(Supplementary Material)
Neural Hierarchical Decomposition for Single Image Plant Modeling

Zhihao Liu[1,2] Zhanglin Cheng[3] Naoto Yokoya[1,2]
[1] The University of Tokyo [2] RIKEN AIP [3] SIAT, Chinese Academy of Sciences

1. Automatic Dataset Generation

Capturing large, high-quality training datasets of real-world plants is universally recognized as challenging [10]. For trees in particular, even high-precision LiDAR scanners are unable to fully scan their intricate branching details due to significant occlusion. Hence, synthesizing plant models with fine details has become a common choice for dataset acquisition in various plant-related tasks, such as foliage segmentation [2] and point cloud reconstruction [5, 10].

For network training, we automatically generate a large number of realistic 3D plant models from scratch based on parametric L-system technique [17] and self-organization growth [12]. The L-system, widely used in the game and movie industries, stands out as a powerful method for creating random, natural-looking plant models. It biologically simulates the plant structures and growth patterns of various species through a set of structural rules, which are typically represented as a sequence of algorithmic symbols (Fig. 1(a)). By randomly selecting procedural parameters, a large number of variants of the same species can be obtained efficiently. For example, Fig. 1(b) shows three synthetic plant models generated by the same species rules. Meanwhile, when synthesizing the plant models, we can also jointly collect other training data, including plant images, segmentation masks, and corresponding box hierarchies (see Fig. 2).

To make our synthetic images more photo-realistic, we employ the physically based rendering (PBR) technique along with global illumination (GI). We render the plants with varying lighting and material settings, and position the camera at random pitch angles and distances. Additionally, as shown in Fig. 3, we randomly add backgrounds to the images. The backgrounds of outdoor and indoor scenes are sourced from two common image datasets: ImageNet [1] and MIT-Indoor-Scene dataset [13]. As part of data augmentation, we dynamically apply post-processing techniques (e.g., Gaussian blur and film grain) to the synthetic photographs during the training process.

As a result, we synthesized a training dataset containing 12 tree species (e.g., Maple, Oak, Cherry), and 9 common houseplant species (e.g., Anthurium, Monstera Deliciosa, Pilea Peperomioides). We trained each species separately. There are 2k unique plants for each species in the training set, with another 1k plants for validation and testing.

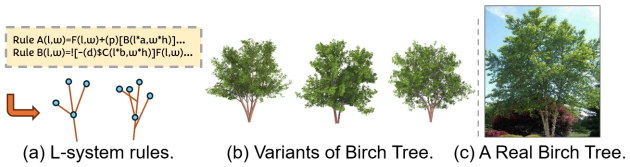


Figure 1. We use L-system to automatically synthesize a large dataset of realistic 3D plant models. (a) The L-system rules define the growth pattern algorithmically. (b) Two Birch trees generated using the same species rules. (c) A real photo of a Birch tree is provided for visual comparison.

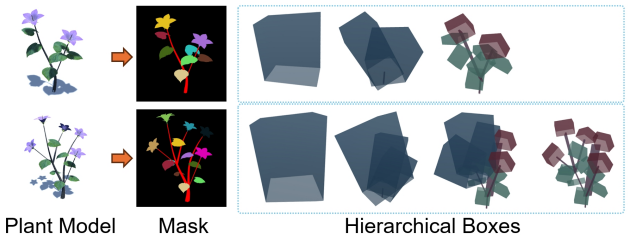


Figure 2. Two synthetic plant variations generated by the same species settings. Their mask images and hierarchical boxes can be simultaneously collected during synthesizing these plant models.

Figure 3. Examples of the synthetic photographs and their corresponding ground-truth segmentation masks. We render plants with different camera transformations, including changes in pitch angles and camera distance, to enhance diversity.

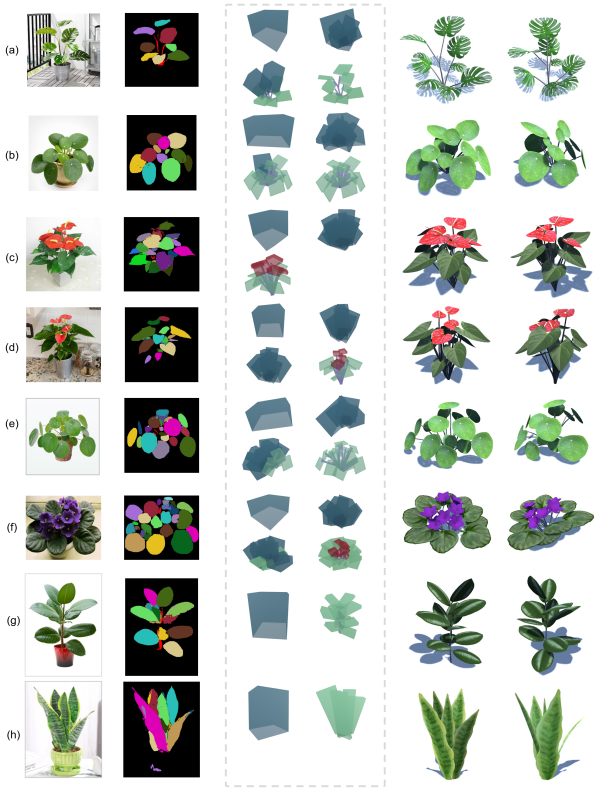


Figure 4. More results of houseplants from real photographs.

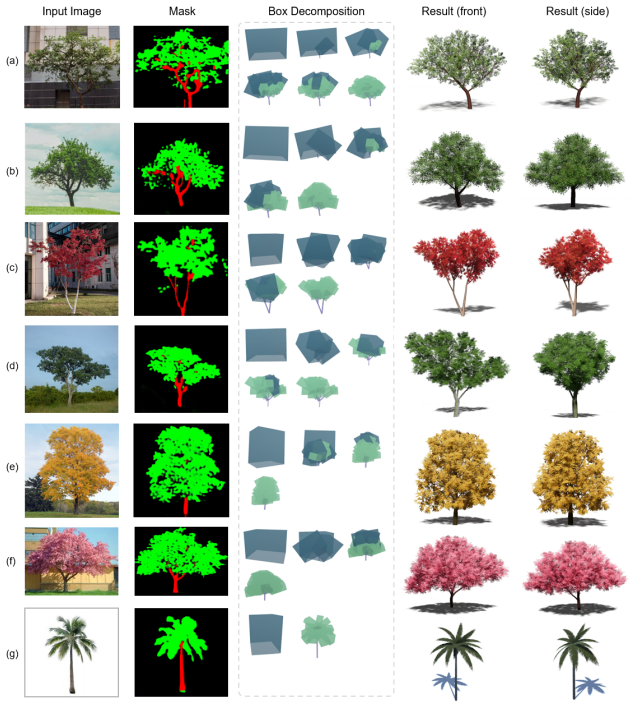


Figure 5. More results of outdoor trees from real photographs.

2. Experiments

2.1. More Results

In Fig. 4 and Fig. 5, we provide more reconstruction results for houseplants and trees, respectively. The results demonstrate that our method can effectively produce realistic 3D plant models from the single images. Table 1 reports the statistics for computing time and decomposition complexity of several reconstructed plants. Overall, our method takes about 158ms on average to infer box structures using our neural network, and 0.79s for synthesizing the final detailed 3D meshes. Thus, the entire reconstruction process is executed with high efficiency.

Table 1. Statistics of several resulting plant models in Fig. 4 and Fig. 5. Time_{HD} is the time for the hierarchical decomposition of box structures; Time_{GC} is the time for the geometry construction of final 3D plant models; Num_{BX} is the number of terminal boxes in the most fine-grained box structures. The upper half of the table is for houseplants, while the lower half is for outdoor trees.

Target Plant	Time_{HD}	Time_{GC}	Num_{BX}
Fig. 4(a)	145ms	0.13s	16
Fig. 4(b)	161ms	0.17s	27
Fig. 4(c)	159ms	0.29s	26
Fig. 5(a)	172ms	1.24s	36
Fig. 5(b)	168ms	1.86s	33
Fig. 5(c)	146ms	1.08s	24

2.2. Ablation Study

We conducted an ablation study to analyze the individual designs of our network. (1) We first examine a variant that directly infers the complete box structure only in a single step from the initial latent vector encoded by the image feature network IFN, rather than expanding the hierarchies progressively. In this case, we increase the number of output boxes for Dec_g to 64, so accordingly the network will face much higher pressure in predicting more nodes and edges simultaneously and accurately. (2) We also compare with a variant that removes the edges in the graphs, where the sub-

Table 2. **Ablation Study.** We compare our full network to two ablated versions: without progressive decomposition and edge connections, respectively. We use the Chamfer and Hausdorff distances to measure the reconstruction quality of the **box structures** for the plants in test dataset.

Condition	Chamfer Dist ↓	Hausdorff Dist ↓
w/o progressive	0.138 (±0.042)	0.242 (±0.106)
w/o edges	0.095 (±0.026)	0.162 (±0.084)
Full network	**0.073** (±0.019)	**0.113** (±0.051)

Table 3. **Quantitative comparison on real-world plants.** We compute the reconstruction error (**Err**) and completeness score (**Comp**) on the real-world plants in Fig. 6, and compare our method with several recent single-image-based reconstruction approaches. We use the dense 3D point cloud P as the approximate ground truth for the plant. The **Err** is the average distance of the points in P to the reconstructed 3D mesh. The **Comp** is the percentage of the points in P that have a distance of less than a given threshold x to the reconstructed mesh. We report the completeness at three threshold values, i.e., 0.10, 0.05, and 0.02. For convenience, the heights of all plants are normalized to 1.00. The results show that our method achieves the best performance.

Methods	**Err**↓	Comp ↑ $x = 0.10$	Comp ↑ $x = 0.05$	Comp ↑ $x = 0.02$
Ours	**0.085**	**78.53%**	**24.86%**	**7.20%**
One-2-3-45[8]	0.124	37.46%	10.39%	6.51%
Wonder3D[11]	0.149	28.59%	9.14%	5.07%
One-2-3-45++[9]	0.136	32.74%	11.25%	6.84%
GeoWizard[3]	0.384	12.63%	4.49%	1.68%
LN3Diff[7]	0.203	19.15%	5.90%	2.77%
DreamGaussian[18]	0.157	25.60%	8.37%	4.39%

decoder Dec_g only infers the graph nodes for the next decomposition step, without edge connections. Table 2 shows the results under different settings. Both ablated variants result in a significant decline in performance.

2.3. Comparison

In this section, we provide more comparison studies against existing single-view-based methods, and even multi-view-based methods. Moreover, quantitative measurements are also included.

Comparison to Single-view-based Methods. In **Fig. 6**, we use several real-world plants to compare our approach with recent single-view based techniques [8, 11, 18]. After reconstructing the 3D plant models from one image, we then observe the reconstructions from another camera view. The input photographs of plants were captured on site using a smartphone camera.

To quantitatively evaluate the reconstructions, we use Colmap [14] to capture dense 3D point clouds for the plants in Fig. 6, serving as approximate ground truths. Then, inspired by prior works on point cloud reconstruction [10, 15], we adopt two metrics to assess the reconstruction quality, i.e., *reconstruction error* and *completeness score*. **Table 3** presents a comparison of the results for these metrics across different methods. The *reconstruction error* (Err) is the average distance of the points in the dense point cloud P to the reconstructed plant mesh M'. The *completeness score* (Comp) represents the percentage of the points in P that have a distance of less than a given threshold x to the reconstruction M'. We report the completeness at three thresholds ($x = 0.10, 0.05, 0.02$), to obverse how much of the

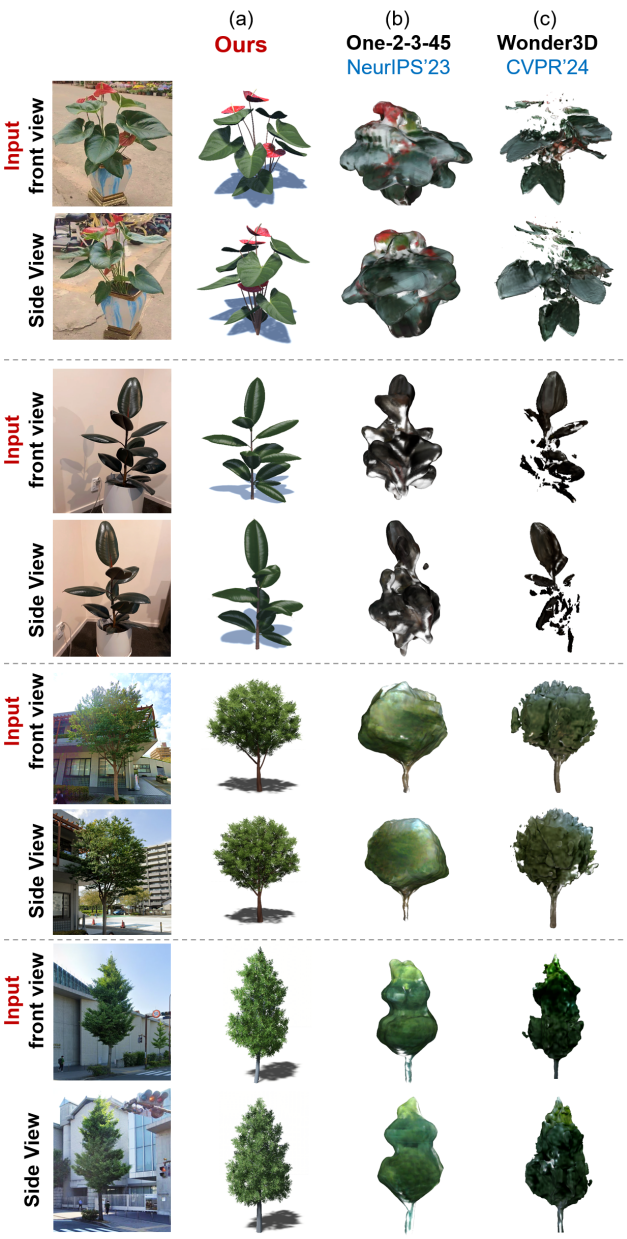


Figure 6. **Comparison to single-image based reconstruction methods on real-world plants.** For each plant, we reconstruct the 3D models from the front view (the first row), and then evaluate the reconstructions from another different view (the second row). From left to right, we show the results of (a) our method, (b-c) two recent diffusion-based approaches[8][11].

area was reconstructed within various degrees of accuracy. For convenience, all plant models were normalized to a maximum height of 1.0 in advance. The results in both Fig. 6 and Table 3 demonstrate that our method can yield 3D plant models with higher geometric quality and consistency than existing single-view-based reconstruction methods in

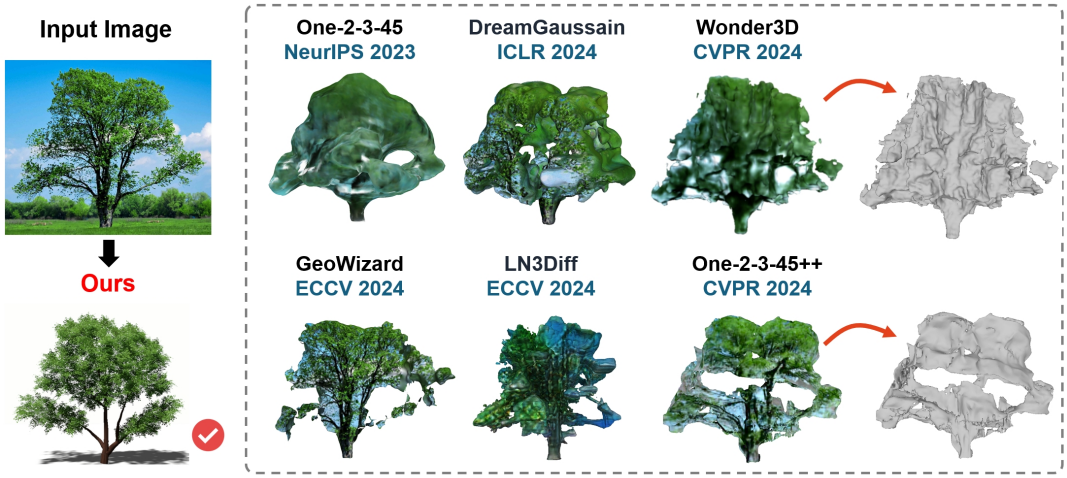


Figure 7. **Comparison with more single-view-based methods using a challenging tree shape.** (a) Our method can produce realistic 3D tree models with high geometric quality. (b) Previous methods [8][18][11][3][7][9] all produce bubble-shaped 3D meshes for trees, limiting their use in practical 3D applications such as games.

generative AI.

Fig. 7 further compares with more single-view-based methods proposed in recent two years [3, 7–9, 11, 18] using a complex tree shape, which has distinct holes and long extended branches. These methods typically follow a similar strategy: they first use diffusion models to synthesize novel views, and then optimize 3D shapes in NeRF-like styles. While these approaches are effective for smooth-surfaced objects, they fail to represent shapes with complex topologies and inner details. Thus, when applied to leafy trees, these methods often produce watertight, bubble-shaped geometries, which cannot meet the requirements for realistic 3D assets in practical applications (e.g., games).

Additionally, Fig. 8 visualizes the underlying geometries of our method more clearly, and compares with a recent state-of-the-art method (i.e., One-2-3-45++ [9]). Our

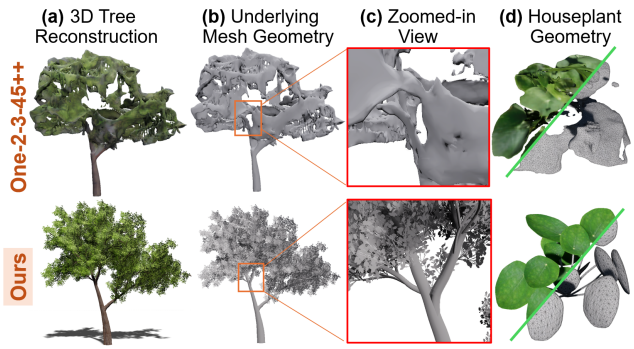


Figure 8. **Underlying Geometry Quality.** We further compare the geometry details more clearly. (a-c) A tree. (d) A houseplant.

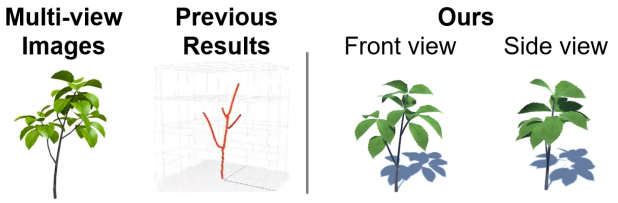
method not only preserves significantly finer geometric details but also produces structured topology, ensuring direct compatibility with downstream applications in computer graphics.

Extra Comparison to Multi-view-based Methods. Reconstruction from multi-view images can directly leverage multi-view stereo techniques to generate 3D point clouds for guiding the subsequent 3D modeling process, making it easier compared to using a single photo alone. Nonetheless, we also conduct additional comparisons with two recent multi-view-based approaches, each specialized for houseplants [5] and trees [4], respectively.

Fig. 9(a) first compares our method with a recent CVPR paper [5] that utilizes a neural network to predict small 3D houseplants from multi-view photographs. However, this approach only produces stroke-like 3D skeletons rather than the complete 3D plant geometries. Fig. 9(b) compares with another recent work [4] for reconstructing outdoor trees from multi-view images. In contrast, our method only requires a single view as input while still achieving a good resemblance to the given photograph. Furthermore, unlike their methods, which are specifically tailored to a single plant category, our approach can adapt to the reconstruction of both houseplants and outdoor trees at the same time.

3. Extended Applications

Not limited to image-based 3D generation, the hierarchical box representation used in this paper can be extended to support a wider range of tasks in plant modeling. Below, we briefly explore its potential in three specific tasks.



(a) Comparison to Isokane et al.[5]'s method

(b) Comparison to Guo et al. [4]'s method

Figure 9. **We additionally compare with multiple-view based methods [5][4],** both of which accept multi-view images as input, and then produce either (a) only rough skeletons of indoor houseplants, or (b) outdoor tree models, respectively. In contrast, our method uses much simpler input (only a single-view image) but can still generate detailed 3D plant geometries. Moreover, our method can adapt to both types of plants at the same time.

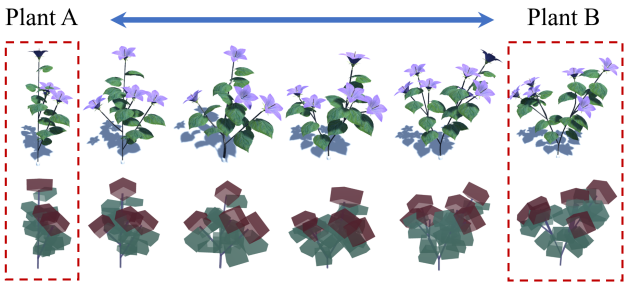


Figure 10. **Extended Application 1**: Shape Interpolation between two plant models (A and B) by directly performing linear interpolation between the feature vectors.

(1) Shape Generation. The plant hierarchies are implicitly encoded in the latent feature space, enabling us to obtain continuous variations between two given plant structures by directly performing linear interpolation on their feature vectors. Fig. 10 shows an example of shape interpolation, where we can observe the smooth transition from plant A to another plant B.

Apart from interpolation, sampling on latent space also allows for easily producing new plants, in line with most generative models. Fig. 11(a) shows a random forest sampled from the latent space trained on the Elm tree dataset. To observe this ability more intuitively, we also did a simple experiment by training on a dataset containing only two species (Oak and Prunus). We select one tree from each species and compute the average of their latent vectors. Based on it, we can generate a new tree that blends the shape features of both species (Fig. 11(b)).

(2) Sketch-based Plant Modeling. Fig. 12 illustrates an interesting application of our method: by training the network on sketch images instead of synthetic photographs, we can easily extend the method into a sketch-based plant modeling system. The training data for the sketches can be readily obtained by applying Canny edge detection to our synthetic plant images with a pure white background.

(3) Scene Reconstruction. In addition to single-plant 3D reconstruction, our method can be applied to reconstruct scenes with multiple plants by using a recent instance segmentation technique [6]. Fig. 13 presents a preliminary example of applying our method to reconstruct a street scene containing multiple trees.

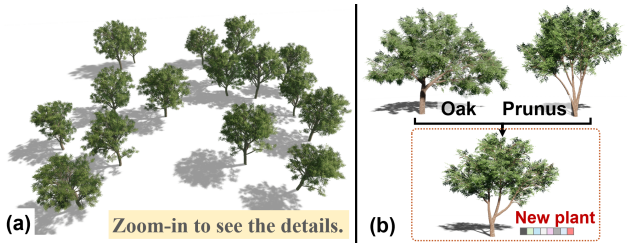


Figure 11. **Explore new plants from latent space.** (a) A randomly-sampled forest. (b) Blend two species into a new tree.

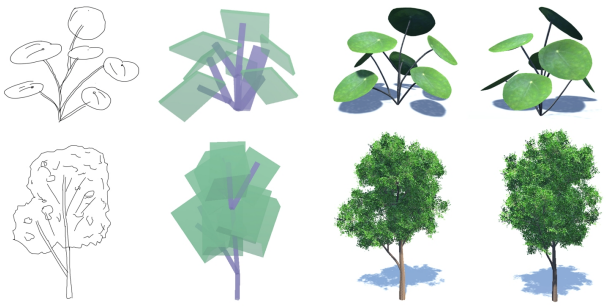


Figure 12. **Extended Application 2**: We can extend the method to a sketch-based plant modeling system by training on sketch images instead of synthetic photographs. The sketch dataset can be constructed using a Canny edge detector.

4. Others

Botanical Parameters. As explained in main paper, terminal boxes are associated with a small set of botanical parameters $\boldsymbol{p}$. They are predicted by sub-decoder Dec_{box}, and then used by parametric modeling modules to generate detailed 3D geometries. Table 4 summarizes the common botanical parameters used in our system, which are designed based on [16] and define the typical appearance features of the

(a) Street photo (b) Front view (c) Top view

Figure 13. **Extended Application 3:** Our method can be used to reconstruct the scenes with multiple plants.

Table 4. List of botanical parameters used in our parametric modeling modules. It can be flexibly expanded by customizing the semantics of reserved parameters $\boldsymbol{P}_{res}$ according to practical needs.

Params	Descriptions
Trees	
n_T	Number of twigs generated on a single node during one growth iteration.
m_A	Mean of angles between twigs.
l_{IT}	Base length of a single internode.
$iter$	Max growth age of the sub-tree.
s_{LF}	Size of leaves attached to the twigs.
α_{GR}	Gravitropism (Influence of gravity on the growth direction).
$\boldsymbol{P}_{res}$	Parameters that are flexibly reserved for different species.
Houseplants	
β_{ABL}	Axial bending angle of a leaf's surface.
β_{RBL}	Radial bending angle of a leaf's surface.
β_{OAF}	Opening angle of a flower.
n_{FP}	Number of flower's petals.
β_{ABP}	Axial bending angle of flower's petals.
β_{RBP}	Radial bending angle of flower's petals.
$\boldsymbol{P}_{res}$	Parameters that are flexibly reserved for different species.

associated plant parts. Moreover, this list can be flexibly expanded by customizing the semantics of reserved parameters $\boldsymbol{P}_{res}$ according to practical needs. For example, $\boldsymbol{P}_{res}$ can indicate the number of serrations when the box is a Monstera leaf.

Limitations. Our approach also has limitations. As shown in Fig. 14, the method may fail when accurate segmentation cannot be achieved. In addition, it cannot process certain species that are not describable by hierarchical boxes, such as willow trees with long, drooping branches. To address these challenges, future work will focus on developing more powerful network architectures, and exploring new representations for plant structures.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 248–255, 2009. 1

[2] Adnan Firoze, Cameron Wingren, Raymond A Yeh, Bedrich Benes, and Daniel Aliaga. Tree instance segmentation with temporal contour graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[3] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 241–258, 2025. 3, 4

[4] Jianwei Guo, Shibiao Xu, Dong-Ming Yan, Zhanglin Cheng, Marc Jaeger, and Xiaopeng Zhang. Realistic procedural plant modeling from multiple view images. *IEEE transactions on visualization and computer graphics (TVCG)*, 26(2): 1372–1384, 2018. 4, 5

[5] Takahiro Isokane, Fumio Okura, Ayaka Ide, Yasuyuki Matsushita, and Yasushi Yagi. Probabilistic plant modeling via multi-view image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2906–2915, 2018. 1, 4, 5

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 4015–4026, 2023. 5

[7] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *European Conference on Computer Vision (ECCV)*, pages 112–130, 2025. 3, 4

[8] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3, 4

[9] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and

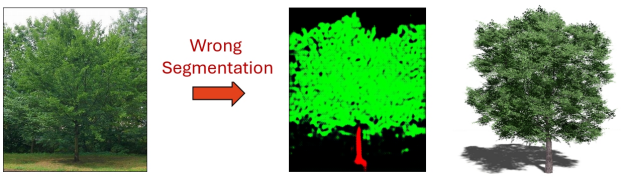


Figure 14. **A Failure Case:** our method fails when the plant in the foreground is indistinguishable from the background.

3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10072–10083, 2024. 3, 4

[10] Yanchao Liu, Jianwei Guo, Bedrich Benes, Oliver Deussen, Xiaopeng Zhang, and Hui Huang. Treepartnet: neural decomposition of point clouds for 3d tree reconstruction. *ACM Transactions on Graphics (TOG)*, 2021. 1, 3

[11] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, Highlight)*, pages 9970–9980, 2024. 3, 4

[12] Wojciech Palubicki, Kipp Horel, Steven Longay, Adam Runions, Brendan Lane, Radomír Měch, and Przemyslaw Prusinkiewicz. Self-organizing tree models for image synthesis. In *ACM Trans. on Graphics (Proc. SIGGRAPH)*, page 58, 2009. 1

[13] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 413–420, 2009. 1

[14] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[15] Neil Smith, Nils Moehrle, Michael Goesele, and Wolfgang Heidrich. Aerial path planning for urban scene reconstruction: a continuous optimization method and benchmark. *ACM Transactions on Graphics (TOG)*, 2018. 3

[16] Ondrej Stava, Sören Pirk, Julian Kratt, Baoquan Chen, Radomír Měch, Oliver Deussen, and Bedrich Benes. Inverse procedural modelling of trees. In *Computer Graphics Forum (CGF)*, pages 118–131, 2014. 5

[17] Jerry O Talton, Yu Lou, Steve Lesser, Jared Duke, Radomír Měch, and Vladlen Koltun. Metropolis procedural modeling. *ACM Transactions on Graphics (TOG)*, 2011. 1

[18] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3, 4