APNet: Urban-level Scene Segmentation of Aerial Images and Point Clouds —- Supplementary Material —-

Weijie Wei Martin R. Oswald Fatemeh Karimi Nejadasl Theo Gevers
University of Amsterdam
{w.wei2, m.r.oswald, f.kariminejadasl, th.gevers}@uva.nl

Abstract

This supplementary material provides additional details about the implementation.

A. Implementation details

A-branch. We adopt HRNet [4] with object-contextual representations (OCR) [5], denoted as HRNet-OCR, as the backbone for A-branch. During training, the OCR loss is preserved while the original 2D segmentation head is removed. The intermediate features, also known as augmented representations as defined in the original paper, from HRNet-OCR are compressed to a total of 128 channels, thereby ensuring alignment with the output of the P-branch.

P-branch. We employ RandLA-Net [2] as the backbone for the P-branch and follow its official configuration for the SemanticKITTI dataset[1] with the following two modifications: Firstly, we double all feature channels in the RandLA-Net to accommodate the additional color features. Furthermore, we double the output channel for the last layer to ensure compatibility with the A-branch. Consequently, the encoder produces outputs with channel dimensions of 64, 128, 256, and 512, respectively. Secondly, we input the same point cloud to RandLA-Net twice and sum up the output features. Although the network does not change, the down-sampling within the network is random, leading to different features for the same point cloud in the end. This technique promotes the consistency of RandLA-Net.

GAF module. We adopt KPConv [3] as the point convolution in the GAF module and adhere to the configuration of the rigid KPConv. Accordingly, one single rigid KPConv encompasses a sphere with a radius of 0.5 meters, centered at the query point. Each kernel point exerts an influence on all support points within a sphere whose radius is 0.24 meters and centered on the kernel point.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9296–9306. IEEE, 2019.

[2] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-net: Efficient semantic segmentation of large-scale point clouds. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11105–11114. IEEE, 2020.

[3] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas Guibas. KPConv: Flexible and deformable convolution for point clouds. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6410–6419, 2019.

[4] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[5] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020.