3D Semantic Occupancy Prediction from Monocular Camera - First report

Max Henking & Baptiste Poffet

7 April 2023

Deep Learning for Autonomous Vehicles, Spring Semester 2023

Definition

This project involves estimating the occupancy and semantic labels of the objects in a 3D scene using sensor data monocular camera images. The occupancy refers to whether a particular voxel or region in the scene is occupied by an object or not, while the semantic labels refer to the specific object categories, such as cars, pedestrians, buildings, and trees.

Inputs

- **Monocular camera images**

Figure 1: Example of a monocular camera image.

Outputs

3D semantic occupancy map of the scene. The 3D occupancy map will be represented as a voxel grid, where each voxel will be labelled with a semantic class. The model also predicts the occupancy probability of each voxel in the voxel grid.

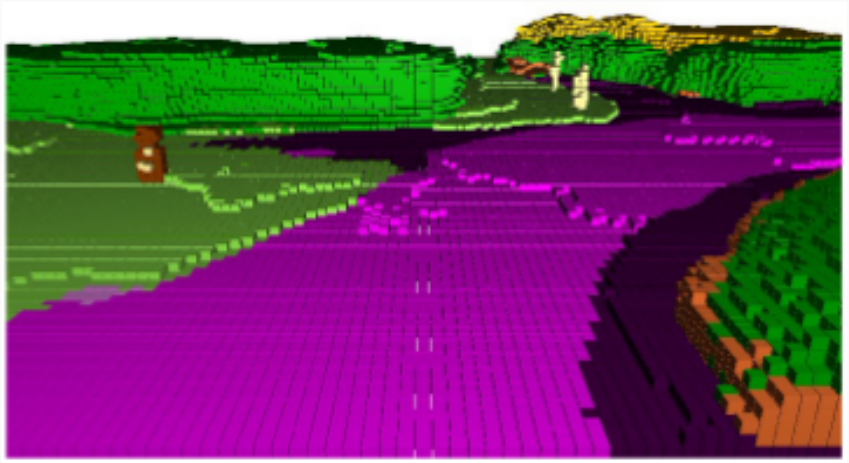

Figure 2: Example of a 3D semantic occupancy map.

Methodology

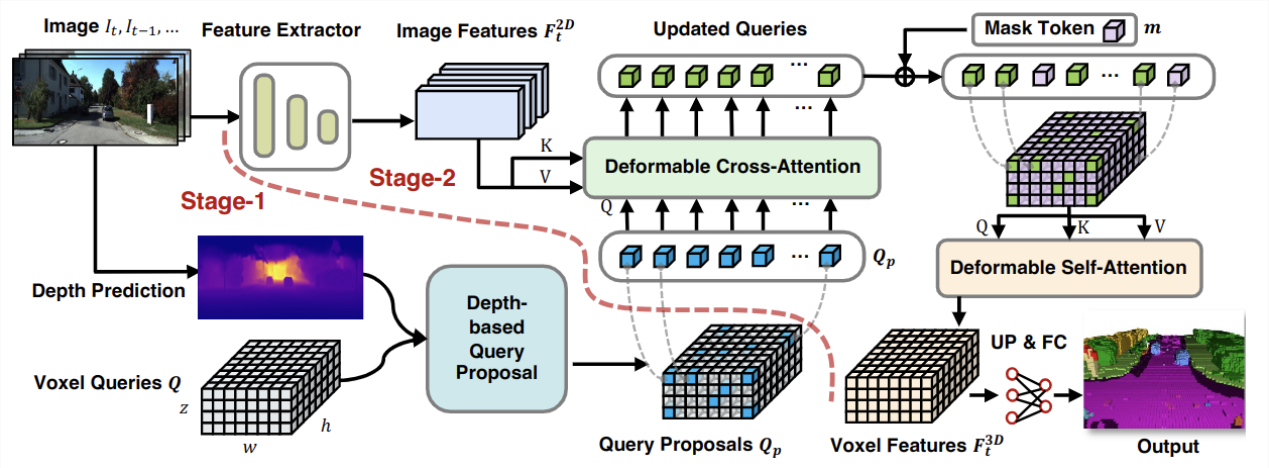

Figure 3: Architecture of our model.

- Stage-1: Depth estimation and binary classification of voxels.
- Stage-2: Semantic classification.

Stage-1

- Prediction of depth Z for each pixel of the image.
- Transformation in 3D using :

$$x = \frac{(u - c_u) \times z}{f_u}, y = \frac{(v - c_v) \times z}{f_v}, z = Z(u, v) \quad (1)$$

- Binary voxels classification (1 if a voxel is occupied by at least one point, 0 otherwise).
- Depth correction using a lightweight UNet model, using as input the binary voxel occupancy map created in the previous step.
- Query proposal (Q_p) based on the selection of depth prediction.

EPFL

Stage-2

- All the non-zero voxels will be classified using several layers of deformable cross-attention.
- Combination of the output of the previous step and the masked voxels to get the complete voxel features.
- Upsample and project the last combination to get the final output with M+1 semantic classes (M classes + 1 class of empty voxels.

Why this method

- It is very recent (2023).
- The results compared to others camera-based semantic scene completion methods are better. [Li et al., 2023]
- *Need to read about other methods to be sure that it is the best we can use*.

Literature

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., and Gall, J. (2019).
SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences.
(arXiv:1904.01416).

Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J. M., Fidler, S., Feng, C., and Anandkumar, A. (2023).
VoxFormer: Sparse Voxel Transformer for Camera-based 3D Semantic Scene Completion.
(arXiv:2302.12251).

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., and Terzopoulos, D. (2021).
Image Segmentation Using Deep Learning: A Survey.
IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1.

Next steps

- Be able to test the model.
- Read articles about other methods and compare.
- Use it with other inputs (ours).
- Find what we can improve (if possible) to get better results.