nuCraft: Crafting High Resolution 3D Semantic Occupancy for Unified 3D Scene Understanding

Benjin Zhu[1], Zhe Wang[2], and Hongsheng Li[1,3,4]

[1] MMLab, The Chinese University of Hong Kong
[2] SenseTime Research
[3] Shanghai AI Laboratory
[4] CPII under InnoHK

Supplementary Materials

A Failure Cases of Previous Occupancy Datasets

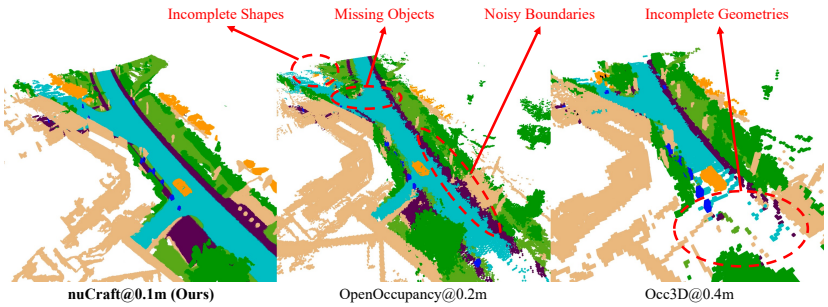

Fig. 1: Failure Cases of OpenOccupancy and Occ3D Datasets

Here we showcase the common failure cases of previous 3D semantic occupancy datasets like OpenOccupancy and Occ3D in Fig. 1. Previous occupancy datasets with lower resolutions (*e.g.*, occupancy with grid size 0.4m) face problems including missing objects, unclear and noisy road boundaries, and incomplete shapes and geometries. As a comparison, our nuCraft dataset overcomes these limitations and presents 3D occupancy with less noises at a higher 0.1m resolution.

B Errors in nuScenes Dataset Inherited by OpenOccupancy and Occ3D

Inaccurate Ego Pose Figure 2 shows a cross-sectional view of aggregated LiDAR point clouds along the Z-axis before and after pose estimation. The inaccurate ego poses provided in the raw nuScenes annotations cause great challenges

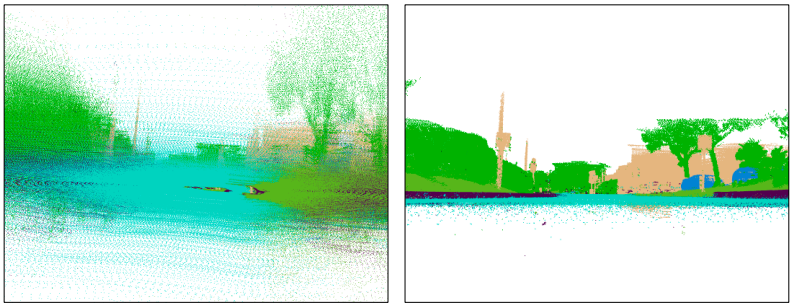

Fig. 2: Cross sectional view of aggregated LiDAR point clouds along $Z-$Axis before and after pose estimation.

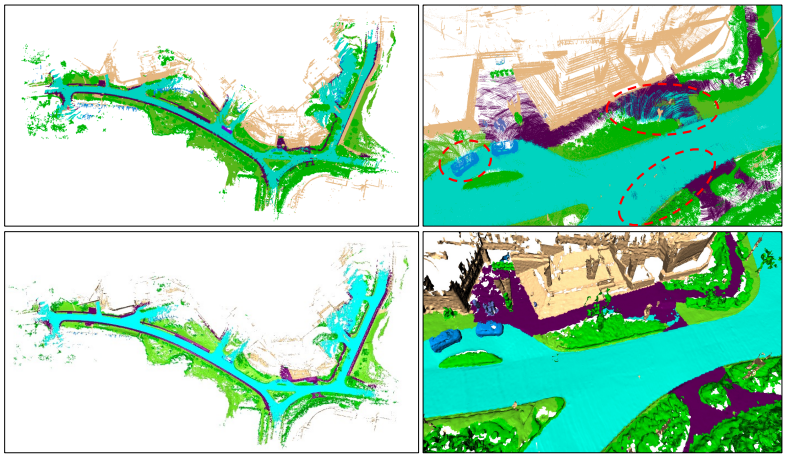

Fig. 3: Given an example scene, the aggregated point clouds (top-left) contains noisy semantic labels (top-right), while the clean reconstructed semantic mesh (bottom-left) demonstrates less noises, complete shapes and clear boundaries (bottom-right).

in obtaining valid dense point clouds. Pose estimation is necessary to better align the LiDAR frames and generate reliable aggregated point clouds for subsequent processing steps.

In-consistent Point Semantic Labels Figure 3 illustrates the difference between the raw point clouds with noisy semantic labels and the reconstructed semantic mesh (before post-processing). The raw point clouds suffer from inconsistent and inaccurate semantic labels. In contrast, the reconstructed semantic mesh exhibits less noise, more complete object shapes, and clearer semantic boundaries. The mesh reconstruction step in nuCraft helps to mitigate the issues with the raw semantic labels and produces a higher-quality representation of the 3D scene.

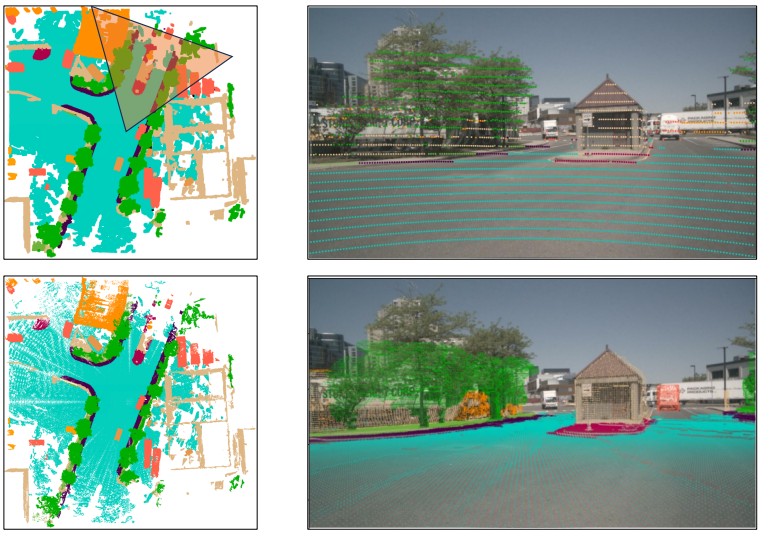

Fig. 4: Visibility Mask. In the first row, we visualize the full semantic occupancy GT in BEV (left), and its corresponding CAM-FRONT (orange triangle) image with its point cloud projection (right). In the second row, we show the computed visibility mask for LIDAR-TOP (left) and CAM-FRONT (right) respectively. For clearer visualization effects, we only vislualize the centroid of occupancy grids, and decrease the point size for the LiDAR visibility mask.

C Visibility Mask of nuCraft

Figure 4 visualizes the visibility mask for camera and LiDAR sensors. To correct misaligned camera poses in partial frames, we utilize the SAM [17] algorithm to obtain object masks and apply the Perspective-N-Points method to compute the transformation between object semantic pixel masks and projected LiDAR points of that object. For the LiDAR visibility mask, we employ a different strategy than Occ3D by reversely starting from the occupancy voxels and computing the occlusion from each voxel to the sensor origin rather than ray marching from sensor origin. This approach allows us to retain more visible occupancy grids than Occ3D.