being either empty or occupied by a certain semantic class. The main challenge is to achieve accurate geometric completion and semantic segmentation simultaneously, which is crucial for obstacle avoidance and safe navigation in autonomous driving applications.

1.2 INPUTS

The inputs of our model will be monocular camera images. Monocular camera images are images captured by a single camera lens, as opposed to stereo images, which are captured by two cameras or lenses positioned slightly apart. In computer vision, monocular images are often used for various tasks such as object recognition, tracking, and scene understanding. However, since monocular images are captured by a single lens, they lack the depth information that can be obtained from stereo images or depth sensors.

FIGURE 1.1
Example of a monocular camera image.

1.3 OUTPUTS

The outputs of our model will be 3D semantic occupancy maps of the scenes. The 3D occupancy map will be represented as a voxel grid. A 3D semantic occupancy map is a representation of a physical environment that combines both spatial and semantic information. It provides not only the spatial layout of the environment (i.e., the locations of obstacles and free space), but also assigns semantic labels to different regions or objects within the environment.

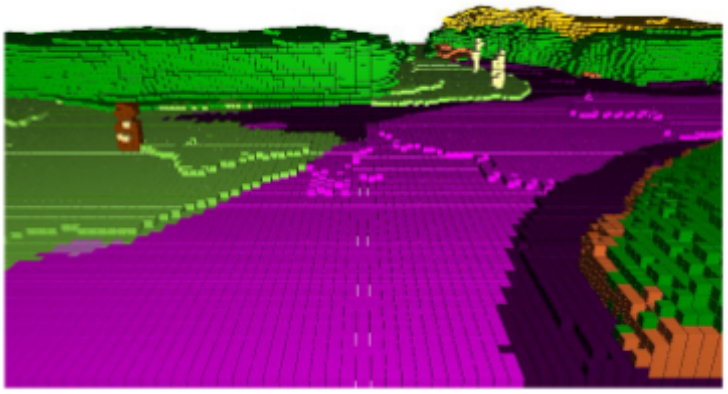

FIGURE 1.2
Example of a 3D semantic occupancy map.

CHAPTER 2

METHODOLOGY

2.1 OVERVIEW OF THE SELECTED METHOD

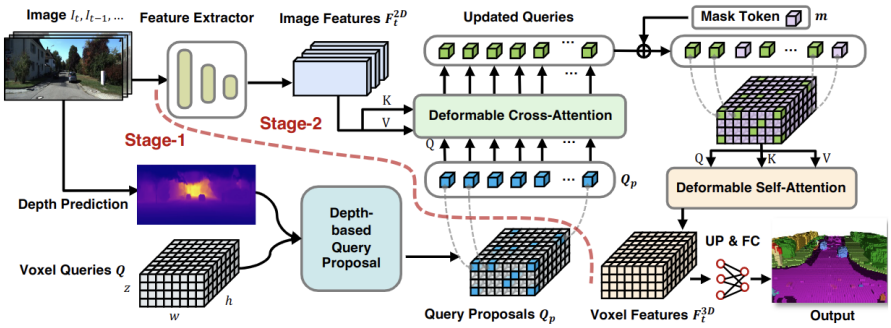

FIGURE 2.1
Architecture of our model.

The method in the article is a two-stage framework designed to predict dense semantic scenes in a 3D voxel grid in front of a vehicle using only RGB images. In the first stage, class-agnostic query proposals are generated by estimating depth from the input RGB images and correcting the depth using a lightweight UNet-like model to predict occupancy. Voxel queries are then selected based on the binary occupancy map resulting from depth correction.

In the second stage, class-specific segmentation is performed. Deformable cross-attention is applied to sample features from hit views around projected 2D image features. Deformable self-attention is used to refine the initial voxel features obtained from the updated query proposals and mask tokens. Finally, the refined voxel features are upsampled and projected to the output space to obtain the final dense semantic map.

Overall, the method combines depth estimation, correction, query proposal generation, and deformable attention mechanisms to refine voxel features, producing a dense semantic map representing the 3D environment in front of the vehicle.

Also, this method is one of the latest research done on the field, has been publicly shared on Github : https://github.com/NVlabs/VoxFormer and the evaluation of the Github project is great.

CHAPTER 3

DISCUSSION

3.1 DATASETS

The dataset used in the article is SemanticKITTI, which is a subset of the KITTI Odometry Benchmark dataset containing dense semantic annotations for each LiDAR sweep across 22 outdoor driving scenarios. SemanticKITTI is specifically designed for semantic scene completion (SSC) tasks and focuses on a volume of 51.2m ahead of the car, 25.6m to the left and right side, and 6.4m in height.

This volume is voxelized into a 3D voxel grid with dimensions of $256 \times 256 \times 32$, where each voxel has a size of $0.2\text{m} \times 0.2\text{m} \times 0.2\text{m}$. The voxel grids are labelled with 20 classes, including 19 semantic classes and 1 free (empty) class. SemanticKITTI provides ground truth semantic voxel grids obtained by voxelizing the aggregated consecutive registered semantic point clouds.

3.2 EVALUATION METRICS

The authors use two metrics to evaluate the performance of their model: Intersection over Union (IoU) and mean Intersection over Union (mIoU). IoU measures the quality of scene completion (i.e., how well the model fills in missing geometry), while mIoU assesses the performance of semantic segmentation (i.e., how accurately the model assigns semantic labels to objects). A good model should perform well in both geometric completion and semantic segmentation.

However, there is a strong interaction between IoU and mIoU, meaning that a high mIoU can be achieved by simply decreasing the IoU. To ensure a thorough evaluation, the authors also examine performance at different ranges ahead of the car: 12.8m x 12.8m x 6.4m, 25.6m x 25.6m x 6.4m, and 51.2m x 51.2m x 6.4m. Understanding short-range areas is crucial, as autonomous vehicles have less time to react to obstacles. In contrast, understanding longer-range areas can be improved as the self-driving vehicle gets closer and gathers more information. The authors report results for different ranges on the validation set and results for the full range on the hidden test set.

3.3 PROPOSED CONTRIBUTION

The objective of this project is to further improve the consistency resulting from the use of the VoxFormer model. We will try to do so by:

- Adding an active object counter: the inclusion of an object counter encourages the model to develop a more structured understanding of the scene, ultimately leading to more consistent and accurate semantic scene completion predictions. It's a good way to analyse if the number of objects stays consistent in time.
- Adding physical constraints for objects: Incorporating physical constraints into the SVT could improve the consistency of the method by ensuring that the generated 3D scene completions adhere to the laws of physics and the inherent properties of objects in the scene (cars, humans, bicyclists,...). This could lead to more realistic and accurate predictions, ultimately enhancing the performance of the model.

BIBLIOGRAPHY

Behley, Jens et al. (16th Aug. 2019). *SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences*. arXiv: arXiv:1904.01416. URL: http://arxiv.org/abs/1904.01416 (visited on 28th Mar. 2023). preprint.

Li, Yiming et al. (25th Mar. 2023). *VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion*. arXiv: arXiv:2302.12251. URL: http://arxiv.org/abs/2302.12251 (visited on 28th Mar. 2023). preprint.

Minaee, Shervin et al. (2021). 'Image Segmentation Using Deep Learning: A Survey'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2021.3059968. URL: https://ieeexplore.ieee.org/document/9356353/ (visited on 28th Feb. 2023).