Supplementary Materials of BioNet and NeFF: Crop Biomass Prediction from Point Clouds to Drone Imagery

Xuesong Li[1,3], Zeeshan Hayder[2], Ali Zia[1,3], Connor Cassidy[1], Shiming Liu[1], Warwick Stiller[1], Eric Stone[3], Warren Conaty[1], Lars Petersson[2], Vivien Rolland[1]
[1]CSIRO Agriculture and Food, [2]CSIRO Data61, [3]Australian National University, Australia
xuesong.li@csiro.au

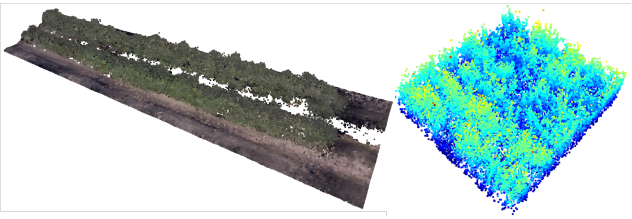

Figure 1. Visualization of a point cloud from the MMCBE (left) and wheat (right) datasets. This figure comes from [4].

1. Implementation

In our NeFF, the SDF network $\mathcal{E}_g$ is modeled by an MLP built by 8 hidden layers, and Relu with Softplus is used as the activation function for all hidden layers. A skip connection is used to connect the input and the fourth layer. The feature filed $\mathcal{E}_f$ consists of an MLP with two hidden layers, while the color filed $\mathcal{E}_c$ has four hidden layers. All hidden layers have the same hidden size of 256. Positional encoding is applied to spatial locations with 6 frequencies and viewing directions with 4 frequencies. We use 4 sparse 3D CNN blocks, as shown in Fig. 2, to compress the 3D features map into a 2D one, and the 3 Transformer encoders have the same defaulting as the original paper [6]. The final prediction MLP includes two hidden layers with 512 and 256 hidden sizes. We use an ADAM optimizer [3] to optimize both networks on a Navida A5500 GPU. The $\mathcal{F}$ and $\mathcal{H}$ functions are trained separately. It takes 10 hours to optimize NeFF for 200K iterations, and around 8 hours to train BioNet for 100K iterations with 4 batch sizes.

2. Difference in two datasets

When comparing our method with baselines in Section 5.2, we find that the MARE suggests that the same approach achieves superior results on the wheat dataset [5] compared to MMCBE [4]. To delve deeper into these differences, we visualized the point clouds from both datasets, as shown in Figure 1. Significant occlusion issues exist in MMCBE, where LiDAR data is often scanned from the side, contrasting with the top-down collection approach in the wheat dataset. This difference can lead to a higher quality of point cloud data for the wheat dataset, thereby enhancing the accuracy of biomass prediction.

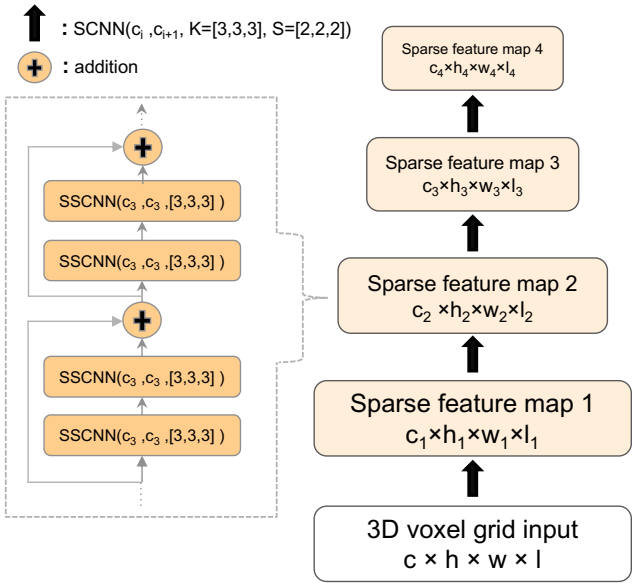

Figure 2. Architecture of the sparse 3D CNN backbone. The input is the voxelized 3D features generated from NeFF. SSCNN stands for submanifold sparse 3D CNN [1], while SCNN stands for sparse 3D CNN. Each pyramidal level consists of two ResNet blocks [2].

3. 3D backbone network architecture

We are using the 3D backbone network to further extract 3D features from the point cloud or distilled feature maps, and the 3D backbone network mainly consists of sparse 3D CNN layers due to their advantages in efficient computation and local geometrical feature extraction. The details of 3D

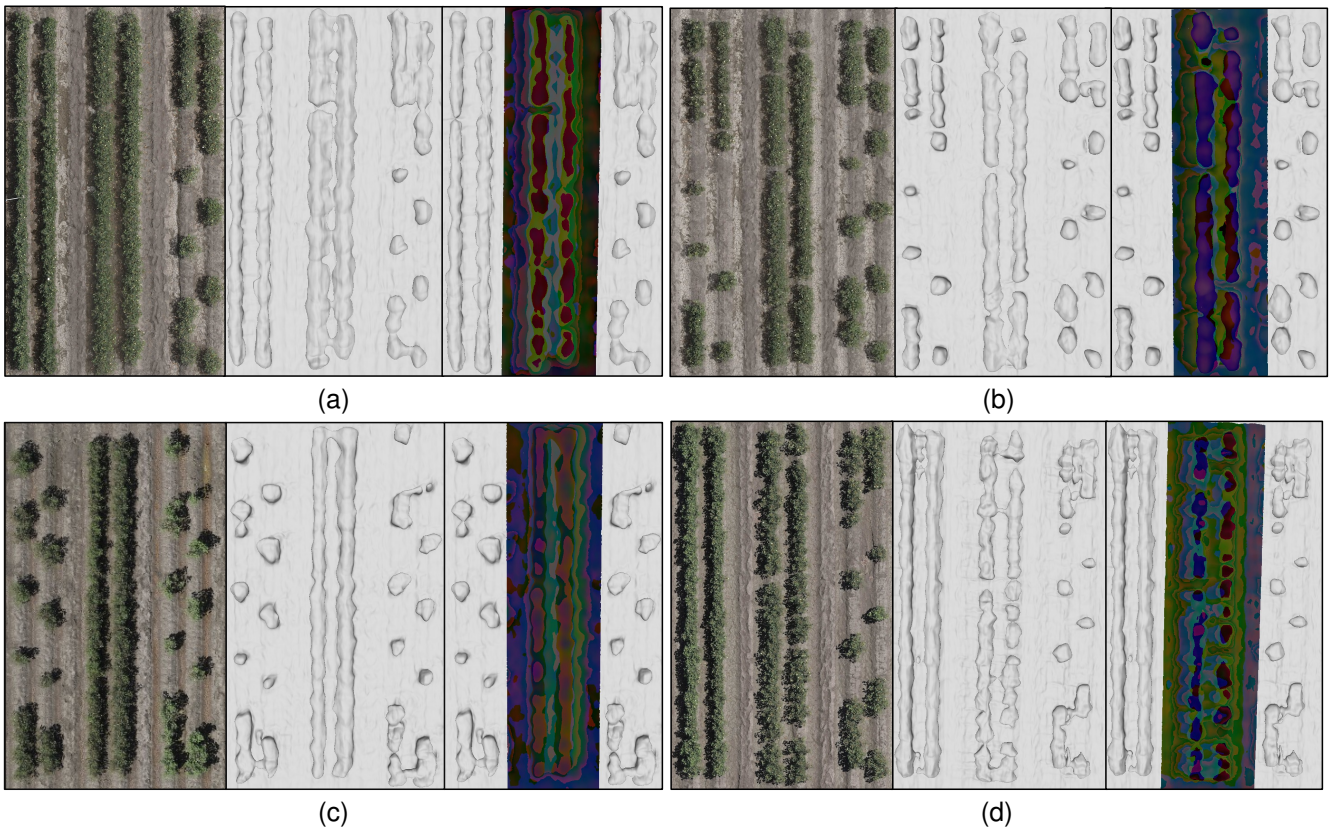

Figure 3. The visualization of 3D features generated by NeFF module. For each fused image, the stitched RGB image is on the leftmost, a 3D reconstruction is in the middle, and the rightmost image shows cropped 3D features (used by the BioNet module for biomass prediction.)

backbone network are shown in Fig. 2. For the MMCBE dataset, we first voxelized the inputs (i.e. point cloud or 3D feature maps) to obtain its voxel representation. The size of the 3D inputs is 16 meters in length, 4 meters in width, and 1.5 meters in height, with a voxelization resolution of 0.08 meters in length, 0.05 meters in width, and 0.075 meters in height, respectively. For the wheat dataset, the size of the 3D input is 1 m^3, the voxelization resolution is 0.008 meters in width and length, and 0.025 meters in height. The value of each voxel is obtained by averaging all points inside a given voxel. The computational operations mainly include regular $3\times3\times3$ convolutional kernel, submanifold CNN [1] and max pooling.

4. Evaluation metrics

Metrics used in the paper are namely mean absolute error *(MAE)*, mean absolute relative error (*MARE*), and root mean square error *(RMSE*), as shown in equation 1 2 3. *MARE* offers the advantage of being more robust against the range of ground truth values observed across our 9 time-points.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|m_i - \hat{m}_i| \tag{1}$$

$$\text{MARE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|m_i - \hat{m}_i|}{m_i} \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m_i - \hat{m}_i)^2} \tag{3}$$

Where N is the total point cloud data points with $i \in \{1, \ldots, N\}$ as the index, $\hat{m}_i$ is the predicted and m_i the ground-truth biomass value. The lower these errors, the better the prediction accuracy. We used RI to indicate the relative improvement when comparing two different methods. The accuracy of two methods are $p1$ and $p2$, (say $p1$ is better than $p2$), the R1 of $p1$ against $p2$ is as follows:

$$RI = \frac{|p1 - p2|}{p2} \tag{4}$$

5. NeFF

To visualize the results from the NeFF, we sample 3D features at a resolution of 1024^3 from the neural feature field (the feature sampling resolution used for BioNet module in 2048^3) in MMCBE dataset. We then employ principal component analysis to compress the high-dimensional features into three channels. These three channels are subsequently normalized to the range [0, 1] and displayed as the RGB color of reconstructed points, as shown in Fig. 3. From this visualization, it is evident that the distilled 3D features can effectively distinguish between the foreground (plants) and background (soil). Furthermore, different parts of the plants can be separated as well. The NeFF generally constructs 3D features for each plot, encompassing several rows of cotton. However, our specific interest lies in the two 13 m cotton rows, as we possess above-ground biomass ground truth data only for every two rows of cotton. Consequently, we must crop the 3D features of the two cotton rows of interest from the entire set of constructed features. These cropped 3D features serve as the input for the BioNet module, which predicts the final biomass. The cropped 3D features are depicted in each rightmost figure in Fig. 3.

References

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

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] Xuesong Li, Zeeshan Hayder, Ali Zia, Connor Cassidy, Shiming Liu, Warwick Stiller, Eric Stone, Warren Conaty, Lars Petersson, and Vivien Rolland. MMCBE: Multi-modality dataset for crop biomass estimation and beyond. *arXiv preprint arXiv:2404.11256*, 2024.

[5] Liyuan Pan, Liu Liu, Anthony G Condon, Gonzalo M Estavillo, Robert A Coe, Geoff Bull, Eric A Stone, Lars Petersson, and Vivien Rolland. Biomass prediction with 3d point clouds from lidar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1330–1340, 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.