

Supplementary Materials of FusionOcc: Multi-Modal Fusion for 3D Occupancy Prediction

Anonymous Authors

1 DETAILS OF POINTS BRANCH

The detailed description of our voxel encoder used in FusionOcc is given in Fig. 2. As the final occupancy task requires the voxel feature map to be consistent with the size and range of the feature map of images in the shared 3D space, we voxelize the point clouds with range: $x \in [-40, 40]$, $y \in [-40, 40]$, $z \in [-1, 5.4]$, where (x, y, z) denotes the coordinates in 3D space. The voxel size is set as 0.05m, and the voxelized points are fed into several Submanifold sparse convolution[1] blocks. Firstly, a 1×1 convolution is used to expand the number of channels from 5 to 16. Then, a 3 times repeated sparse block is adopted to down-sample the feature sized of $[1600, 1600, 128, 16]$ to $[200, 200, 16, 64]$. Finally, another 1×1 convolution is utilized to reduce the number of channels from 64 to 32 to obtain the feature sized of $[200, 200, 16, 32]$.



Figure 1: Integrate point clouds from different numbers of adjacent frames and finally perform random sampling on the merged point clouds. 2,4,8 denotes the number of adjacent frames being merged.

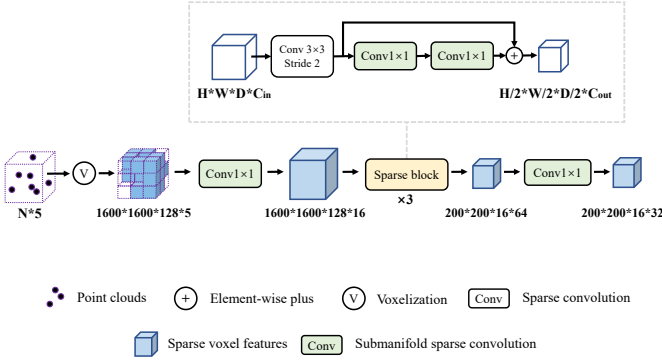


Figure 2: Details of our voxel encoder.

As verified by experiments in the paper, relatively dense point clouds will greatly promote the model's performance on accuracy. We merge point clouds of several adjacent frames and adopt a random sampling to reduce the amount of computation (see Fig. 1). The merged point clouds are much denser than the original one and will provide more details of objects in the scene.

2 DECLARATION

Our code will be made publicly on GitHub. The video in the supplementary material consists of clips from various weather conditions in the validation set of Occ3d-nuScenes.

REFERENCES

- [1] Benjamin Graham and Laurens Van der Maaten. 2017. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307* (2017).